VOLUME 46 • SPECIAL ISSUE • NOVEMBER 2023

ISSN 0971-3034

THE INDIAN JOURNAL OF TECHNICAL EDUCATION

Published by
INDIAN SOCIETY FOR TECHNICAL EDUCATION
Near Katwaria Sarai, Shaheed Jeet Singh Marg,
New Delhi - 110 016

INDIAN JOURNAL OF TECHNICAL EDUCATION

Volume 46 • Special Issue • November 2023

Indexed in the UGC-Care Journal list

INDIAN JOURNAL OF TECHNICAL EDUCATION

Published by
INDIAN SOCIETY FOR TECHNICAL EDUCATION
Near Katwaria Sarai, Shaheed Jeet Singh Marg
New Delhi - 110 016

Editorial

Industry-Institution Partnership is a Key to Produce Ready for Job Graduates: Every year, India is producing a large number of graduates compared to many other countries. But the point is how many of them are employable? Many of them are not able to fulfil the requirements to secure jobs in the industries/companies. According to the "State of Working India 2023", in 2021-22 over 42 % of India's graduates under the age of 25 years were unemployed!

Those were the days, only higher percentage or grades secured by the graduates in the examinations considered for jobs. Now the scenario is changed. Sometimes, even the top academic performers are not able to get the jobs, even if they get may be with a lower salary package. Why so? Many universities and institutions in India are not updating the curriculum to suit the job requirements of industries. Hence, there is a need to revamp the curriculum by introducing emerging technologies, recent processing methods adopted in industries, hands-on learning experiences, internships, etc. Also, the essential expertise on soft skills like communication skills, listening skills, speaking ability, time management, goal setting, conflict management, teamwork, positive thinking, entrepreneurship, leadership and so on are more important. The industries/companies will provide the inputs to educational institutions to prepare the curriculum to meet their requirements, accordingly the institutions can educate their students.

It is need of the hour that every academic institution should have a collaboration with various kinds of industries/companies to educate the students both theoretically and practically with required skills that will help to produce ready for job graduates. This is a win-win situation for the students as well as the industries/companies.

New Delhi **Editor**

30th November 2023

CMR Technical Education Society

(Accredited by NBA & NAAC with 'A' Grade Approved by AICTE
Affiliated to JNTUH)
Kandlakoya (V) Medchal Road, Hyderabad 501 401

Contents

Motorcyclists Riding Without A Helmet: Automatic Number Plate Detection

Pasupunuti Vivek, Chityala Vishnuvardhan
Putta Suchay
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

Ch Rekha
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ ch.rekha445@gmail.com

ABSTRACT

Several solutions exist with the potential to fix the current problems with Indian traffic legislation. One traffic violation that has contributed to an upsurge in accidents and deaths in India is the practice of riding motorbikes or mopeds without protective headgear. Currently, the majority of traffic violations are recorded using CCTV footage. If the biker is not wearing a helmet, the traffic authorities need to find the exact moment the offence happens and focus on the licence plate. The increasing number of bike riders and the frequency of traffic offence make this a time- and labor-intensive task. To automate the identification of this traffic infringement and the extraction of the vehicle's licence plate number, this research study project designs a system to identify non-helmeted motorcyclists. The main point is that Item Discovery makes advantage of Deep Learning at the 3 degree level. The things that may be identified include a person, a motorbike or scooter, a helmet, and a license plate. The first two levels use YOLOv2, while the third level uses YOLOv3. Then, the number plate number is extracted using optical character recognition (OCR). Actually, we built a replacement system that can identify helmets and eliminate licence plate numbers by using additional of the aforementioned approaches.

KEYWORDS: *Helmet detection, Number plate recognition, You only look once, Deep learning, Optical character recognition, Convolutional neural networks.*

INTRODUCTION

Headgear reduces the likelihood that the skull will fracture, effectively reducing head activity to zero. In the event of an accident, the padding inside the helmet will not only absorb the force but also gradually immobilize the head. It also protects the brain from severe injuries by distributing the impact across a larger area. The most critical function is to mechanically protect the cyclist's head from whatever may have touched it [1]. Wearing a high-quality full-face safety helmet may lessen the likelihood of injury. The purpose of enforcing rules on website traffic is to encourage restraint, which in turn reduces the likelihood of major injuries and deaths. On the other hand, these regulations are not followed religiously in practice. As a result, we need to find workable answers to these difficulties. Manual traffic security utilizing CCTV is one possibility. On the other hand, a substantial amount of human resources is required to finish several iterations before the objective may be attained [2]. Therefore, this inefficient manual technique of helmet locating is not manageable in cities with huge people and lots of vehicles operating on the highways. Following this, we provide a system for full safety helmet identification and licence plate removal that makes use of YOLOv2, YOLOv3, and optical character recognition. Helmet discovery systems often begin with gathering datasets, then move on to object localization, background removal, and semantic network item classification [3].

RESEARCH CONTEXT

Observation 1

The research paper "Headgear Visibility Category with Motorcycle Detection and Tracking" by J. Chiverton

[1] presents a method for the automated categorization and tracking of bike riders who are wearing and who are not wearing headgear. Using a combination of static pictures and individual photo structures extracted from video clip data, the approach employs support vector machines that were trained on pie charts representing the head regions of motorcyclists. By using background reduction, motorcycle riders may be quickly extracted from video footage, and the trained classifier is then incorporated into a radar system. Before a trained classifier can categorize the riders, their heads are separated. A track is a series of closely spaced locations that each motorcyclist builds. To fully identify these tracks, we then take the mean of all the classifier findings. Research shows that the classifier can reliably detect whether motorcyclists are using protective headgear in static images. The effectiveness and validity of the category method are further shown by radar testing.

Observation 2: Rattapoom Waranusast, Nannaphat Bundon, Vasan Timtong, and Chainarong Tangnoi detail their system in their research article "Machine Vision Strategies for Motorcycle Helmet Discovery"; the system can detect motorcyclists and tell whether they are wearing helmets. [2] in In order to distinguish moving items, such as bikes, the system employs data collected from neighbouring buildings and the K-Nearest Neighbour (KNN) classifier. After that, we split and total up the riders' heads on the recognised bike based on the estimation profile. The method classifies the head as either wearing or not wearing a safety helmet using characteristics extracted from four regions of the segmented head. The experiment found that the average accurate discovery price for the close lane was 84%, the far lane was 68%, and both lanes were 74%.

Monitoring 3: A technique for automatically recognising motorcyclists without helmets and a method for automatically detecting vehicles were presented in the term paper "Motorcyclist's Headgear Using Discovery Using Image Handling" by writers Thepnimit Marayatr and Pinit Kumhom. The work also classified motorbikes on public highways. [5] To begin with, we use the back subtraction approach to remove the background from the foreground images, and then we improve them using the threshold and mathematical morphology method. This allows us to detect moving cars in real-time. Differentiating motorcycles from other types of vehicles is the second stage. Attribute removal is applied to location, and semantic network classification is further requested. Finding a hat in the last stage yields Hough transform. The testing results showed that headgear discovery had an accuracy rate of 98.22% and motorbike classification a rate of 77%.

Observation 4: Xinhua Jiang's research paper "A Research of Low-resolution Safety Helmet Image Acknowledgment Incorporating Statistical Features with Artificial Neural Network" examined the safety helmet recognition method in low-resolution video clip images and also assessed the low-resolution safety helmet recognition problem at the cross-coordinate by deducing the relationship between different attributes and acknowledgment price [8]. After locating the heads in the surveillance footage, it extracted analytical characteristics using a neighbourhood binary pattern and a grey level co-occurrence matrix. Lastly, the test sample's recognition rate was computed using a back-propagation artificial neural network. Using the GLCM statistical characteristics in combination with the BP man-made semantic network, the experiment correctly identifies the helmet.

Observation 5:

"Automatic Discovery of Bike-Riders without Headgear Using Surveillance Videos in Real-time" [9] proposed a system for automatically detecting helmet less bikers using security video clips in real-time. The authors of the study were Kunal Dahiya, Dinesh Singh, and C. Krishna Mohan. To begin identifying motorcyclists in surveillance film, the suggested technique employs object division and history reduction. The next stage is to determine whether the biker is wearing a cap by using an aesthetic function and a binary classifier. They also provide a combined method for masking violations, which makes the recommended technique more reliable. To evaluate the efficacy of their method, they laid up a comparison of three popular feature representations: regional binary patterns (LBP) for classification, scale-invariant attribute change (SIFT), and pie chart of oriented slopes (HOG). It is possible that 93.80% of the real-world surveillance data was discovered, according to the testing findings.

EXISTING REGIME

To detect illegal internet traffic, the present method relies heavily on CCTV records. In order to identify motorcyclists who aren't wearing protective headgear, traffic officials need to focus on the frame containing the infraction and examine the licence plate. The increasing number of bikers and the prevalence of traffic violations online make this a laborious and time-consuming task. This task has been accomplished by recent studies using CNN, R-CNN, LBP, HoG, HaaR traits, etc. The efficiency, precision, and rate of discovery of items and categories are all negatively impacted by these jobs [3, 4].

Problem Statement

Machine learning (ML) is a subfield of AI in which a "trained" model may potentially learn to solve problems autonomously given enough training data. A mathematical model of certain data, called "training material," is what machine learning algorithms use to make predictions or assessments. Object detection apps also make advantage of them. Thus, upon training with a specific dataset, a Headgear discovery version may be used. Motorcyclists who choose not to wear protective headgear may be easily identified by this innovative design. Using the tracked routes, we are able to extract the rider's licence plate and save it as an image. After receiving this picture, an optical character recognition (OCR) model deciphers the text and, as a result, delivers the number plate number as maker-encoded text. Also, it can be done in real time using a camera.

MODULES

Details Collected

For training, 50,000 models were used across 5 classes with 11,000 small YOLOv3 photos. Due to the great accuracy of all object course detection and the mean average precision (mAP) reaching a continuous maximum value of 75%, training was terminated after 50,000 iterations. identification of protective headgear.

In order to train for the customised courses, the YOLOv3 model is fed annotated photographs. You may use the weights you make while exercising to fill up the version. Afterwards, an image is provided as input. Among the five courses that were taught, the model could identify all of them. Here you may find details on certain motorcyclists. Even without a helmet, we can readily get the rider's other course information. The licence plate may be extracted using this.

Recovery of Licence Plates

The associated person course is identified when the biker without a helmet is found. To do this, we check to see whether the works from the no-headgear course are within the person course. When looking for a certain motorbike and licence plate, same steps are used. Once the licence plate's uses are determined, it is clipped and stored as a new picture.

Recognition of Licence Plates: With the removed licence plate in hand, an OCR model is trained. After optical character recognition (OCR) detects text in an image, it produces the recognised strings as part of the machine-encoded message. Without a doubt, the optical character recognition module will provide a confidence score and a list of possible licence plate numbers. This degree of certainty shows how sure it is that it can still recognise the given licence plate. After that, a text file is created to save the licence plate with the highest level of certainty for future reference.

RECOMMENDED METHODS

The purpose of this study is to determine if motorcyclists really use protective headgear. If they aren't, we decal the bike's licence plate. A YOLO CNN version is available, complete with test images of the licence plate and train footage [5].

The following components are being followed or used to put the aforementioned plan into action.

1) The first photo will be sent into the algorithm, and YOLOV2 will be used to identify whether the picture contains a person and an electronic bike. We shall proceed to step 2 if YOLO design identifies both.

2) Here, we'll apply the YOLOV3 design to the task of identifying helmet wearers. In such a case, the software will malfunction. Proceed to step 3 if the rider is not wearing a helmet after that application.

The third step is to remove the licence plate data using the Tesseract OCR API in Python. Later on, OCR will be able to get the vehicle number from the supplied picture.

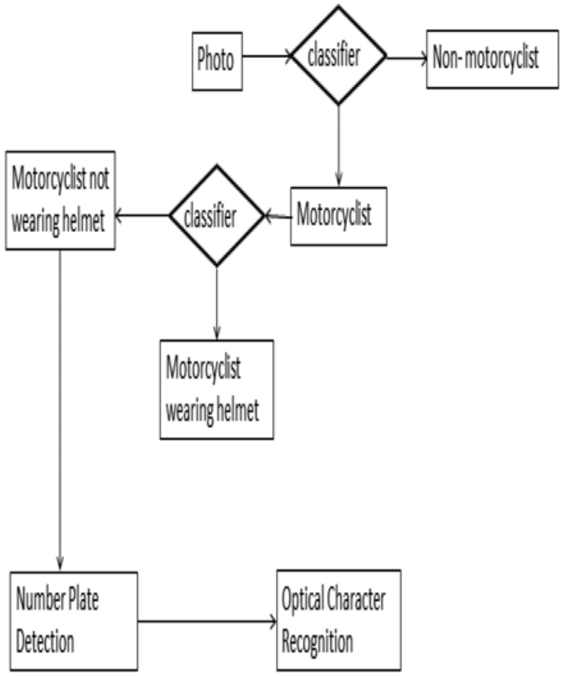


Photo
classifier
Non-motorcyclist
Motorcyclist
classifier
Motorcyclist not wearing helmet
Motorcyclist wearing helmet
Number Plate Detection
Optical Character Recognition

Figure 1. Flow chart for the proposed methodology

Figure 1 up there explains the essence of the procedure outlined below.

1. Bicycle and Non-Bicycle Classifiers

Both the "Motorcycle" and the "Person" classes in the "YOLOv2 Things Discovery Version." get the framework as an input. This process generates an image that includes the required route detection, discovery confidence inside the bounding box, and chance value [6].

Using the characteristics supplied by Picture AI collection, just the recognized objects are extracted, stored as separate photographs, and called sequentially with class name and photo number. The picture of a person, for instance, would be preserved as person-1, person-2, etc., whereas a motorbike, for instance, would be preserved as motorcycle-1, motorcycle-2, etc [7, 8]. The information about these erased photos is stored in a dictionary that may be accessed for further analysis at a later time.

2. Indicators for motorcyclists using protective headgear and those opting not to do so: The person's picture is fed into the helmet detection model once the motorbike and rider have been recognised. There were a few false positives discovered when checking the headgear model. Consequently, after being cropped, the original photo was reduced to only the upper quarter. This prevents the possibility of false positives caused by riders not wearing their safety helmets or by riders accidentally leaving their helmets on their motorcycles while they ride.

Thirdly, optical character recognition: The Transportation Office may find out which motorcyclists are at fault by using Google's Tesseract Optical Character Recognition to decipher the licence plates.

You can see how well each classifier performed on the test data by looking at their recall, precision, and precision metrics. In our experiment, the effectiveness of each classifier is determined by its accuracy. The accuracy is determined using the following formula.

$$Accuracy = \frac{No.\ of\ samples\ correctly\ classified}{Total\ number\ of\ samples}$$

RESULT ANALYSIS

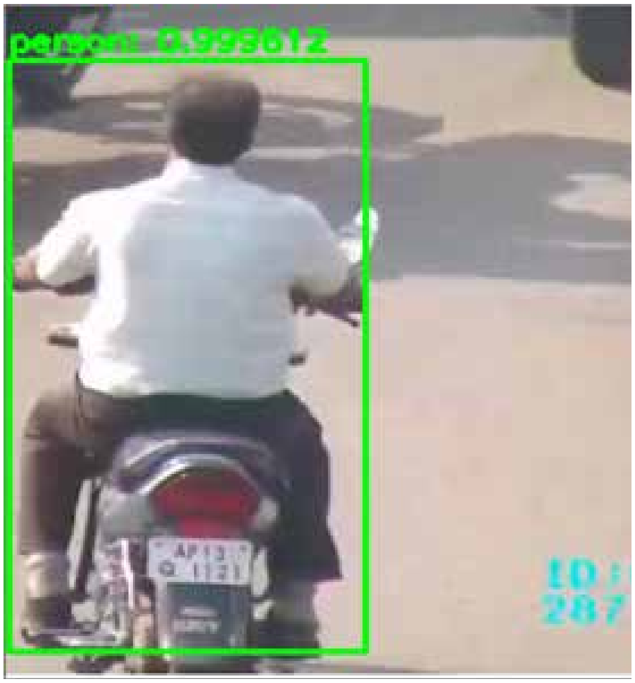


Figure 2. Detection of motorcycle and person (not wearing helmet)

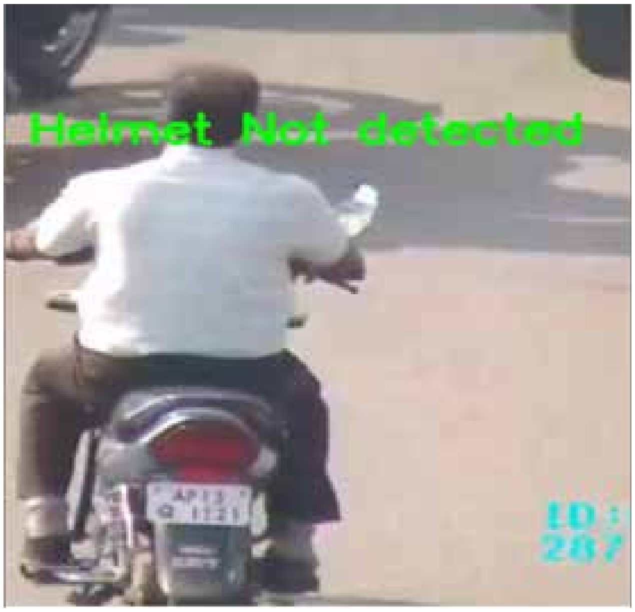
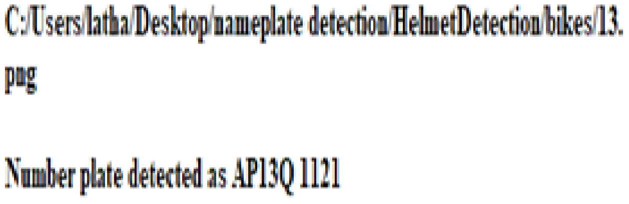


Figure 3. Detection of helmet


C:/Users/latha/Desktop/nameplate detection/HelmetDetection/bikes/13.
png

Number plate detected as AP13Q 1121

Figure 4. License plate registration number is obtained using optical character recognition

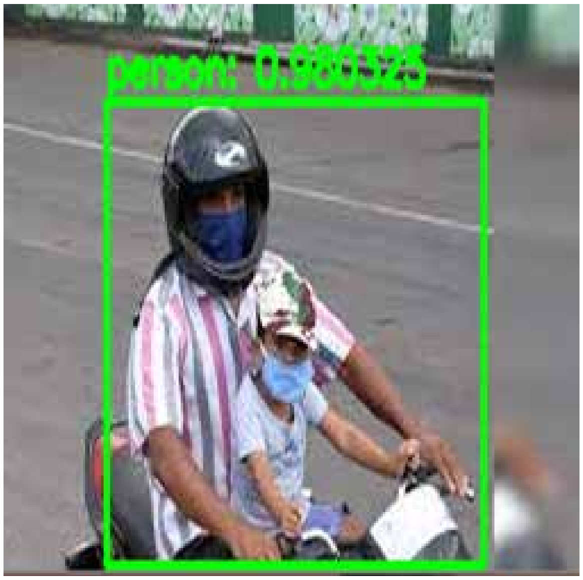


Figure 5. Detection of motorcycle and person (wearing helmet)

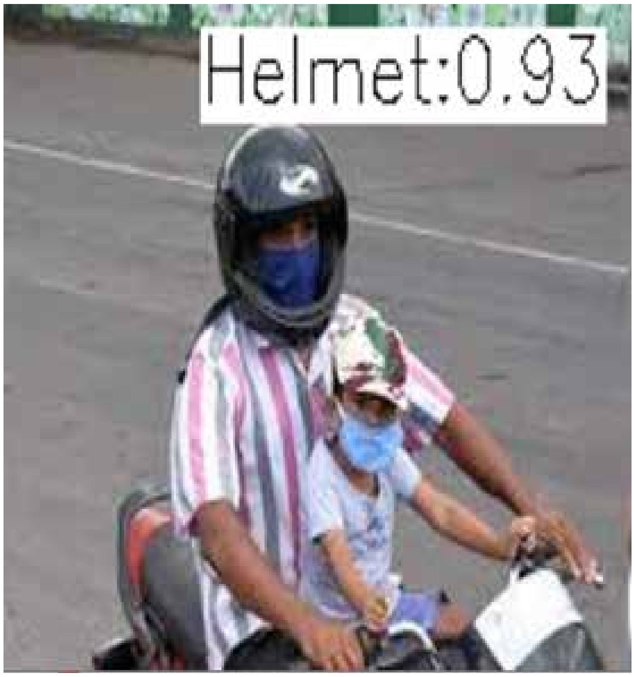


Figure 6. Detection of helmet with probability value

Our experiment's safety helmet vs. non-helmet classifier achieved a rate of 99.64 percent, while the motorbike vs. non-motorcycle classifier achieved a rate of 99.78 percent. Consequently, 99.42% was the total accuracy for identifying motorcyclists who were not wearing protective headgear [9].

FINAL CONCLUSION

The input is camera data, and the output is a method for discovering riders without helmets. If the motorcyclist in the picture is not wearing a helmet, the number plate number of the bike will be shown. A YOLO-style object finding notion is used for bike, person, helmet, and licence plate detection. The number may be extracted from the number plate using optical character recognition (OCR) if the rider is not using a protective helmet. In the sake of finding new uses for the framework, not only are the personalities stripped away, but so is the framework itself. We have accomplished all of the goals set forth for this project.

ACKNOWLEDGMENT

We appreciate St. Peter's Design University's help with the research lab and the subsequent assistance in making this report more aesthetically pleasing.

REFERENCES

1. J. Chiverton, “Helmet Presence Classification with Motorcycle Detection And Tracking”, IET Intelligent Transport Systems, Vol. 6, Issue 3, pp. 259–269, March 2012.

2. Rattapoom Waranusast, Nannaphat Bundon, Vasan Timtong and Chainarong Tangnoi, “Machine Vision techniques for Motorcycle Safety Helmet Detection”, 28th International Conference on Image and Vision Computing New Zealand, pp 35-40, IVCNZ 2013.

3. Romuere Silva, Kelson Aires, Thiago Santos, Kalyf Abdala, Rodrigo Veras, Andr´e Soares, “Automatic Detection Of Motorcyclists without Helmet”, 2013 XXXIX Latin America Computing Conference (CLEI). IEEE,2013.

4. Romuere Silva, “Helmet Detection on Motorcyclists Using Image Descriptors and Classifiers”, 27th SIBGRAPI Conference on Graphics, Patterns and Images.IEEE, 2014.

5. Thepnimit Marayatr, Pinit Kumhom, “Motorcyclist‟s Helmet Wearing Detection Using Image Processing”, Advanced Materials Research Vol 931- 932,pp. 588-592,May-2014.

6. Amir Mukhtar, Tong Boon Tang, “Vision Based Motorcycle Detection using HOG features”, IEEE International Conference on Signal and Image Processing Applications (ICSIPA).IEEE, 2015.

7. Abu H. M. Rubaiyat, Tanjin T. Toma, Masoumeh Kalantari-Khandani, “Automatic Detection of Helmet Uses for Construction Safety”, IEEE/WIC/ACM International Conference on Web Intelligence Workshops(WIW).IEEE, 2016.

8. Xinhua Jiang “A Study of Low-resolution Safety Helmet Image Recognition Combining Statistical Features with Artificial Neural Network”. ISSN: 1473-804x

9. Kunal Dahiya, Dinesh Singh, C. Krishna Mohan, “Automatic Detection of Bike-riders without Helmet using Surveillance Videos in Real-time”, International joint conference on neural network(IJCNN). IEEE, 2016.

10. Maharsh Desai, Shubham Khandelwal, Lokneesh Singh, Prof. Shilpa Gite, “Automatic Helmet Detection on Public Roads”, International Journal of Engineering Trends and Technology (IJETT), Volume 35 Number 5- May 2016, ISSN: 2231-5381

11. R. Smith, “An Overview of the Tesseract OCR Engine,” Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Parana, 2007, pp. 629- 633.

12. Z. Xu, X. Baojie and W. Guoxin, “Canny edge detection based on Open CV,” 2017 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI), Yangzhou, China, 2017, pp. 53-56.

13. J. Deng, W. Dong, R. Socher, L. J. Li, Kai Li and Li FeiFei, “ImageNet: A large-scale hierarchical image database,” 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, pp. 248-255.

14. Soo-Chang Pei, Ji-Hwei Horng, “Circular arc detection based on Hough transform”, Pattern Recognition Letters, Volume 16, Issue 6, 1995, Pages 615-625.

Blockchain Based Certificate Validation

Shrigiri Divya, Bollepalli Lahari
Maddineni Sri Harsha
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

K Srujan Raju
Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ ksrujanraju@gmail.com

Tabeen Fatima
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

ABSTRACT

We are digitising all academic credentials as part of this project. We will definitely store them on a blockchain web server to make them far more safer and prevent fraudulent. Nobody can hack or alter the data stored on the blockchain web server since it enables immutable storage. A warning will be sent to the user in the event that verification fails at the next block storage space due to information being changed. The use of hash codes to validate purchases across several servers is a crucial feature of blockchain technology. Because the hash codes for the same information change if one web server modifies it, all of the other web servers will also detect the modification. For instance, in Blockchain modern technology, data is stored on multiple servers. If someone tries to alter the data on one server, only that server's hash code will be changed. The other servers will remain unchanged. When the confirmation time comes around, this changed hash code will be detected, allowing future malicious customer modifications to be prevented. The data is considered original and unmodified if the hash codes used to record it on a blockchain don't change, and new blocks containing transaction data may be added to the blockchain at any time. With each new data store, we will verify the hash codes of all blocks.

KEYWORDS: *Blockchain, Server, Tamper, ML.*

INTRODUCTION

Building and releasing the system that addressed these issues is the primary goal of the project. A thorough assessment of the system's security is part of the research. According to the results, the implementation is safe, reliable, and practically suitable. Possible clues to important architectural issues with other blockchain-based systems' security features might be found here. Here, we take a close look at the implementation from the standpoint of the system architecture and database. The engineering-centric design of the system is described by its system and database architectures [1].

The primary business logic, including certificate application, assessment, signing, and issuance, is handled by certificate-issuing apps. The majority of the community's signatories trigger the certificate-issuing software to add the Merkle tree and all certificate hashes to Blockchain [2]. Revocation of a certificate is a possible step in the issuing process. Certification applications, tests, signatures, and issuance are all handled by the issuing apps, which are vital to the company. A Merkle tree is included in the certificate hash by the algorithm that creates the certificates. This allows the Merkle root to be added to the Blockchain. A certificate's revocation may be handled by the same software that issued it [3].

A primary goal of the verification application is to guarantee the validity and authenticity of all issued certificates. It mostly consists of a website and an Android app. Using the same method, they compare the transaction message they received from the blockchain API with the verification data on the receipt. A simplified explanation of how it works is as follows: Ascertain this: Prior to the authentication code being deemed valid, the following matters must be addressed: There are a number of requirements that must be met, including compatibility with the local certificate, presence in the Merkle tree, blockchain presence of the Merkle root, non-revocation of the certificate, and validity of the expiration date [6]. Notably, the Android app streamlines certificate delivery by enabling quick document verification via QR code scanning [5]. Blockchain, a distributed ledger system, stores authentication data and functions as a trust architecture. The Merkle root, comprised mostly of authentication data, is often constructed from the hashed data of many certificates. We chose MongoDB as our database because of its scalability, high availability, and support for certifications that are based on JSON [8]. The widespread use of mobile devices, the proliferation of the Internet, and other IT developments have caused a shift in people's daily routines. There has been a dramatic uptick in the widespread use of virtual money, or digital currency created for use only on the internet. Internet accessibility has contributed to the appreciation of several digital currencies, the most prominent of which being Ripple, Bitcoin, and Ether [2]. The revolutionary currency's underlying blockchain technology is starting to get some attention. The immutable record keeping and distributed ledger properties of blockchain technology could be very useful in many other areas as well [9].

Emerging technologies like distributed ledger technology, or blockchain, are changing the way financial transactions are recorded. Once everyone is in agreement, the transaction is added to a block that contains the records of several previous transactions. With each connection-related block, you may get the previous block's hash value. Decentralisation occurs when data is stored across several nodes, a process known as distributed data storage, which is the building block of a blockchain [1]. A correct database requires cooperation among the nodes. In blockchain technology, a block is considered valid when it has been confirmed by several nodes.

LITERATURE SURVEY

Existing System

Due to the centralised storage and human verification of certificates, the verification process is very time-consuming. No commercial sector (bank) can rest certain that the certifications they get are secure. However, the data may be altered, removed, or amended. It is easy to hack a certificate and create a copy of it. When students go to their interviews, they should carry their certificates. Certificates are not secure.

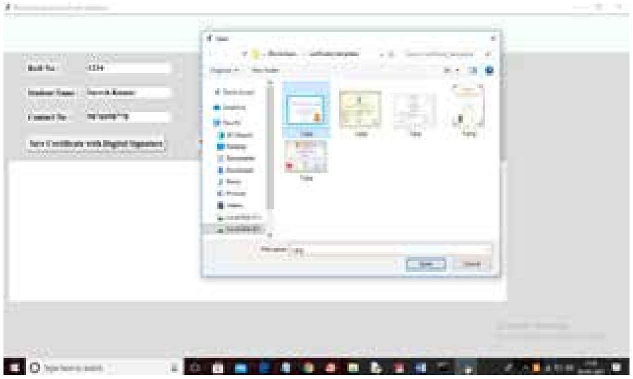
Proposed System

A blockchain certificate system was created using applicable technologies in this research. The EVM executes the system's application, which was developed on the Ethereum platform. Three distinct user types participate in the system: A certification unit or school may issue a certificate, use the system, and search the database. The method allows the authorities to give certificates to pupils whenever they meet specified standards. Students are free to ask questions about their certificates after they have received them.

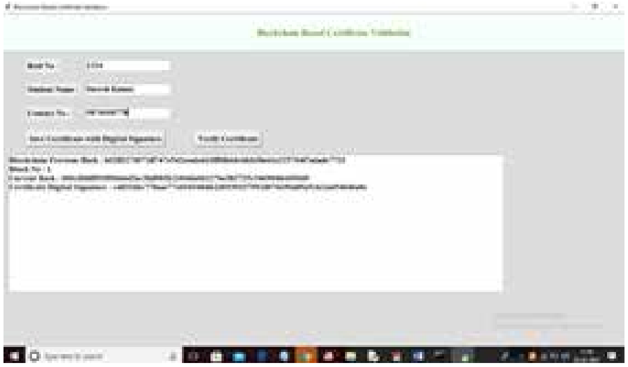
METHODOLOGY

Building the System A blockchain certificate system was created using applicable technologies in this research. The EVM executes the system's application, which was developed on the Ethereum platform. Three distinct user types participate in the system: A certification unit or school may issue a certificate, use the system, and search the database [9]. The method allows the authorities to give certificates to pupils whenever they meet specified standards. Students are free to ask questions about their certificates after they have received them. The programme Process One example of a distributed ledger technology is blockchain. Here are the steps involved in how the system that was created for this research works: Degrees are awarded and student information is entered into the system by schools. After that, a blockchain is immediately updated with the student's unique identifier. All data is validated by the certificate system. After their information is properly confirmed, students get e-certificates with a QR code instead of the traditional physical copy. A digital copy of the certificate and an inquiry number are also sent to each graduate. Graduates may apply for jobs by sending their serial numbers or QR codes from electronic certificates to specific firms. When the system verifies the serial numbers, the firms get an email or other notification. They can see whether the certificate has been altered or faked thanks to the QR code [10].

Working

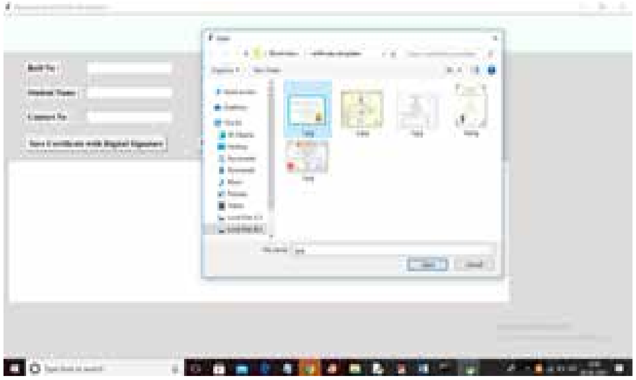


The following page will appear when you input the student's information in the previous screen, click the "Save Certificate with Digital Signature" button, choose and upload the "1.jpg" image, and then click the "Open" button:

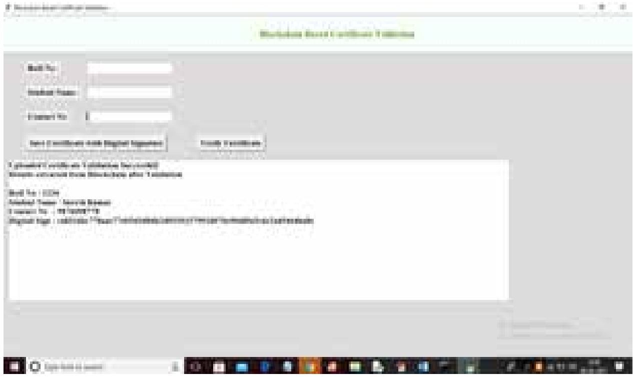

Blockchain Based Certificate Validation
Roll No
Student Name
Contact No
Save Certificate with Digital Signature
Verify Certificate

```
Blockchain Previous Hash : [illegible]
Block No : 1
Current Hash : [illegible]
Certificate Digital Signature : [illegible]
```

Before storing new blocks, blockchain verifies both old and new hash codes to ensure data is not changed. The preceding screen shows that it generates an older hash using block number 1 and its current hash. With each certificate upload, it keeps creating additional blocks. You can see the process compare the new record's hash against the old document's hash as it runs. By sharing the same or similar photographs and hitting the "Verify Certificate" button, the versifier could receive the following result after storing the data to the blockchain.

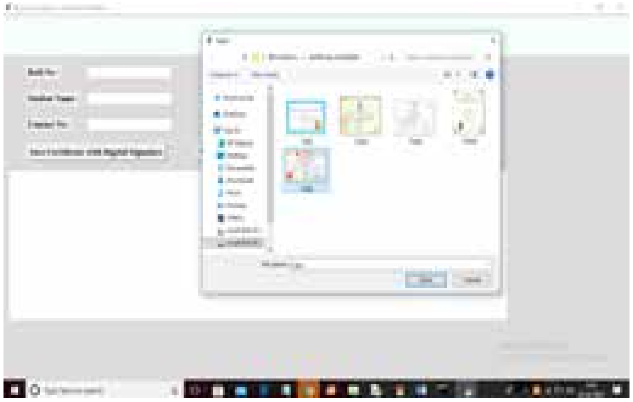


With the '1. jpg' file selected and submitted in the previous screen, click the 'Open' button to get the following output.

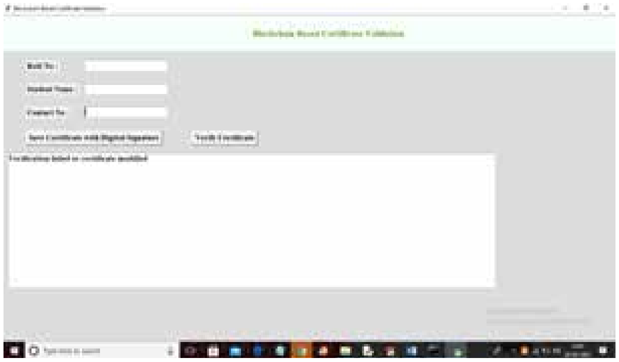

Blockchain Based Certificate Validation
Roll No
Student Name
Contact No
Save Certificate with Digital Signature
Verify Certificate

```
Uploaded Certificate Validation Successful
Details extracted from Blockchain after Validation

Roll No : 1234
Student Name : [illegible]
Contact No : [illegible]
Digital Sign : [illegible]
```

Above, we uploaded the same and suitable photo so the software could detect the digital trademark, then we retrieved the relevant data from the blockchain, and finally, we tried it again with a different image.



Click the "Open" button after selecting and publishing the "5. jpg" files in the previous screen to get the following output.


Blockchain Based Certificate Validation
Roll No:
Student Name:
Contact No:
Save Certificate with Digital Signature
Verify Certificate

Verification failed or certificate modified

The supplied certificate does not match the certificates kept on the blockchain, which is why confirmation failed on the above page. Any other kind of certification may be converted to an electronic signature using the same approach.

CONCLUSION

The credential system developed by the MIT Media Laboratory in June 2016 using blockchain technology is more secure, reliable, and harder to forge than current alternatives that depend on third-party adjudication. However, there are several major verification concerns and a vulnerable abrogation method that limit the project's utility and prevalence. In an attempt to solve these difficulties and make the idea more feasible, we proposed and built a set of unique cryptographic mechanisms, including multi-signature, a BTC-address-state-based retraction device, and believable coordinated recognition.

Compared to the other methods, the multi-signature technique makes the issue of forging much worse as it demands the trademark of the majority of academic board members on each progress document. It greatly improves the security of private important storage as several devices and people may access the unique secrets. Not only that, but cancellation treatment based on Bitcoin addresses is much more trustworthy and accessible than earlier methods, which greatly enhances the stability of certification retraction. Furthermore, this method lessened the likelihood of cancellation failure as the termination process follows the same multi-signature methodology and involves several individuals. Reliable federated identity showed the certification's legitimacy in a new way by using the trusted route and federated identification. Digital best security and agreement proof are two adjacent domains that can benefit from our work's methodology. In contrast to the traditional model dependent on a third party, our approach allows both companies to link their contract to the blockchain using multisignature, hence removing concerns related to credential forging.

In addition, we built a certificate system that adhered to all of the specified protocols and was based on the blockchain using Java and JavaScript. This approach has partially resolved the issue with Blockcerts and has made the idea of a blockchain-based certification more practical. Lastly, we conducted a series of security tests that took into account four distinct perspectives: method, data, network, and functionality. The assessment's findings make it very evident that the system meets the security standards necessary for the business application.

Lastly, I should include a few of important cautions, although they are outside the scope of this article: Thousands of people working in the cryptocurrency ecosystem maintain the Bitcoin blockchain, which our employment is founded upon. Because the blockchain ecosystem and business models are impacted by all types of stakeholders, it would be naïve to think that Bitcoin will continue to operate flawlessly indefinitely. With the help of additional blockchain resources, such as Hyperledger and Ethereum, we want to one day eradicate these features of instability.

REFERENCES

1. Tengyu Yu, Blockchain operation principle analysis: 5 key technologies, iThome, https://www.ithome.com.tw/news/105374

2. JingyuanGao, The rise of virtual currencies! Bitcoin takes the lead, and the other 4 kinds can't be missed. Digital Age, https://www.bnext.com.tw/article/47456/bitcoinether-li tecoin-ripple-differences-betweencryptocurrencies

3. Smart contractswhitepaper, https://github.com/OSELab/learning-blockchain/blob/ master/ethereum/smart-contracts.md

4. Gong Chen, Development and Application of Smart Contracts, https://www.fisc.com.tw/Upload/b0499306-1905- 4531-888a-2bc4c1ddb391/TC/9005.pdf

5. Weiwei He, Exempted from cumbersome auditing and issuance procedures, several national junior diplomas will debut next year.iThome, https://www.ithome.com.tw/news/119252

6. Xiuping Lin, “Semi-centralized Blockchain Smart Contracts: Centralized Verification and Smart Computing under Chains in the Ethereum Blockchain”,Department of Information Engineering, National Taiwan University, Taiwan, R.O.C., 2017.

7. Yong Shi, “Secure storage service of electronic ballot system based on block chain algorithm”, Department of Computer Science, Tsing Hua University, Taiwan, R.O.C., 2017.

8. Zhenzhi Qiu, “Digital certificate for a painting based on blockchain technology”, Department of Information and Finance Management, National Taipei University of Technology, Taiwan, R.O.C., 2017.

9. Weiwen Yang, Global blockchain development status and trends,

10. Benyuan He, “An Empirical Study of Online Shopping Using Blockchain Technology”, Department of Distribution Management, Takming University of Science and Technology, Taiwan, R.O.C., 2017.

Crime Data Analysis and Prediction using Decision Tree

Gaini Vamshi, K Shiva, Neeradi Kanakaraju
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

P Santhuja
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ santhuja.pandu@gmail.com

ABSTRACT

A methodical way to detect crimes is via crime analysis and prediction. This technique may identify areas with a high crime likelihood and show those areas where crimes are more likely to occur. Data mining is a technique for discovering new and valuable information in large amounts of unstructured data. By analysing the current datasets, we may anticipate the extraction of new information. All around the globe, people are dealing with the perilous issue of crime. The standard of living, economic development, and national prestige are all impacted by criminal activity. In order to safeguard society against criminal activity, people must find innovative ways to enhance crime analytics and use cutting-edge solutions. We provide a system that can examine a given area, identify potential crimes, and forecast the likelihood of specific crimes occurring there. Using a number of data mining approaches, this article describes several forms of criminal analysis and crime prediction.

KEYWORDS: *Crime, Act379, Act302, Data mining.*

INTRODUCTION

A society's standard of living and economic development may be negatively impacted by crime, which is a prevalent social concern [1]. According to research, it is one of the most important factors in deciding where to settle down and what to avoid while travelling [2]. A number of negative social outcomes may result from crime: people become afraid to go out at night, social bonds weaken as a result of frequent avoidance of certain areas, and the community's reputation takes a hit. People may be scared away from a neighbourhood or decide not to stay if they hear rumours about how dangerous it is. The economy takes a hit because of this. Governments and taxpayers bear the brunt of crime's economic impact due to the increased expenditure on law enforcement, courts, and prisons, as well as intangible expenses like victims' psychological distress and diminished quality of life. A large number of crimes nowadays are a major issue in many nations. Crime and criminal behaviour are indeed the subjects of scientific investigation, which aims to shed light on the nature of crime and reveal trends in the criminal justice system. Problems with storage and analysis might arise when dealing with crime data due to its rapidly increasing size. Because this data is inconsistent and inadequate, problems with selecting appropriate methods for analysis emerge. In order to improve crime data analysis, researchers are motivated to study this kind of data due to these difficulties. Problems with storage and analysis might arise when dealing with crime data due to its rapidly increasing size. Because this data is inconsistent and inadequate, problems with selecting appropriate methods for analysis emerge. These problems encourage researchers to study this kind of data in order to improve crime data analysis [3]. The goal of this study is to determine whether a county has a low, medium, or high probability of having violent crimes by applying an appropriate machine learning algorithm to crime data.

LITERATURE SURVEY

A. A Criminological and Criminal Activity Analysis Clinical studies of criminal behaviour, law enforcement, and the criminal justice system are the tools used by criminologists in their quest to identify characteristics of criminal behaviour. [4] Among the most crucial, information extraction techniques may provide substantial support for this subject. Crime analysts operate within the field of criminology and are tasked with investigating and detecting criminal activities and any connections between them. Data extraction techniques work well in criminology due to the large number of datasets documenting criminal behaviour and the complex relationships between various types of information. It is necessary to first define the characteristics of crime in order to build further study. Data mining provides a wealth of useful information that can be of use to law enforcement. [5] Due to the complexity of the process requiring human intelligence and expertise, data mining might be used by the police to aid in their investigation of illegal tasks [6]. The goal here is to incorporate years of human expertise into computer models via the use of data mining. The Second Stage: Criminal Activity Predictability There is a lot of evidence to support the idea that criminal action can be predicted statistically [7], as criminals often stick to what they know would keep them safe. If they have a knack for successfully completing illicit tasks, they tend to repeat it, often in the same spot. While this isn't always the case, it happens often enough for these techniques to be effective. Efforts to characterise criminal activities using concepts such as reasonable judgement, regular activity, and crime pattern. A composite thought is an aggregation of various ideas.

Section C. Assessing Methods for Classifying The bulk of forecasts based on past data are derived via category algorithms. When it comes to creating predictions within certain classes, classification is a supervised method. Assuming adequate training conditions, this method may forecast course tags. Some examples of category algorithms include k-Nearest Neighbours, Support Vector Machines, Weighted Voting, and Artificial Neural Networks. Applying any of these techniques to a dataset will help you find a group of models that can predict the unknown class label. Two parts of the dataset are used for category analysis: the training set and the examination set. The training set is also known as the reliant set, and the examination collection is also known as the independent set. The prediction version is linked to the test collection after the training set has been run through the machine learning algorithm. The purpose of using the following classification algorithms is to make predictions about criminal activities.

METHODOLOGY

Machine learning is the process of spontaneously discovering significant patterns in data. Over the last two decades, it has become an indispensable tool for almost every project requiring data extraction from large datasets. Machine learning powers everything around us. With the help of mobile ads, search engines learn to provide us with the best results, anti-spam software filters our emails, and software that detects credit card fraud protects our wallets. Technology has advanced to the point where digital video cameras can recognise faces and intelligent personal assistant applications on smartphones can understand spoken instructions. Unfortunately, modern cars are equipped with collision prevention technologies that are derived from AI algorithms. Artificial intelligence is widely used in several therapeutic domains, such as bioinformatics, medicine, and astronomy. The complexity of the patterns that need to be identified makes it impossible for a human programmer to provide a clear, granular interpretation of how to complete such jobs under these situations, in contrast to more traditional computer system programmes. Instead than mindlessly following predetermined blueprints, we learn and improve many of our skills via trial and error, much like other sentient beings. Recent advances in artificial intelligence have mostly focused on providing programmes with the capacity to learn and adapt.

Random Forests maximises the design for test data by building numerous classifiers from the training data and then using the sum of their predictions. It is a commonly used set finding approach. As a variance reduction tool, the Random Woodlands method uses a random manner to divide choices in order to avoid overfitting on the training data. In an arbitrary forests classifier, a set classifier $h(x|\theta1)$, $h(x|\theta2)$,..., $h(x|\theta k)$ is built up from a collection of classifiers. A model random vector is used to choose K trees, and each tree in the family, denoted as $h(x|\theta)$, belongs to a certain category. Further, the specification vectors □k are chosen arbitrarily. The training dataset is used to construct each category tree in the ensemble, with each tree using a unique subset $D\theta k(x, y) \subset D(x, y)$.

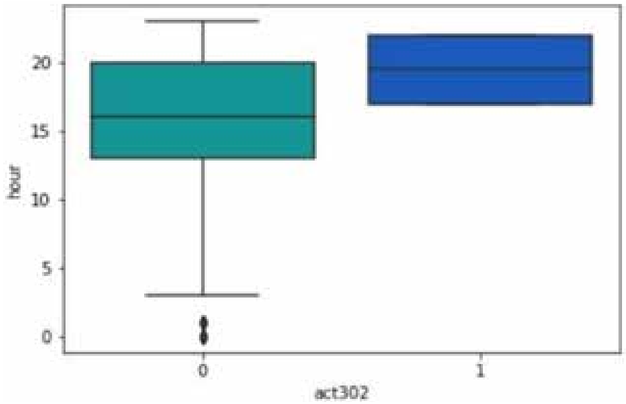


Figure 1. Act of 302 results

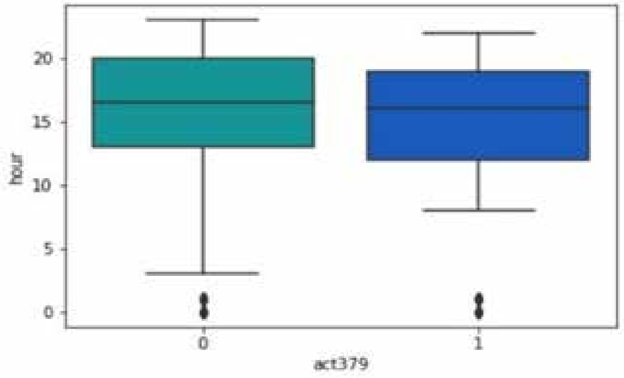


Fig.3.2.Act of 379 results

CONCLUSION

Predictive models for monthly crime rates by crime category were the primary topic of this article. Rising poverty, poor government oversight, widespread corruption, and other issues are all contributing to India's alarmingly high crime rate. In order to take the appropriate measures to decrease crime, the suggested model is very beneficial for both investigative agencies and police officials. This project's interactive visualisations aid in the investigation of criminal networks. Improving this study in the future will include teaching bots to use machine learning to identify high-crime regions. Better predictions are possible with the help of machine learning's advanced concepts because of the similarities between machine learning and data mining. It is possible to boost prediction by improving data privacy, dependability, and accuracy.

REFERENCES

1. O. Karan, C. Bayraktar, H. Gümü¸skaya, and B. Karlık, "Diagnosing diabetes using neural networks on small mobile devices," Expert Syst. Appl., vol. 39, no. 1, pp. 54–60, 2012.

2. F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao, "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," in Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2017, pp. 1903–1911.

3. Z. Liang, G. Zhang, J. X. Huang, and Q. V. Hu, "Deep learning for healthcare decision making with emrs," in Proc. IEEE Int. Conf. Bioinformatics Biomed., 2014, pp. 556–559.

4. C. M. Bishop, Pattern Recognition and Machine Learning. Berlin, Germany: Springer, 2006.

5. Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in Proc. 26th Annu. Int. Conf. Mach. Learn, 2009, pp. 41–48.

6. V. A. Convertino et al., "Use of advanced machine-learning techniques for noninvasive monitoring of hemorrhage," J. Trauma Acute Care Surgery, vol. 71, no. 1, pp. S25–S32, 2011.

7. P. Gueth et al., "Machine learning-based patient specific prompt-gamma dose monitoring in proton therapy," Phys. Med. Biol., vol. 58, no. 13, p. 4563, 2013.

8. J. Labarère, P. Schuetz, B. Renaud, Y.-E. Claessens, W. Albrich, and B. Mueller, "Validation of a clinical prediction model for early admission to the intensive care unit of patients with pneumonia," Academic Emergency Med., vol. 19, no. 9, pp. 993–1003, 2012.

[9. O. Hasan et al., "Hospital readmission in general medicine patients: A prediction model," J. General Internal Med., vol. 25, no. 3, pp. 211–219, 2010.

10. A. K. Diehl, M. D. Morris, and S. A. Mannis, "Use of calendar and weather data to predict walk-in attendance." Southern Med. J., vol. 74, no. 6, pp. 709–712, 1981.

11. J. F. Fernandez, O. Sibila, and M. I. Restrepo, "Predicting ICU admission in community-acquired pneumonia: Clinical scores and biomarkers," Expert Rev. Clin. Pharmacology, vol. 5, no. 4, pp. 445–458, 2012.

12. H. Zhai et al., "Developing and evaluating a machine learning based algorithm to predict the need of pediatric intensive care unit transfer for newly hospitalized children," Resuscitation, vol. 85, no. 8, pp. 1065–1071, 2014.

Stroke Disease Identification System by using Different Types of Machine Learning Algorithms

Mamidipally Rishitha Reddy
Maripedda Praveen, Guglavath Srishanth
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

Sultana Saba
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ sabasultana014@gmail.com

ABSTRACT

As a sub field of AI, machine learning (ML) enables programmed to learn from examples and make precise predictions about future outcomes with little human intervention. The purpose of this study is to catalogue and evaluate the various Machine Learning methods already in use for the purpose of Stroke Prediction. In order to assess the Machine Learning methods utilised for Stroke Predictions, we have looked at the prior literature. In terms of anticipated outcomes, most studies focused on mortality rate and functional result. Neural networks, decision trees, support vector machines, and random forests were the most popular methods. But a handful of classifiers and predictors completed rudimentary reporting criteria for instruments in the medical industry, and none of them was useful.

KEYWORDS: *AI, Stroke predictions, Random forest, ML.*

INTRODUCTION

Stroke affects a large population and is becoming ever more common in underdeveloped nations. A number of variables influence the prevalence of different kinds of stroke. A correlation between risk variables and stroke types is established using predictive algorithms. Algorithms based on machine learning aid in the early detection and avoidance of these stroke instances. Because stroke is a complex medical disorder, it is very difficult to anticipate stroke symptoms and breakouts by considering risk factors [1]. This has piqued the curiosity of those working in the tech industry, who are hoping to find a way to use machine learning to reliably gather datasets on a regular basis and provide reliable diagnostic findings for stroke patients. In addition, there has been a deluge of published papers detailing machine learning algorithms that claim to solve the problem [2] . This survey article aims to find the best machine learning methods for stroke prediction so that we can better understand the issue and find effective solutions [3].

LITERATURE SURVEY

Testing for thrombophilia in young patients hospitalized for ischemic stroke: The role of thrombophilia in this condition is still up for debate. Our objective was to investigate the role of hereditary and acquired thrombophilias in the development of ischemic stroke, TIA, and amaurosisfugax in younger individuals.

From January 1, 2004, to December 31, 2012, a total of 685 individuals with ischemic stroke, transient ischemic attack, or amaurosisfugax were enrolled at Aarhus University Hospital in Denmark for thrombophilia study. The patients' ages ranged from 18 to 50 years. Medical records and the Danish Stroke Registry provided the clinical data. The laboratory information system yielded the findings of the thrombophilia examination. Ischemic stroke (N = 377) and transient ischemic attack (TIA) or amaurosisfugax (N = 308) both had absolute thrombophilia prevalences and risk ratios (OR) with 95% CI. Publicly available statistics were used to determine the prevalence of thrombophilia in the whole population.

This work introduces a prototype for classifying stroke using a combination of text mining tools and machine learning techniques; the goal is to categorise the illness using these approaches. With the use of appropriately trained machine learning algorithms, machine learning may be positioned as a major tracker in domains like as healthcare, data management, and surveillance. The data mining approaches used in this study provide a comprehensive overview of information tracking from both a semantic and a syntactic standpoint. The plan is to use the case sheets as a source of data for symptom mining and then use it to train the system. Sugam Multispecialty Hospital in Kumbakonam, Tamil Nadu, India, provided 507 patient case sheets for the data gathering phase. The next step was to use tagging and maximum entropy approaches to mine the case sheets. The suggested stemmer then uses these datasets to categorise the strokes by extracting features that are common and distinctive. Following that, a number of machine learning techniques, including ANNs, SVMs, boosting and bagging, and random forests, were used to the processed data. using a better classification accuracy of 95% and a reduced standard deviation of 14.69, artificial neural networks trained using a stochastic gradient descent approach fared better than the other techniques.

Predicting strokes using svm: Recognising strokes early is crucial for prompt prevention and treatment. Results from studies demonstrate that metrics derived from different risk factors provide useful information for stroke prognosis. This study delves into the several physiological markers that are used as stroke risk factors. The International Stroke Trial database was used for data collection, and Support Vector Machine (SVM) was trained and tested with success. Here, we used support vector machines (SVMs) with various kernel functions and discovered that a linear kernel achieved a 90% success rate.

A database of worldwide stroke trials

One of the biggest randomised studies ever performed on individual individuals with acute stroke is the International Stroke Trial (IST). Data on 19,435 individuals who had an acute stroke and had 99% full follow-up are included in the IST dataset. Participants' ages ranged from 80 and above for more than 26.4% of the total. There was a lack of thrombolytic treatment and inadequate background stroke care. A pilot phase ran from 1991 to 1993 as part of this clinical investigation, which ran the full gamut from 1991 to 1996. There is full data at the baseline and more than 99% at the follow-up in this big prospective randomised controlled trial. An analyzable database was given with data retrieved for each randomised patient on the factors examined at randomization, as well as at two further time points: one at six months and another fourteen days following randomization or before discharge. The purpose of the study was to determine whether acute ischemic stroke patients' clinical outcomes were affected by the early administration of aspirin, heparin, or neither.

Classification-based analysis and prediction model for stroke disease

Today, data mining is crucial for the medical industry's illness prediction efforts. As a potentially fatal illness, stroke has risen to the position of third most common killer in both developed and developing nations. Stroke is a major contributor to the high rate of permanent disability in the United States. The intensity of the stroke determines how long it takes for the patient to make a full recovery. Many studies have compared the efficacy of predictive data mining in order to make illness predictions. Here, a variety of characteristics are employed to forecast the occurrence of stroke illness using classification techniques such as Neural Networks, Decision Trees, and Naive Bayes. To reduce the number of dimensions, our study use a principal component analysis technique, which then finds the traits that contribute most to the prediction of stroke illness and can tell you whether a patient has the condition or not.

Existing System

Cell death due to inadequate blood supply to the brain is the hallmark of a medical illness known as a stroke. It has recently surpassed all other causes of mortality on a global scale. Upon examining the people who have suffered from stroke, many risk variables that are thought to be associated with its causation have been discovered. Numerous studies have attempted to forecast and categories stroke illnesses based on these risk variables. Data mining and ML algorithms provide the backbone of the majority of the models. In this study, we used four ML algorithms to identify the most likely or actual kind of stroke based on a patient's current health status and medical history. In order to address this issue, we have compiled a large number of submissions from healthcare facilities. A real-time medical report may benefit from the results, as shown by the categorization result, which is good. Current system algorithms include Naive Bayes, J48, KNN, and Random Forest.

Proposed System

A disruption in blood flow to an area of the brain is the leading cause of brain tissue damage known as a stroke. Consequently, the patient's quality of life may be diminished due to a reduction in some functions associated with the afflicted area. In this study, we address the issue of stroke identification in CT scans by using CNNs that have been optimised using Particle Swarm optimisation (PSO). We took into account both ischemic and hemorrhagic strokes, and we made our dataset public to encourage studies into brain detection of both conditions. For every instance in the dataset, there are three distinct picture kinds: the original CT scan, an image with the skull segmented, and a third image with the radio logical density map. Finding encouraging findings, the results demonstrated that CNNs are well-suited to handling stroke detection.

METHODOLOGY AND IMPLEMENTATION

The following modules have been developed to carry out this project:

The first step is to upload the stroke dataset to the programme using this module.

Dataset Preprocessing & Features Selection: In this module, we will clean the dataset by replacing missing values with 0. Then, we will use a label encoding algorithm to convert non-numeric values to numeric values. After that, we will select features from the dataset. Finally, we will split the dataset into two parts: train and test. The application will use 80% of the data for training and 20% for testing.

Thirdly, train the Naïve Bayes algorithm: the data mentioned earlier will be used to train the model, and then it will be tested on test data to determine its correctness.

The fourth step is to train the J48 method, which will take the data from the training set and use it to create a model. Then, using the test set, we will determine the model's correctness.

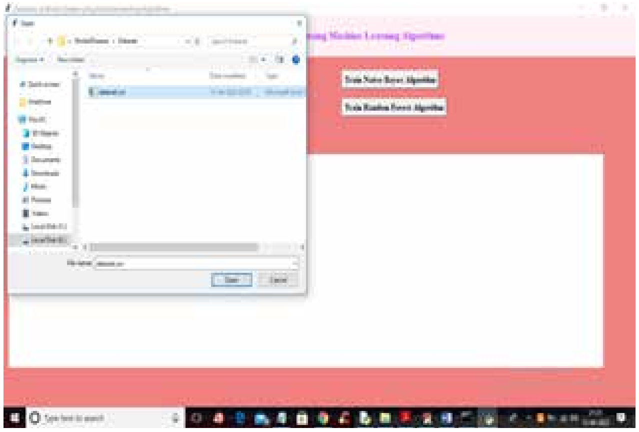
Fifthly, train the KNN algorithm: this involves feeding the aforementioned training data into the algorithm to create a model, and then applying the model to test data in order to determine its accuracy.

The sixth step is to train a model using the Random Forest technique. This model will be fed the training data from step five and then tested on test data to determine its correctness.

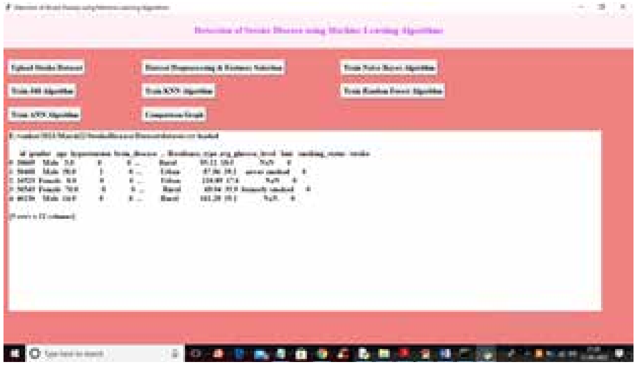
The seventh step is to train an ANN algorithm, which will take the data from step one and use it to build a model. Then, using the model on test data, we can determine how accurate it was.

Comparison Graph: We will use this module to create a graph that compares the algorithms' accuracy.

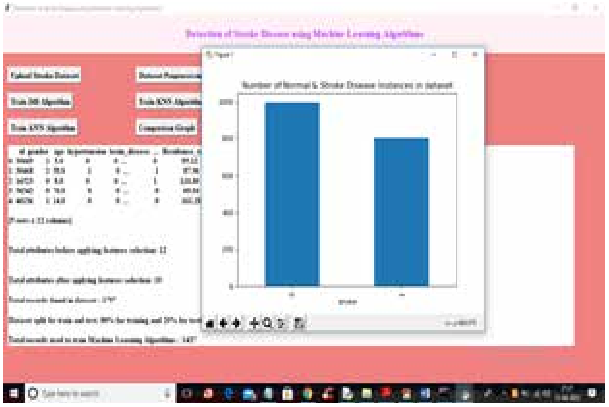
OPERATION



To load the dataset and obtain the output below, pick and upload the dataset.csv file on the top screen, then click the "Open" button.

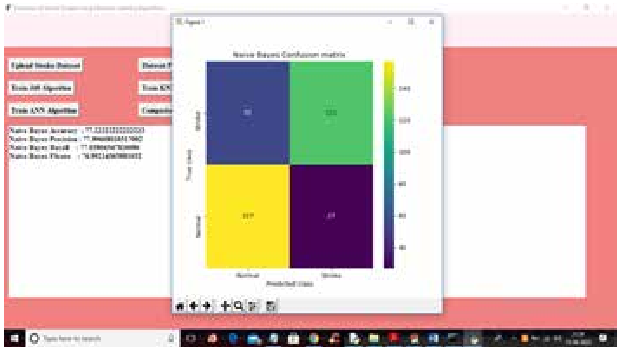

Detection of Stroke Disease using Machine Learning Algorithms
Upload Stroke Dataset
Dataset Preprocessing & Features Selection
Train Naive Bayes Algorithm
Train J48 Algorithm
Train KNN Algorithm
Train Random Forest Algorithm
Train ANN Algorithm
Comparison Graph

The dataset is imported on the screen above, and it has a lot of missing and non-numeric data. To process the dataset and obtain the output below, click the “Dataset Preprocessing& Features Selection” button.

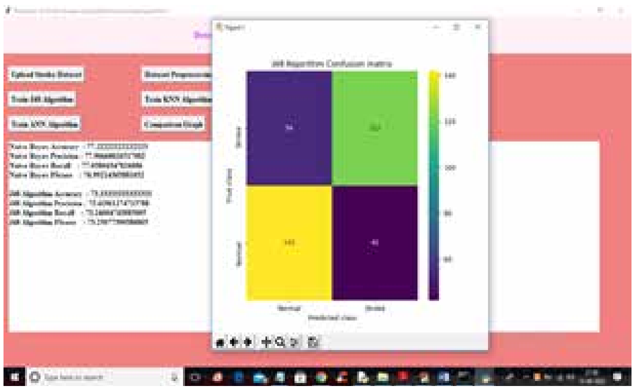


In above graph x-axis represents 0 (normal) and 1 (stroke) and y-axis represents number of instances available in those categories in dataset and now close above graph and see below screen.

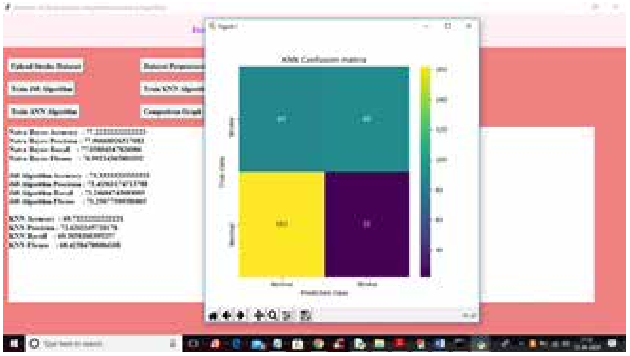
The entire dataset has been transformed to numeric format on the screen above, and it has been divided into train and test datasets. Click the “Train Naïve Bayes Algorithm” button to train Naïve Bayes on the above dataset and obtain the output below.



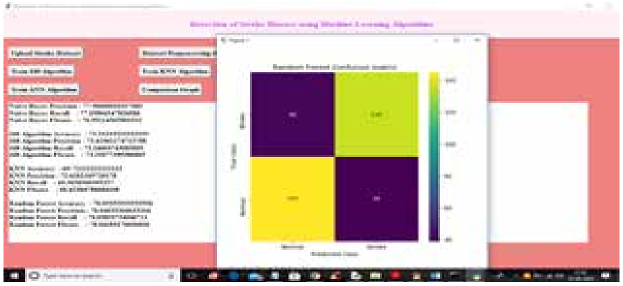
In above screen with Naïve Bayes we got 77% accuracy and in confusion matrix graph we can see number of correct and incorrect prediction by Naïve Bayes. Now click on ‘Train J48 Algorithm’ button to get below output.



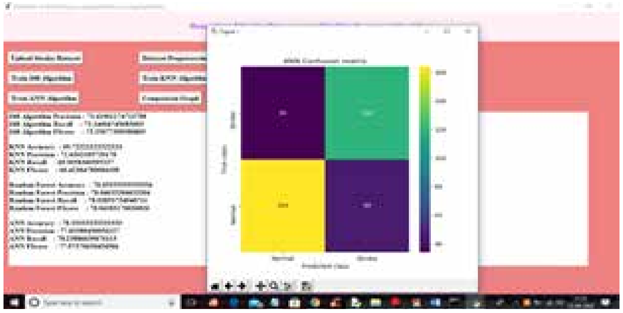
The top screen shows our 73% accuracy rate with J48, and the confusion matrix graph shows the amount of right and wrong predictions made by J48.Now, close the preceding graph, and then select “Run KNN Algorithm” to obtain the output shown below.



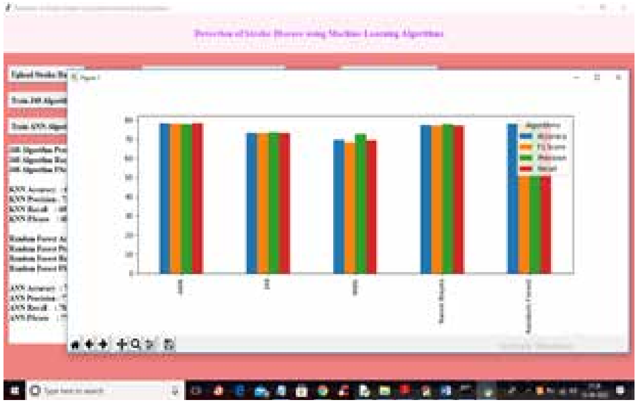
The above screen shows our 69% accuracy with KNN, while the confusion matrix graph shows the amount of right and wrong predictions made by KNN. Now, close the preceding graph, and then select “Run Random Forest Algorithm” to obtain the output shown below.



We obtained 78% accuracy with Random Forest on the screen above, and the confusion matrix graph shows the number of right and wrong predictions made by Random Forest. Now, close the above graph, and then select "Run ANN Algorithm" to obtain the output below.



The ANN achieved 78.33% accuracy in the screen above. The confusion matrix graph shows the amount of accurate and inaccurate predictions made by the ANN, and the method as a whole demonstrated the high accuracy of the ANN. Now shut the graph above, and to view the graph below, click the "Comparison Graph" button.



The x-axis in the graph above denotes the names of the algorithms, and the y-axis shows accuracy along with additional metrics like precision, recall, etc. Various color bars correspond to distinct metrics, and ANN achieved great accuracy across all techniques.

CONCLUSION

Neurologists may get assistance from ANN when it comes to identifying strokes from CT head scan images. Acquired data for training dataset also affects the achieved accuracy. In this study, we tested N photos of each stroke type and found that our suggested strategy achieved an accuracy of 78.33%. The categorization result is significantly influenced by the quantity of photographs utilized in the training procedure. When more images are used in the training process, accuracy increases. Here, every colored bar denotes a distinct statistic, and ANN performed admirably in every instance.

The study may be expanded upon in future work by including other categorization algorithms. Also, by supplementing the current dataset with certain non-stroke data, stroke prediction may be achieved.

REFERENCES

1. S. H. Pahus, A. T. Hansen, and A.-M. Hvas, "Thrombophilia testing in young patients with ischemic stroke," Thrombosis research, vol. 137, pp. 108–112, 2016.

2. P. Govindarajan, R. K. Soundarapandian, A. H. Gandomi, R. Patan, P. Jayaraman, and R. Manikandan, "Classification of stroke disease using machine learning algorithms," Neural Computing and Applications, pp. 1–12.

3. L. T. Kohn, J. Corrigan, M. S. Donaldson, et al., To err is human: building a safer health system, vol. 6. National academy press Washington, DC, 2000.

4. R. Jeena and S. Kumar, "Stroke prediction using svm," in 2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), pp. 600–602, IEEE, 2016.

5. P. A. Sandercock, M. Niewada, and A. Członkowska, "The international stroke trial database," Trials, vol. 13, no. 1, pp. 1–1, 2012.

6. M. S. Singh and P. Choudhary, "Stroke prediction using artificial intelligence," in 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON), pp. 158–161, IEEE, 2017.

7. S. Y. Adam, A. Yousif, and M. B. Bashir, "Classification of ischemic stroke using machine learning algorithms," Int J ComputAppl, vol. 149, no. 10, pp. 26–31, 2016.

8. A. Sudha, P. Gayathri, and N. Jaisankar, "Effective analysis and predictive model of stroke disease using

classification methods," International Journal of Computer Applications, vol. 43, no. 14, pp. 26– 31, 2012.

9. G. Kaur and A. Chhabra, "Improved j48 classification algorithm for the prediction of diabetes," International Journal of Computer Applications, vol. 98, no. 22, 2014.

10. I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

11. P. Sewaiwar and K. K. Verma, "Comparative study of various decision tree classification algorithm using weka," International Journal of Emerging Research in Management &Technology, vol. 4, pp. 2278– 9359, 2015.

DFR TSD A Deep Learning Based Framework for Robust Traffic Sign Detection Under Challenging Weather Conditions

Gangaboina Madhusudhan, Yelmareddy Indu
Mukka Praneeth Raj
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

K Srujan Raju
Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ ksrujanraju@gmail.com

ABSTRACT

A thorough understanding of autonomous truck technology relies heavily on long-term traffic signal detection and recognition (TSDR). Numerous interesting ways have been suggested in the current literature, and a great amount of study has been conducted due to the importance of this task. However, most of these methods have only been tested on datasets that are clean and devoid of challenges. They haven't taken into account the performance degradation caused by various CCs that obstruct real-world traffic-sign photos. We examine the TSDR issue under CCs and zero in on the performance degradation that comes along with them in this research. For this reason, we advise using a TSDR architecture that is built on a Convolutional Semantic network (CNN) for prior enhancement. Among the components of our modular approach are a convolutional neural network (CNN) difficulty classifier, an encoder-decoder CNN style called Enhance-Net, and two distinct CNN styles called sign-discovery and category. In order to improve the web traffic indicator regions (rather than the complete picture) in challenging photos that are accurately detected, we propose a new training process for Enhance-Net. To evaluate the efficacy of our approach, we used the CURE-TSD dataset, which includes traffic video clips captured under different CCs.

KEYWORDS: *CNN, TSD, TSDR, CC, Traffic video.*

INTRODUCTION

Driver assistance systems and driverless vehicle technology rely on web traffic indicator detection and recognition. Ensuring the safe and widespread use of this technology relies on a TSDR algorithm that can withstand many real-world challenges with reliability and durability [1]. But, in addition to the wide range of website traffic indicators to detect, images of traffic captured in the field are often obscured by various unfavourable weather conditions and other forms of activity, and they are far from perfect [2].

LITERATURE SURVEY

Points of view and survey on vision-based traffic indication detection and analysis for smart driver assistance systems T. B. Moeslund, A. Mogelmose, and M. M. Trivedi wrote the article.

online traffic signal recognition (TSR) discovery systems for driver assistance vehicles are detailed in this article, which also provides a literature review on online traffic signal detection. We break down the many steps involved in discovering website traffic signs into their component parts and explain how current tasks contribute to each one: division, function extraction, and last indication identification. While TSR has been around for a while, there are still some unanswered questions about the literature that we want to address. These include the fact that there isn't enough use of publicly accessible imagery datasets and that European website traffic signs are over-represented. In addition, we go review several potential future paths for TSR research that include integrating context and localization. On top of that, we unveil a fresh public database including US website traffic indicators.

[2] An assessment of the methods for the autonomous detection and recognition of website traffic indicators, U. Raghavendra, A. Gudigar, and S. Chokkadi wrote it.

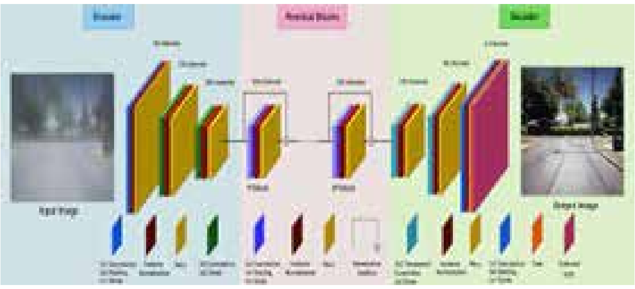
It would seem that ITS has made significant progress in every way. In order to meet the needs of motorists for both safety and convenience, ITS rely on the detection and identification of traffic indicators. Within the framework of a vision-based vehicle driver assistance system, this study provides a critical analysis of three major actions in the Automatic Traffic Indication Discovery and Acknowledgment (ATSDR) system: segmentation, detection, and identification. Also, several experimental setups of the photo capture system are the main focus. In addition, potential future study hurdles are discussed in order to make ATSDR much more successful. This opens up a lot of options for the researchers to conduct a full examination of ATSDR and include future features into their study.

EXISTING SYSTEM

Most people think of website traffic indicator detection as a challenge with picture segmentation and recognition. Using the CURE-TSD dataset, the authors of SegU-Net [2] have published a benchmark result. Despite their promising results on the challenge-free dataset, they show that merging the challenging dataset significantly reduces performance. The reason for this is because the picture enemies obscure the finer characteristics of small website traffic sign locations, both in terms of colour and shape. For practical purposes, a comprehensive approach that can withstand the negative effects of these CCs is crucial [6]. There are a lot of nice CNN-based approaches to website traffic indicator detection and categorization that have been suggested in the literature, but none of them account for CCs. Consequently, these methods' efficacy is highly dependent on the clarity of the captured image.

PROPOSED SYSTEM

To mitigate this problem, we provide Enhance-Net, a deep convolutional neural network (CNN) picture enhancer that pre-processes these traffic photos. Temel et al. [6] noted that the twelve different types of CCs in the dataset do not have to be resolved simultaneously and also brought attention to the fact that benchmark formulae are more vulnerable under difficult weather conditions like Rain, Snow, and Haze. Furthermore, two CCs that occur often in real-life circumstances are lens blur and dirty lens. Thus, we emphasise five distinct CCs—Rain, Snow, Haze, Dirty lens, and Lens blur—related to five different degrees of magnitude in this study. Training a single network to overcome all of these challenges might result in less-than-ideal performance because to the diverse nature of these hurdles. Therefore, to guarantee the highest potential improvement for each CC, we use five separate Enhance-Nets trained on a single kind of problem at a time. We use the modular TSDR approach proposed in [5] to classify and identify signs in the enhanced website traffic images. There are four distinct parts to our method: the obstacle classifier, the improvement blocks, the sign detector, and the indicator classifier.



METHODOLOGY

There are a lot of algorithms out there that can detect website traffic indicators, but they're all trained on perfect images and only use those for discovery. Unfortunately, the quality of outdoor photos can sometimes degrade over time, making it impossible for existing algorithms to detect signs in photos affected by bad weather. In order to avoid unnecessary complications, the author of this research has presented a CNN algorithm that comprises two parts. [5] The first section will clean up any weather-related debris, and the second will undoubtedly identify signs of online usage. [6]

The DFR-TSD method is a suggestion system that uses a Deep Learning CNN formula to find traffic signals effectively. After being trained on the 'CURE-TSD' dataset, the proposed formula achieves a precision of up to 99%. [7]

In order to carry out this project, we have really assembled the following parts:

Make a Tonne of Website Visitors Sign Up Convolutional Neural Network (CNN) Model: We will generate and load website traffic indicator using this component. CNN model

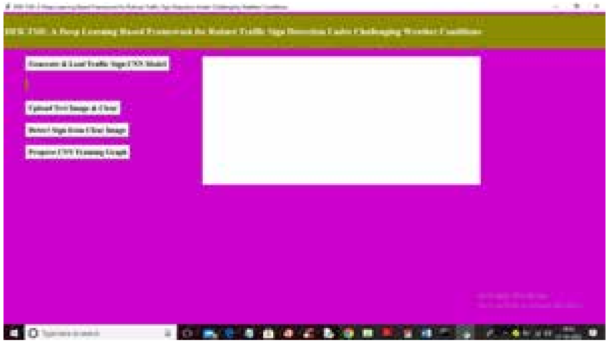
Second, we'll utilise this module to submit an inspection image and then apply convolutional neural network (CNN) design to remove climate-related images.

Thirdly, Locate Indicator in Clear picture: The clear picture will now be sent into the CNN sign recognition algorithm to locate the indicator.

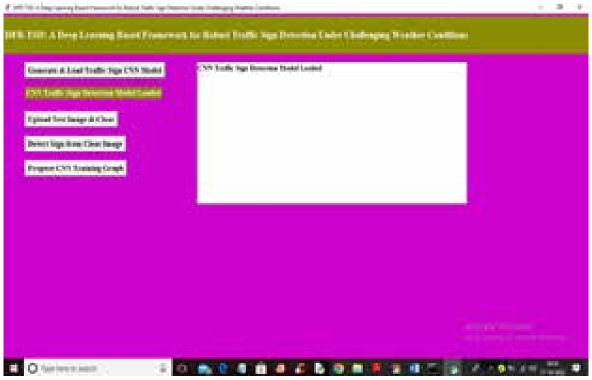
The Suggest Convolutional Neural Network Training Graph is where we'll make sure to plot the CNN training loss and precision graphs.

SCREEN SHOTS

To run project double click on 'run.bat' file to get below screen

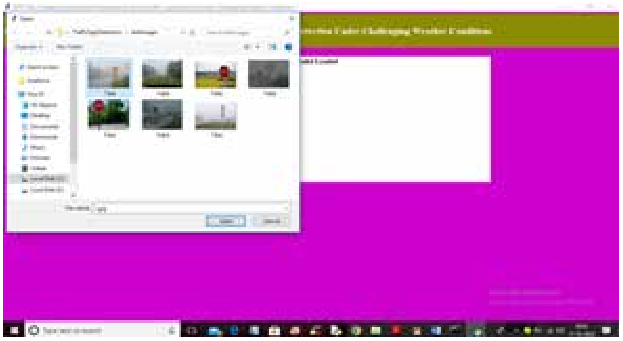

DFR TSD: A Deep Learning Based Framework for Robust Traffic Sign Detection Under Challenging Weather Conditions
Generate & Load Traffic Sign CNN Model
Upload Test Image & Clear
Detect Sign from Clear Image
Prepare CNN Training Graph

From the previous screen, you may produce and load a CNN model by clicking the "Generate & Tonnes Website Traffic Sign CNN Model" button. This will lead to the next screen.

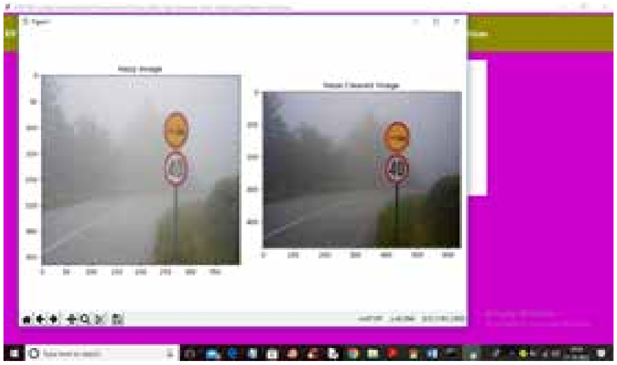

DFR TSD: A Deep Learning Based Framework for Robust Traffic Sign Detection Under Challenging Weather Conditions
Generate & Load Traffic Sign CNN Model
CNN Traffic Sign Detection Model Loaded
Upload Test Image & Clear
Detect Sign from Clear Image
Prepare CNN Training Graph

CNN Traffic Sign Detection Model Loaded

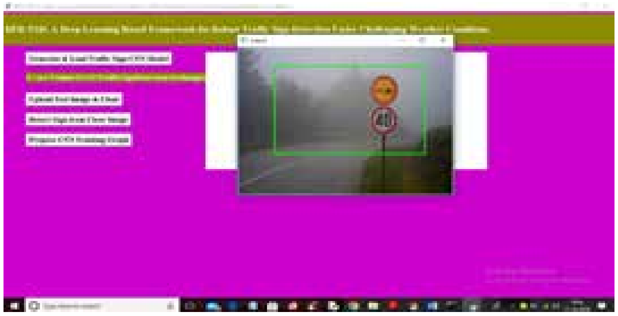
To submit a test picture affected by weather conditions and subsequently clean it up, click the "Upload Test Image & Clear" option on the previous screen once the CNN model has loaded.



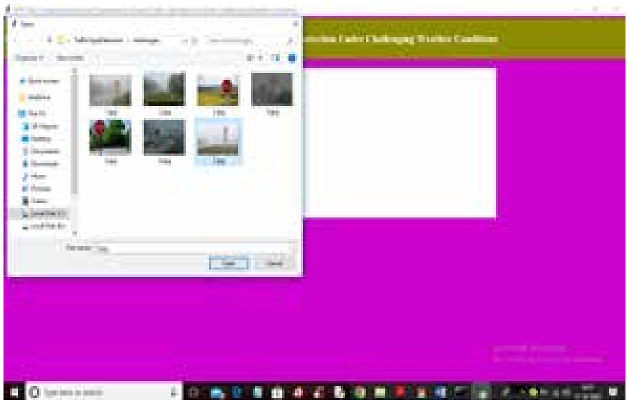
After selecting and publishing the "1. jpg" file in the previous screen, you may load and clean the picture by clicking the "Open" button. The result will be shown below.



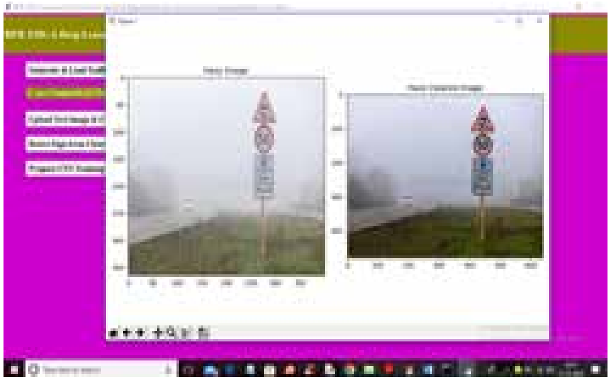
Above, you can see the difference between the first photo—a weather-influenced one (caused by things like overcast haze, rain, bad lighting, or a bad camera lens) and the second—a clean one. To obtain the output listed below, click the "Discover Indication from Clear Photo" button.



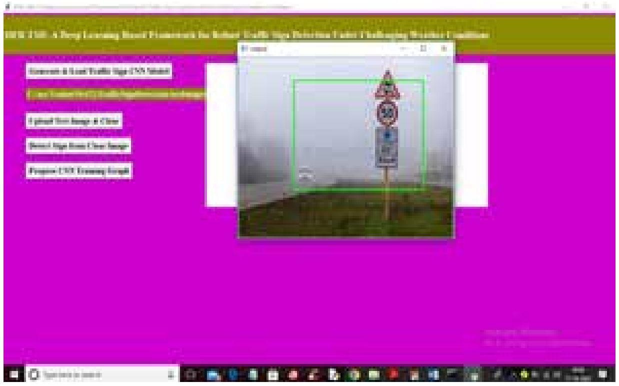
After detecting the traffic sign and drawing a bounding box around it, the CNN model moved on to testing additional images.



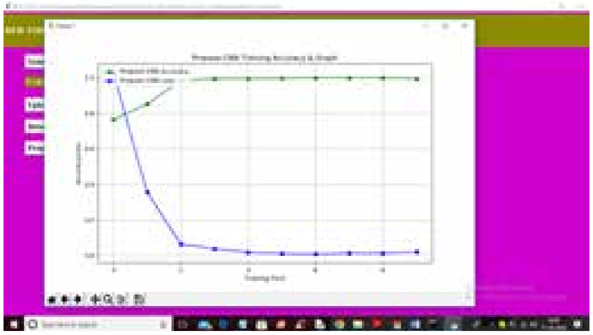
After choosing and uploading 7.jpg on the previous screen, click on the "Open" button to get the following output:



Click the "Spot Indicator from Clear Image" button to get the output; the first image is the weather-impacted one and the second is the clean one.



The following result will be shown when you click the "Propose CNN Training Graph" button; the previous screen indicated online traffic indications, and you may post and see more images in the same manner.



The figure above shows the relationship between the variables "Training Date" (x-axis) and "Precision and loss values" (y-axis), with the green line representing accuracy and the blue line representing loss. As we can see in the graph above, the accuracy and loss both increased with each passing day of boosting.

Please be aware that no detection technique is 100% accurate, and that some bounding boxes may not always predict with 100% accuracy.

CONCLUSION

We have presented a robust framework for TSDR under different CCs that is modular and based on deep convolutional neural networks (CNNs) in this research. We have shown how the presence of different CCs destroys the efficiency of the current TSDR formulae and proposed a deep CNN-based solution that fixes the problem. An obstacle classifier based on VGG16 architecture discovers and labels the problem, then sends the image to the correct Enhance-Net, which restores the functions needed for accurate identification of the regions indicated by web traffic. The Enhance-Nets are trained using our suggested unique loss function and training process, which includes web traffic indication area focused MAE in both pixel and attribute domain with the indication detection loss as a limitation. This sets them apart from the previous complete image enhancement-based techniques. The improvement of the indication zones based on their exact detection is effectively ensured by this. Additionally, we have shown experimentally that zones including traffic signs are more critical for improvement in order to get better detection performance. Finally, we conclude that our approach's modular structure is successful by comparing it to other end-to-end qualified deep CNN-based object recognition networks and finding that our technique is superior to all of them. Our modular design allows us to build each part of our structure separately.

REFERENCES

1. A. Mogelmose, M. M. Trivedi, and T. B. Moeslund, "Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey," IEEE Trans. Intell. Transp. Syst., vol. 13, no. 4, pp. 1484–1497, Dec. 2012.
2. A. Gudigar, S. Chokkadi, and U. Raghavendra, "A review on automatic detection and recognition of traffic sign," Multimedia Tools Appl., vol. 75, no. 1, pp. 333–364, Jan. 2016.
3. S. Xu, "Robust traffic sign shape recognition using geometric matching," IET Intell. Transp. Syst., vol. 3, no. 1, pp. 10–18, Mar. 2009.
4. J. F. Khan, S. M. A. Bhuiyan, and R. R. Adhami, "Image segmentation and shape analysis for road-sign detection," IEEE Trans. Intell. Transp. Syst., vol. 12, no. 1, pp. 83–96, Mar. 2011.
5. I. M. Creusen, R. G. J. Wijnhoven, E. Herbschleb, and P. H. N. de With, "Color exploitation in hog-based traffic sign detection," in Proc. IEEE Int. Conf. Image Process., Sep. 2010, pp. 2669–2672.
6. D. Temel, M.-H. Chen, and G. AlRegib, "Traffic sign detection under challenging conditions: A deeper look into performance variations and spectral characteristics," IEEE Trans. Intell. Transp. Syst., vol. 21, no. 9, pp. 3663–3673, Sep. 2020.
7. J. Canny, "A computational approach to edge detection," in Readings in Computer Vision. Amsterdam, The Netherlands: Elsevier, 1987, pp. 184–203.
8. X. Baró, S. Escalera, J. Vitrià, O. Pujol, and P. Radeva, "Traffic sign recognition using evolutionary AdaBoost detection and forest- ECOC classification," IEEE Trans. Intell. Transp. Syst., vol. 10, no. 1, pp. 113–126, Mar. 2009.
9. P. G. Jiménez, S. M. Bascón, H. G. Moreno, S. L. Arroyo, and F. L. Ferreras, "Traffic sign shape classification and localization based on the normalized FFT of the signature of blobs and 2D homographies," Signal Process., vol. 88, no. 12, pp. 2943–2955, Dec. 2008.
10. F. Zaklouta and B. Stanciulescu, "Real-time traffic sign recognition in three stages," Robot. Auto. Syst., vol. 62, no. 1, pp. 16–24, Jan. 2014.
11. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1097–1105.
12. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, arXiv:1412.7062. [Online]. Available: http://arxiv.org/abs/1412.7062

Smart Voting System through Facial Recognition

Mamidi Poojitha, Maddi Gowtham Reddy, Kamble Pallavi
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

Jonnadula Narasimharao
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ jonnadula.narasimharao@gmail.com

ABSTRACT

The political election is a potentially pivotal event in every modern democracy, but a major concern for democracies is the widespread mistrust of electoral systems among large segments of the global population. Concerns about the reliability, usability, security, integrity, and transparency of electronic voting systems are common. When it comes to this area, Estonia is in the forefront and may be considered cutting edge. Nevertheless, blockchain technology is still in its infancy and is used by a small number of firms. In addition to delivering advantages like immutability and decentralisation, blockchain technology can solve all of the above difficulties. Current blockchain-based e-voting solutions have a number of issues, the most important of which are their skilled approach or lack of comparison and testing. In this article, we present an electronic voting system that is built on the blockchain and can be used for any kind of poll. Blockchain makes full use of it, and all operations may be handled inside it. Once the voting process begins, the platform operates autonomously and decently, removing any opportunity to influence the outcome.

KEYWORDS: *Block chain, E voting, Secured voting.*

INTRODUCTION

Several nations have started using electronic ballot methods recently. When Estonia held its national elections using an electronic voting system, it was the world's first. Following this, digital voting was adopted by Norway for their council elections and by Switzerland for their state-wide elections [1]. In order to compete with traditional ballot systems, electronic ballots must meet all of the same standards, including those for security and anonymity [2]. In order to guarantee that an e-Voting system is accessible to people and safe against external influences that might alter votes or prevent voters' ballots from being tampered with, it must have better security. Voters' anonymity is ensured by several digital voting systems with the use of Tor [3]. But because several intelligence services throughout the globe control different portions of the Internet that might allow them to detect or intercept votes, this approach does not provide overall privacy or integrity. Voters will not have to worry about someone misusing the selection process because of our determination to create safe political elections [4]. Recently, blockchain has been used as an example of a protected invention that is used in an online setting. To oversee all aspects of political elections, our electronic voting system employs blockchain technology. The main benefit is that trust in the central authority that conducted the elections is not necessary. Because of our system, this power cannot influence the election. Another issue with electronic voting is that voters are confused since the system is not transparent about how it works [5]. For this problem, blockchain technology provides an entirely transparent solution by making all recorded information and operations, including the specifics of how this data is handled, visible to all users. This state-of-the-art technology outshines the age-old electronic voting platform devoid of blockchain in every respect when it comes to security [6].

RELATED STUDY

Block-chain-Enabled E-Voting, There may be fewer voter frauds and more voter access with blockchainenabled e-voting (BEV). Voters who are eligible to do so use a computer or smartphone to cast their ballots in an anonymous manner. The BEV system employs a private encryption key and tamper-proof identification cards. This article provides an overview of many BEV implementations, discussing the potential benefits and drawbacks of the method.

"Ballot Refine with Block-chain Technology: Auditable Block-chain Voting System," When it comes to digital voting, various countries use different approaches. Every one of them has its own set of pros and cons. Inadequate techniques of system verification and accounting is one of the most prevalent and serious problems. This problem has an answer in blockchain technology, which has been getting a lot of attention as of late. Electronic voting procedures and components of an audit-and confirmation-qualified online voting system are defined in this article as the Auditable Blockchain Ballot System (ABVS). ABVS is able to do this by using both blockchain technology and a paper audit trail that has been confirmed by voters.

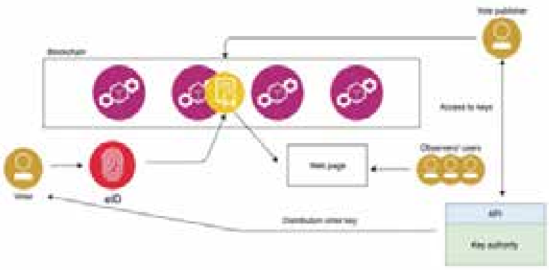
Bitcoin: A Peer-to-Peer Electronic Cash System: It might be possible to bypass banks entirely with a peer-to-peer version of electronic currency, allowing for instantaneous online payments. While digital trademarks can help, they won't solve the problem entirely if a trusted third party is still needed to avoid duplication of expenditure. We provide a solution to the problem of double spending that makes use of a decentralised network of peers. The network verifies the purchase time by adding it to a distributed ledger of hash-based proof-of-work, creating an immutable record that can only be updated by updating the proof-of-work itself. The longest chain is evidence of the observed sequence of events and also of its origin from the largest pool of CPU power. They will generate the longest chain and outperform opponents as long as most of their CPU power is controlled by nodes that aren't collaborating to attack the network. Minimal framework is required by the network itself. Assuming the longest proof-of-work chain as proof of what happened while they were gone, nodes may join and quit the network at whim, and messages are sent on a best-effort basis.

Current Setup

One example of a safe technology utilised in an online environment is blockchain, which has become more prominent in recent years. All election operations are managed by our e-voting system, which employs blockchain technology. Having no need to have faith in the governing body that created the voting system is its main advantage [5, 6]. Our structure does not allow this power to influence the outcomes of political elections. Lack of openness on the system's effectiveness is another issue with e-voting that causes people to lose faith in it [7].

PROPOSED SYSTEM

Any kind of political election, such as those for the position of head of state or trainee parliament, may benefit from the suggested blockchain ballot method since it takes into account all voting requirements. Using a public blockchain, the technology allows for much more extensive political elections. Any user may instantly verify the stored information (ballots), but other blockchains can alter the public blockchain. This person stands in for everyone who cares about the blockchain voting process. We identify three primary roles in our proposed system: ballot publisher, critical authority, and voter. Any one of these three people might stand in for a business, a client, or an organisation. The functions of ballot publisher and essential authority may be consolidated into one position, since they can be fulfilled by the same entity. The person takes part in the election by using a ballot system. The ballot author executes the arrangement, which is part of the wise agreement. Before publishing the smart agreement, the vote publisher must be able to recognise any cypher tactics. It is crucial that the vote author and the key authority work closely together. A voter and vote author get all cypher secrets from the important authority, which produces and distributes them. No unauthorised entity should be able to access the distribution route, so it must be protected.



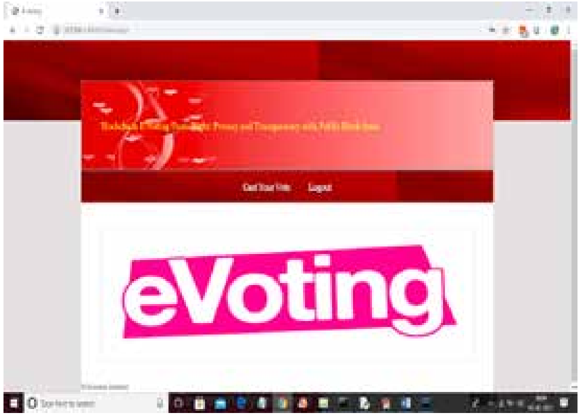
METHODOLOGY

Since Blockchain provides secure and tamper-proof data storage, we are using public python Blockchain APIs to store and manage ballot information in our project. To carry it out, we have really developed components that adhere to these standards.

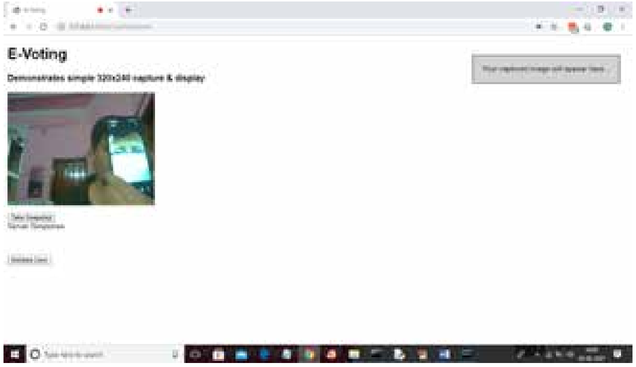
The admin section allows users to see party data and the vote tally, as well as upload new candidates and party information. Enter "admin" as the username and "admin" as the password to access the system as an administrator.

The user module requires the user to register with the app by entering their username and password, and then they must submit a photo of their face taken by a camera. After completing the registration process, users may verify their identity by going to the login page. Once they've successfully logged in, they'll be able to cast their vote according to their capabilities.

RESULTS EXPLANATION



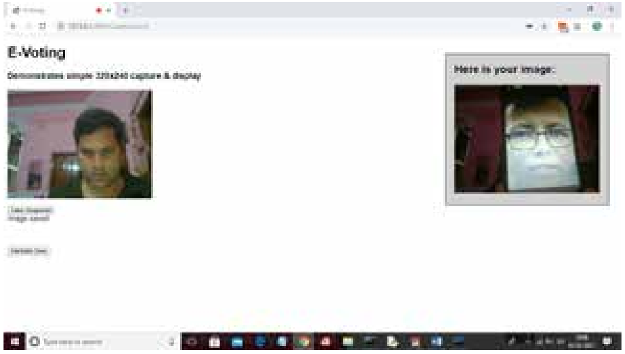
In above screen user can click on 'Cast Your Vote' link to get below webcam screen.



E-Voting

Demonstrates simple 320x240 capture & display

The camera is active on the above screen; to take a snapshot of the person's face, we need to display their face and then click the "Take Snapshot" button.

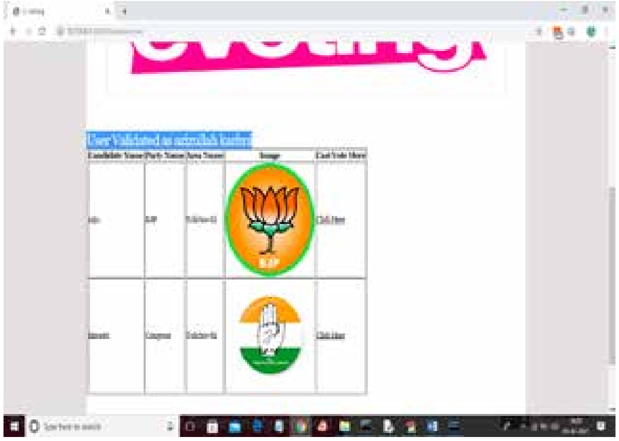


E-Voting

Demonstrates simple 320x240 capture & display

Here is your image:

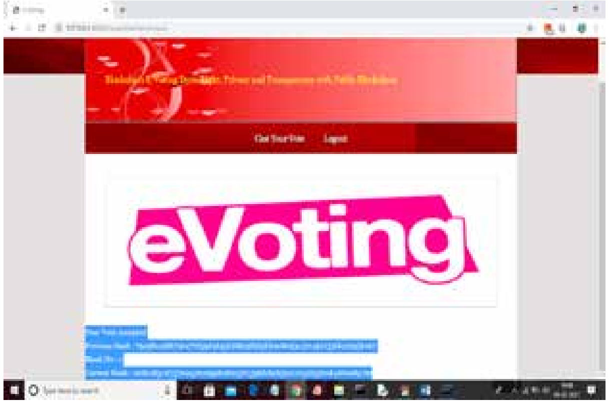
In above screen person face is capture and now click on 'Validate User' button to validate user.



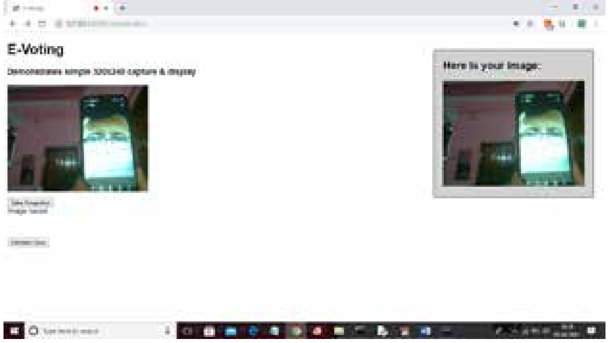
User Validated as azizullah karimi

Candidate Name	Party Name	Area Name	Image	Cast Vote Here
[illegible]	BJP	[illegible]		Click Here
[illegible]	Congress	[illegible]		Click Here

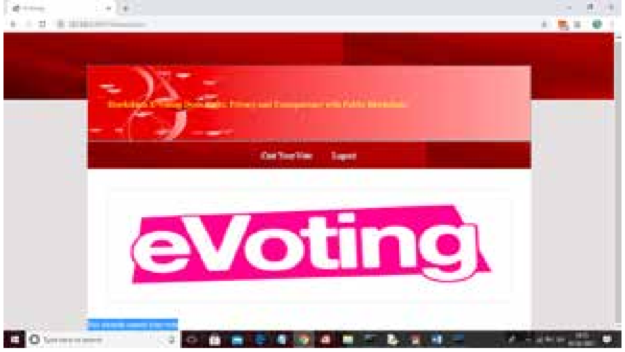
The user's identity, "azizullahkarimi," is shown in the blue screen above. A selection of candidates is then shown. The person may cast their vote by clicking on the "Click on this link" option, which takes them to the next page.



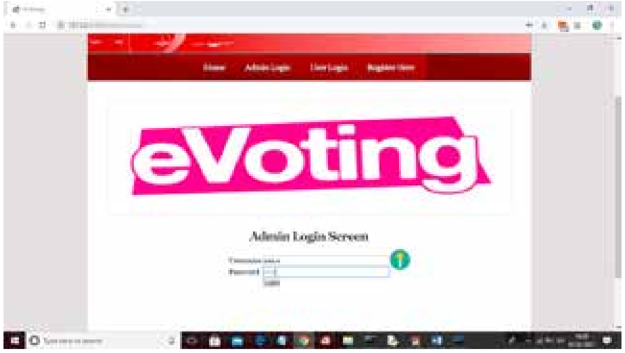
Because this is the first vote, the block will be added to the Blockchain with the number 1 in the display above. As we can see, the Blockchain has generated a chain of blocks with validation of both the previous and current hash codes. Try again using the same person to vote this time.



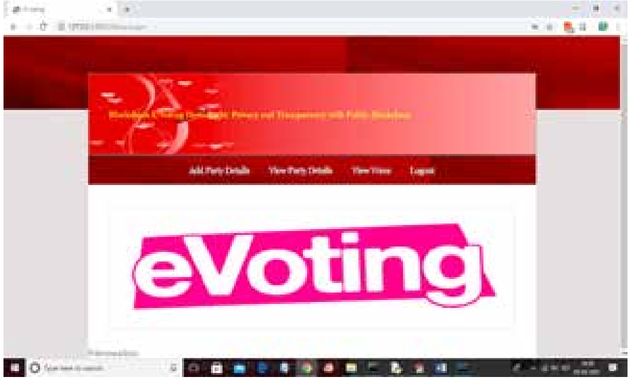
In above screen same user trying again and below is the result.



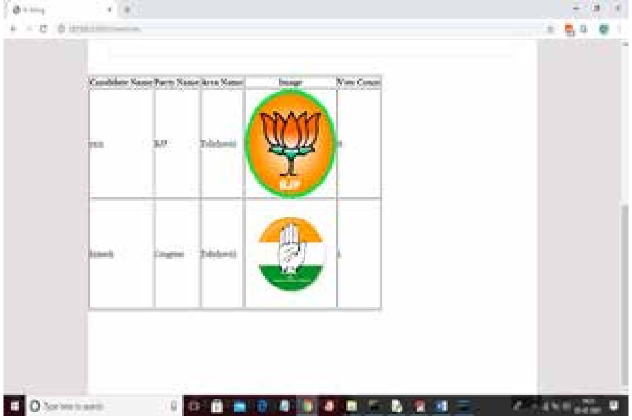
Repeated attempts by the same user will result in the following message: "You currently casted you elect." To see the ballot total, the user must logout and log in as "admin."



In above screen login as admin and after login will get below screen.



In above screen admin can click on 'View Votes' link to get below screen.



Candidate Name	Party Name	Area Name	Image	Vote Count
[illegible]	BJP	[illegible]		[illegible]
[illegible]	Congress	[illegible]		[illegible]

CONCLUSION

Public blockchain offers more advantages in this kind of voting system due to its transparent data and the fact that everyone may see them in real time, even when there are small variances in network delays. While an exclusive blockchain may be somewhat faster, the system's stability suffers as a result of its partial streamlining, since it only operates where the authority wants it to. A single voice over typically takes 6.32 seconds (mean 6.34 seconds), Hyperledger Composer 6.05 seconds (average 6.04 seconds), and Ethereum Ropsten 17.75 seconds (median 17.93 seconds), according to the data. Both the block time and the consensus method in use impact these durations.

REFERENCES

1. Ahmed Ben Ayed,"A Conceptual Secure Block Chain-Based Electronic Voting System",2017 IEEE International Journal of network &Its Applications(IJNSA),03 May 2017.

2. Rifa Hanifatunnisa, Budi Rahardjo," Blockchain Based E-Voting Recording System Design",IEEE 2017.

3. Kejiao Li, Hui Li,Hanxu Hou, Kedan Li,Yongle Chen," Proof of Vote: A High-Performance Consensus Protocol Based on Vote Mechanism & Consortium Blockchain", 2017 IEEE 19th International Conference on High Performance Computing and Communications; IEEE 15th International Conference on Smart City; IEEE 3rd International Conference on Data Science and Systems.

4. Ali Kaan Koç, Emre Yavuz, Umut Can Çabuk, Gökhan Dalkilic," Towards Secure E-Voting Using Ethereum Blockchain",2018 IEEE.

5. Supriya Thakur Aras, Vrushali Kulkarni," Blockchain and Its Applications – A Detailed Survey", International Journal of Computer Applications (0975 – 8887) Volume 180 – No.3, December 2017.

6. Freya Sheer Hardwick, Apostolos Gioulis, Raja Naeem Akram,Konstantinos Markantonakis," E-Voting with Blockchain: An E-Voting Protocol with Decentralisation and Voter Privacy",IEEE 2018,03 July 2018.

7. Kashif Mehboob Khan, Junaid Arshad, Muhammad Mubashir Khan," Secure Digital Voting System based on Blockchain Technology",IEEE 2017.

8. Huaiqing Wang, Kun Chen and Dongming Xu. 2016. A maturity model for blockchain adoption. Financial Innovation, Springer, Open Access, DOI 10.1186/s40854-016-0031-z

9. Buterin, Vitalik. 2015, On Public and Private Blockchains. [Online] https://blog.ethereum.org/2015/08/07/on-public-andprivate-blockchains/ [10] Zyskind et. al. 2015. Decentralizing Privacy: Using Block chain to Protect Personal Data, 2015 IEEE Securityand Privacy Workshops (SPW), San Jose, CA, USA, July 2015 [Online].Available: http://dx.doi.org/10.1109/SPW.2015

10. Jianliang Meng, Junwei Zhang,Haoquan Zhao, "Overview of the Speech Recognition Technology", 2012 Fourth International Conference on Computational and Information Sciences.

Driver Drowsiness Detection using Machine Learning

Kolli Swetha Priya, Rachha Dinesh
Ravula Suryateja
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

K Shilpa
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ shilpamtech555@gmail.com

ABSTRACT

Accidents involving sleepy drivers are on the rise; as it is well-known that fatigue and lack of concentration are major contributing factors in many accidents, this research is mainly focused on improving methods for detecting sleepiness. State of the driver in actual driving situations. Attempts to reduce these road accidents are the goal of chauffeur drowsiness detection systems. A variety of methods have been used to detect sleepiness or careless driving, and the secondary data collected is centred on prior studies on systems for detecting drowsiness. In the event of an accident, our programme will be able to identify the driver's drowsiness using the camera's captured picture, and we will be investigating how this data may be used to improve driving safety. one that ensures the safety of vehicles by reducing the likelihood of accidents caused by drowsy drivers. Gathering a human image from the camera and investigating how that data may be used to enhance driving safety is the primary objective. Identify the drowsy driver or not by collecting images from the live webcam feed and processing them using a machine learning algorithm. As soon as it detects that the driver is becoming tired, it will activate the buzzer alarm system and sound the alert. In the event that the chauffeur remains unconscious, they will inform their loved ones of their whereabouts by email and text message. Therefore, this energy is more effective than the challenge of detecting drowsiness while driving. Dlib for face and eye extractions.

KEYWORDS: *Eye detection, Face extraction, Driver drowsiness.*

INTRODUCTION

Modern times have seen a steady upturn in the automotive sector all around the globe. The subsequent explosion in vehicle ownership has led to a corresponding rise in traffic accidents worldwide [1]. Road accidents have shown to be a threat that has significantly diminished public and driver safety. According to the World Health Organization's Worldwide Standing Record on traffic Safety, three major factors that contribute to traffic accidents are inattention, alcoholism, and lack of sleep [2]. Households throughout the world face a significant risk due to the subsequent casualties and expenses. Due to their high cost and limited accessibility, the existing technologies used to detect sleepiness are not widely employed in most automobiles, especially those that are not premium models [3]. There is, therefore, an increasing need for a practical and intelligent technology for detecting sleepiness that the many vehicles in the industry can quickly adjust to. A number of novel advancements have been achieved in the field of artificial intelligence and AI; these advancements use a variety of algorithms to teach the version to be intelligent and capable of managing itself [4].

Inspiration

One of the leading causes of fatalities in traffic accidents is sluggish driving. People who drive long hours, particularly at night, or who operate buses that go cross-country or that run overnight are more likely to have this problem. Travellers in every nation face the cloudy headache of chauffeur tiredness [6]. Fatal and injury traffic accidents due to driver weariness occur often each year. Consequently, due to its enormous practical relevance, the identification of driver fatigue and related symptoms is a dynamic area of study [7]. Purchase, processing, and caution are the three main components of a typical sleepiness discovery system. Below, the purchasing system records the driver's frontal face on video, which is then sent to the processing block to detect sleepiness. The caution system will notify the driver via an alarm or warning system if it detects that they are becoming sleepy. [8]

Objective

With the use of a camera and an AI system called LBPH (Regional binary Pattern Histogram), we are able to detect whether a driver is drowsy in this role. This app will use the built-in webcam to read driver photos, then use the OPENCV LBPH formula to extract facial features from those photos. If the driver is blinking his eyes for 20 frames in a row or yawning, the app will send a sleepy message to the driver. If the range is better to sleepiness, the application will tell the driver; otherwise, we utilise the LBPH pre-trained drowsiness version and the Euclidean distance function to continually examine or anticipate the EYES and MOUTH distance better to drowsiness. Python and OpenCV were the tools I used for my project, which aimed to identify driving lanes. I applied the outcome of a processing pipeline I built to a video stream after it dealt with a number of sensitive photos.

LITERATURE SURVEY

A Smart Video-Based Drowsy Driver Detection System with Multiple Light Sources and Installed Applications

A smart video-based slow chauffeur detection system, unaffected by various lights, is created in this work. A driver's spectacles won't be a barrier for the proposed technology to identify drowsiness. The proposed method is divided into two separate computational procedures—the slow driver detection and the chauffeur eyes detection—by a near-infrared-ray (NIR) camera. Slow standing detection may achieve an accuracy of up to 91%, while the typical rates for open/closed eye detection without spectacles are 94% and with glasses 78%, respectively. After software optimisations, the processing speed with the 640 × 480 layout video may reach up to 16 frames per second (fps) when using an FPGA-based integrated platform.

"Motorist Fatigue Detection based on Eye Monitoring and Dynamic Layout Matching"

For safe driving, it is advised to use a system that can detect driver drowsiness in real-time using vision technology. Using the quality of skin tones, the driver's face may be found from shadow photos collected in automobiles and trucks. After that, the eye areas are located using edge discovery. In the next frame, the captured images of the eyes are not only used as dynamic templates for eye monitoring, but they are also utilised for fatigue detection, which in turn helps to develop warning alarm systems for road safety. Running the system test is a Pentium III 550 with 128 MB of RAM. It seems like the experiment turned out very well. On four test movies, the system achieved an average correct price for eye placement and tracking of 99.1 percent, and it could monitor 20 structures per second. While a rate of 100% is ideal for tiredness detection, the test videos show an average accuracy rate of 88.9%.

EXISTING SYSTEM

In every major city throughout the globe, traffic congestion is a serious problem right now. In the past, several approaches of collecting traffic data—like infrared light sensor units and induction loops—had been suggested, each with its own set of advantages and disadvantages. In recent times, photo processing has shown promising results in acquiring real-time traffic data from CCTV footage installed at traffic lights. The collection of online traffic statistics may be accomplished in a number of ways. While some of them take the time to tally the entire number of pixels [3], others calculate the number of automobiles [4-6]. In terms of collecting traffic data, several methods have shown promising outcomes. But if the extravehicular spacing is very small, it might not account for rickshaws or auto-rickshaws, the usual means of website traffic in South Asian nations, as lorries, and it might give wrong results when calculating the number of cars.

One of the disadvantages is the traffic bottleneck. In this case, we can detect online activity by use of an infrared light sensor.

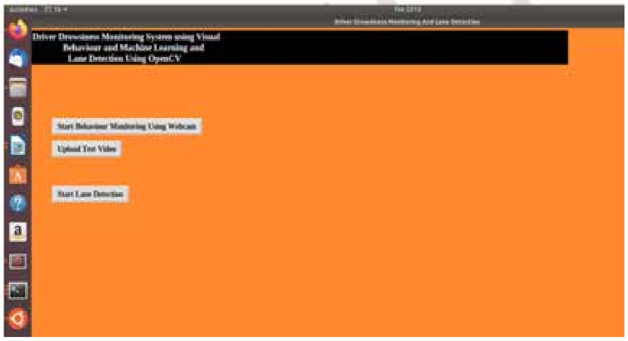
Great traffic day, which was not without its flaws The use of image processing to extract real-time traffic data from CCTV video installed at traffic lights has shown promising results.

PROPOSED SYSTEM

In this study, we are able to detect driver drowsiness by monitoring their visual actions using a camera and an artificial intelligence formula called LBPH (Neighbourhood binary Pattern Histogram). In order to detect if the driver is showing signs of sleepiness, such as blinking his eyes for 20 frames in a row or yawning, this app will use the built-in camera to analyse the driver's photos. Then, it will use the OPENCV LBPH algorithm to extract facial features from the photos. If the range is more specific to sleepy, the application will notify the driver; otherwise, we use the LBPH pre-trained sleepiness design and the Euclidean distance function to continuously examine or anticipate the distance between the eyes and the mouth.

METHODOLOGY

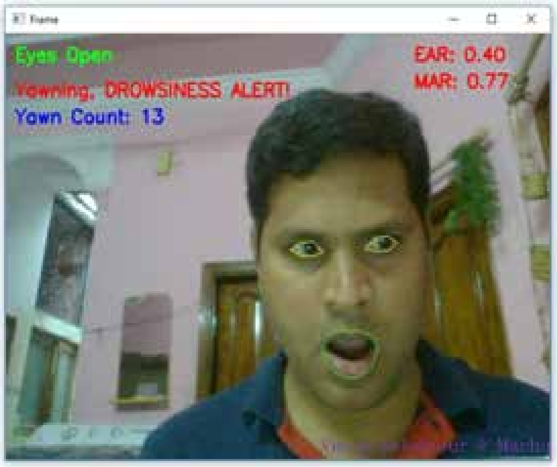
An important contributor to traffic accidents and fatalities is drowsy driving. As a result, there is a lot of activity in the field of identifying driver weariness and its symptoms. Conventional approaches often concentrate on either vehicles, conduct, or physiological factors. A few of approaches need pricey sensors and data processing, while others are obtrusive and divert the driver's attention. Consequently, this research establishes a low-cost, somewhat accurate method for detecting driver sleepiness in real time.


Driver Drowsiness Monitoring And Lane Detection

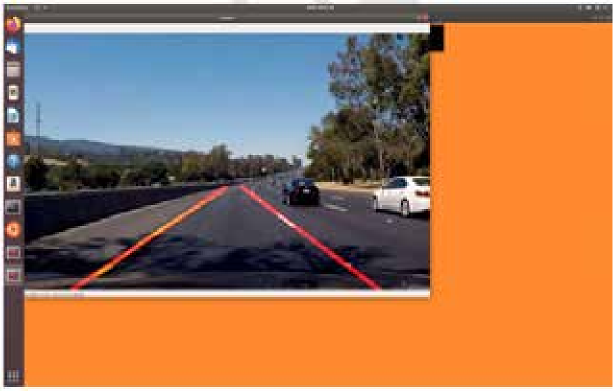
Driver Drowsiness Monitoring System using Visual Behaviour and Machine Learning and Lane Detection Using OpenCV

Home page.

To connect the programme with the camera, click the "Beginning Behaviour Tracking Of Using Cam" button in the upper right. This will bring you to the page below, where you may start streaming your webcam. Above, we can see the live feed from the webcam; if the user's eyes are closed, the app will show all of the frames to let us know.



Eyes Open
Yawning, DROWSINESS ALERT!
Yawn Count: 13
EAR: 0.40
MAR: 0.77



CONCLUSION

Based on aesthetic behaviour and artificial intelligence, this study proposes a low-cost, real-time driver tiredness monitoring system. Here, measurements of visual habits including eye aspect ratio, mouth opening ratio, and nose size proportion are computed from the live video clip captured by a webcam. The development of an adaptive thresholding method has allowed for the real-time detection of driver fatigue. Using the synthetic data that is created, the industrialised system operates perfectly. After everything is said and done, the feature values are saved and put into AI classification formulae. The following Bayesian classifiers, FLDA, and LBPH, have been examined.

To be improved in the future

When compared to Bayesian classifiers, FLDA and LBPH perform better. While both FLDA and LBPH have a uniqueness of 1, their sensitivity values are 0.896 and 0.956, respectively. Work will undoubtedly be carried out to include FLDA and LBPH into the industrialised system in order to conduct the category (i.e., sleepiness discovery) online due to their superior accuracy.

REFERENCES

1. W. L. Ou, M. H. Shih, C. W. Chang, X. H. Yu, C. P. Fan, "Intelligent Video-Based Drowsy Driver Detection System under Various Illuminations and Embedded Software Implementation", 2015 international Conf. on Consumer Electronics - Taiwan, 2015.
2. W. B. Horng, C. Y. Chen, Y. Chang, C. H. Fan, "Driver Fatigue Detection based on Eye Tracking and Dynamic Template Matching", IEEE International Conference on Networking,, Sensing and Control, Taipei, Taiwan, March 21-23, 2004.
3. S. Singh, N. P. papanikolopoulos, "Monitoring Driver Fatigue using Facial Analysis Techniques", IEEE Conference on Intelligent Transportation System, pp 314-318.
4. B. Alshaqaqi, A. S. Baquhaizel, M. E. A. Ouis, M. Bouumehed, A. Ouamri, M. Keche, "Driver Drowsiness Detection System", IEEE International Workshop on Systems, Signal Processing and their Applications, 2013.
5. M. Karchani, A. Mazloumi, G. N. Saraji, A. Nahvi, K. S. Haghighi, B. M. Abadi, A. R. Foroshani, A. Niknezhad, "The Steps of Proposed Drowsiness Detection System Design based on Image Processing in Simulator Driving", International Research Journal of Applied and Basic Sciences, vol. 9(6), pp 878-887, 2015.
6. R. Ahmad, and J. N. Borole, "Drowsy Driver Identification Using Eye Blink Detection," IJISET - International Journal of Computer Science and Information Technologies, vol. 6, no. 1, pp. 270-274, Jan. 2015.
7. A. Abas, J. Mellor, and X. Chen, "Non-intrusive drowsiness detection by employing Support Vector Machine," 2014 20th International Conference on Automation and Computing (ICAC), Bedfordshire, UK, 2014, pp. 188- 193.
8. A. Sengupta, A. Dasgupta, A. Chaudhuri, A. George, A. Routray, R. Guha; "A Multimodal System for Assessing Alertness Levels Due to Cognitive Loading", IEEE Trans. on Neural Systems and Rehabilitation Engg., vol. 25 (7), pp 1037-1046, 2017.

Early Detection of Cancer Using AI

Mohammad Abdul Sayeed, Palepu Pravallika Phani, Satharam Shiva Kumar Reddy
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

G Kalpana Devi
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ gkalapana15@gmail.com

ABSTRACT

Due to the complex nature of the cell, early detection of lung cancer is a monumentally difficult and time-consuming task. Malignant cells start to proliferate rapidly, which leads to the growth of cancer in the body. Predicting early-stage lung cancer cells is a crucial part of photo processing, which is also useful for lung cancer prevention treatment. This proposed system aims to provide a computer-based medical diagnostic system for early detection of lung cancer by using image processing techniques and a constructed neural network classifier. The preprocessing step of this system involves applying masks using morphological operations and the thresholding approach to remove background and surrounding tissue. Several image enhancing algorithms are also part of this activity. The area-based division technique is used to compute the area of rate of interest (ROI). The desired nodule is extracted using the circle fit technique. The feature extraction stage involves drawing out properties such as Span, Mean Strength, Area, Euler Number, and ECD. This step concludes with the training of the Artificial Semantic Network (ANN) for classification using the back breeding technique.

KEYWORDS: *ROI, ECD, ANN, AI, Cancer.*

INTRODUCTION

Uneven and uncontrolled cell growth in the lungs is a key factor in the development of lung cancer. Cigarette smoking is one of the contributing causes. It is quite likely that treatment will be possible if it is discovered early on. Among the most important steps in detecting lung cancer cells is testing. Screening is the process that is used to detect and identify the imperfection. On computed tomography (CT) scans or chest X-rays, a nodule appears as a round white mass. [1] Nodules may be either benign or cancerous, depending on their cause. Nodules in the lungs that are smaller than 3 centimetres in diameter are considered pulmonary or non-cancerous. Sometimes, these imperfections are referred to be innocuous. Malignant blemishes are those with a diameter of more than 3 cm and are toxic. A deadly mole is likely a malignant nodule, thus it's important to diagnose it as soon as possible. It is necessary to keep an eye on these imperfections throughout time to see whether they are becoming worse. There is a possibility of acquiring cancer cells if the size of the spot changes and it is growing. Therefore, it is necessary to notice a flaw. [2] The long-term endurance price of lung cancer cells is particularly low compared to other kinds of cancer cells. Consequently, it provides a vital research system in the field of medical image processing and the early detection of lung cancer is of the utmost importance [3].

LITERATURE SURVEY

Using Image Segmentation Techniques for Lung Blemish Detection, This piece was written by Nanusha.

In order to carry out the robot needle biopsy, it is necessary to first identify and then segment lung abnormalities using computer tomography (CT) pictures. Thresholding, active contour, differential driver, area expanding, and landmark are some of the common division formulae that were studied in this article. We used them to four CT scans of various types of lung imperfections to analyse their efficacy on blemish finding. [5] When it comes to segmenting blemish follow vessels, region growth outperforms the other methods, whereas thresholding, energetic shaping, and differential driver do admirably when it comes to segmenting individual nodules. A differential operator is the best choice for segmenting semi-transparent nodules. While Landmark does a good job of segmenting imperfections surrounded by blood vessels and partially opaque nodules, it is not very sensitive when used on single imperfections. [6]

Analysing and Segmenting CT Images of Abnormal Lungs: Current Methods, Challenges, and Emerging Trends Written by Awais Mansoor, Ph.D., among others

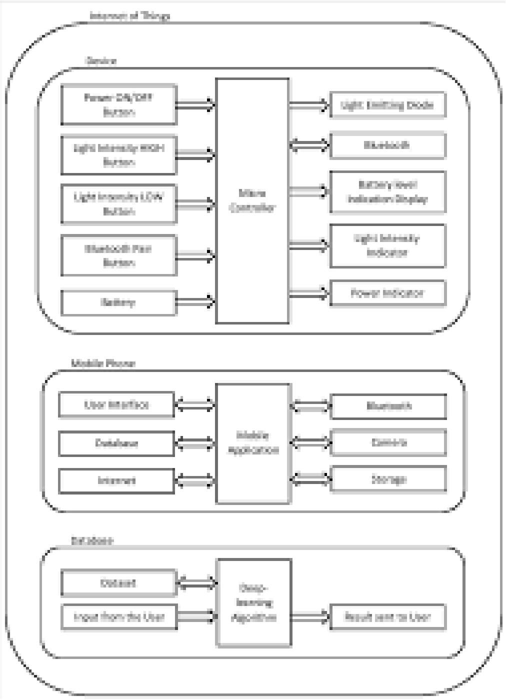
Our goal is to help radiologists better understand their options for decision support in daily practice by reviewing and explaining the capabilities and performance of currently available techniques for dividing lungs with pathologic problems on breast CT pictures. We will also provide illustrations to aid in this process. A crucial first step in radio logic pulmonary image processing is segmentation, a computer-based method that determines the boundaries of the lung from surrounding thoracic tissue on computed tomographic (CT) pictures. For the metrology of lung abnormalities, several algorithms and software application systems provide image division regimens; nevertheless, almost all of the current photo segmentation approaches work effectively only when the lungs show little or no pathologic problems. Because of their inaccurate division approaches, computer-aided detection technologies may fail to accurately depict regions with moderate to high levels of condition or anomalies that have a tough appearance in the lungs.

EXISTING SYSTEM

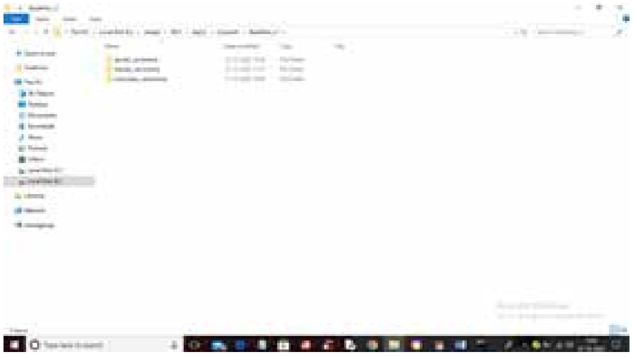
Finding a nodule is a crucial first step in diagnosing lung cancer. Prior to removing the required nodules, the image enhancing pre-processing is repeated. There is a clearing of the items associated to the image's borders. Image pre-processing methods include grey thresholding for binaryization and picture background approaches. To separate the nodules from the lung, an algorithm based on regulatory ions is used. We identify and segment nodules with areas ranging from 75 pixels to 1000 pixe ls for further processing [6, 7].

PROPOSED SYSTEM

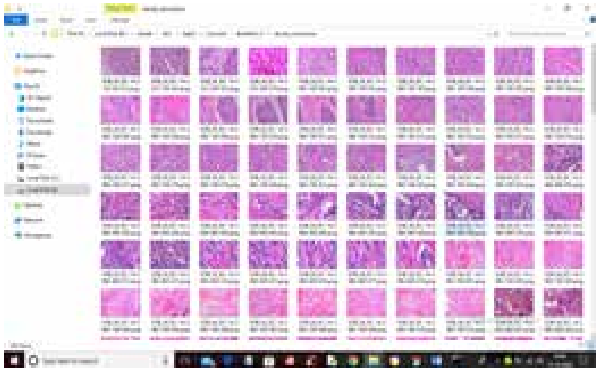
Reducing picture noise is a common use case for the median filter. The median filter determines whether two adjacent pixels in the picture are comparable by looking at how closely they are to each other. Within this filter, the median pixel value of the surrounding pixels is substituted for the original pixel value. In order to improve contrast, the histogram equalisation approach is used to modify picture intensity. It is a visual representation of the values of the intensity of the pixels in the picture. The data structure that keeps track of the occurrences of each pixel's intensity level in the picture is one possible interpretation [8].



Our goal here is to predict cancer cases using an AI algorithm called a Neural Network, which we'll be training using a plethora of optimizer techniques including ADAM, SGD, and Slope Descent Mini Batch. Below is a presentation displaying the cancer information derived from the photographs you provided, which include three different types of cancer cells or stages. These photos were used to train our AI algorithms. [9]



You can see three distinct forms of cancer in the dataset; to see photos of each, open any folder on the top screen.



So, we are training AI with three different optimizers utilising the photos up top.

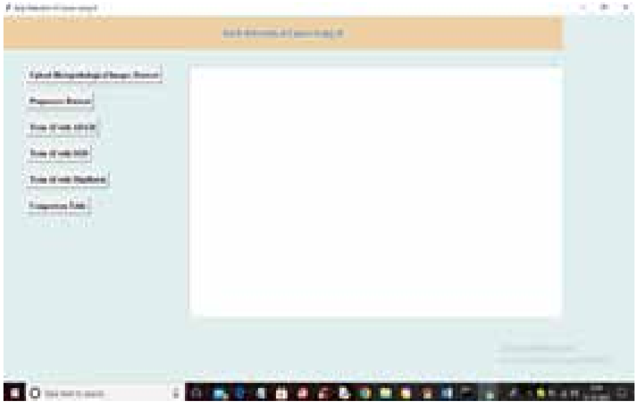
We have developed the following modules to carry out this project.

1) Histopathological Image Dataset Publication: This component will undoubtedly be used to publish the dataset to the programme.
2) Preprocess Dataset: This module will be used to read all the photographs, resize them to the same size, and then normalise their pixel values. Following processing, the dataset will be immediately divided into test and train sets.
3) Use ADAM to train AI: Our AI formula with ADAM as the optimizer will be fed training data using this component. To find the accuracy of the predictions, we will use the test data from the trained version after training.
4) Use SGD to Train AI: As a matter of fact, we will feed the Training Data to the AI algorithm using the optimizer as utilising this module. In order to determine the accuracy of our predictions, we shall use certified design assessment data after training.
5) Use MiniBatch to Train AI: This module will be used to input Training Data into an AI formula while using MiniBatch as the optimizer. We will determine the accuracy of the prediction using the certified model's test data after training.

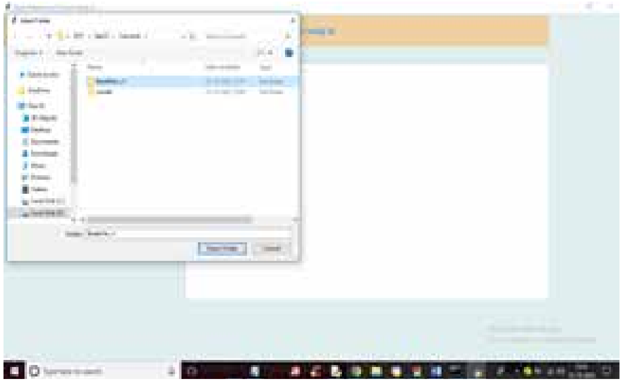
We will provide performance in tabular form and plot all algorithm precision using this component, which is called a comparison table.

SCREEN SHOTS

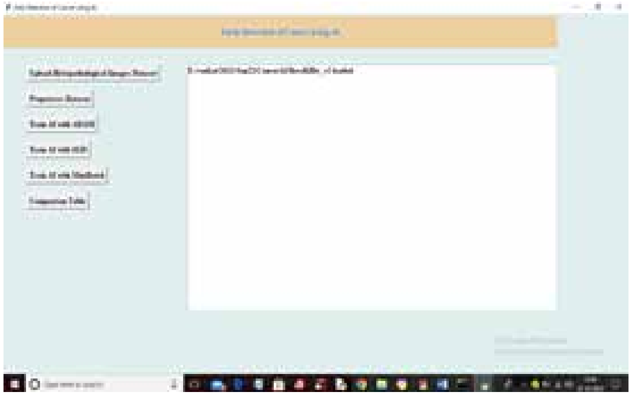
To run project double click on 'run.bat' file to get below screen



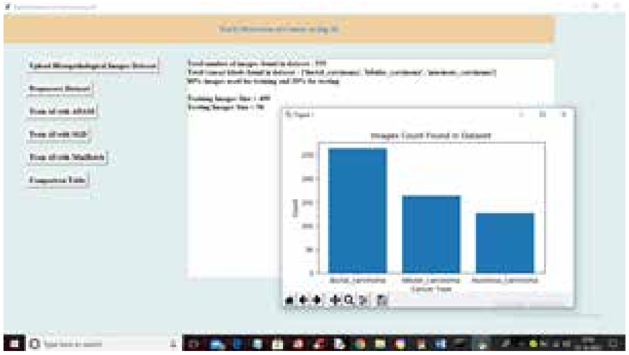
Select the "Upload Histopathology Images Dataset" option on the previous page to upload the dataset. Then, you will see the following result.



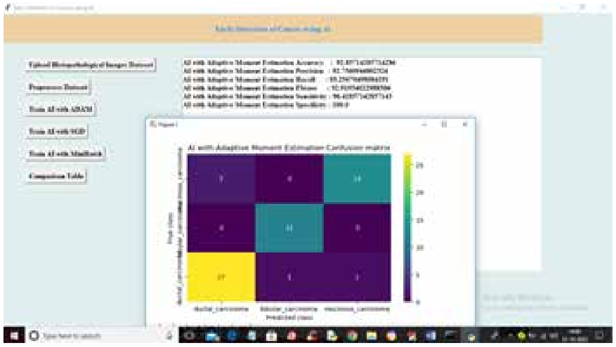
Pick the whole "Dataset" folder to upload on the previous page, then hit the "Select Folder" button to load the dataset. The results will be shown below.



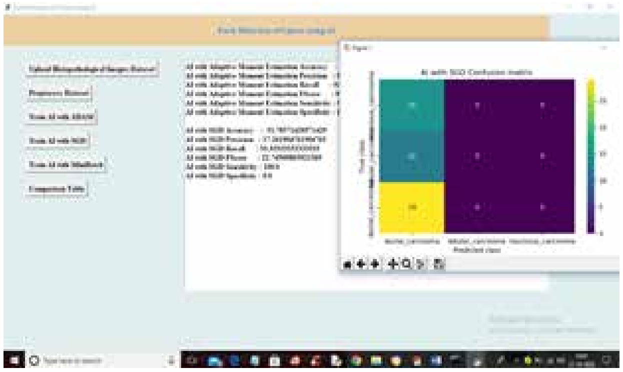
Once the dataset is imported in the above page, you may read and process the photos by clicking on the "Preprocess Dataset" button. The outcome will be shown below.



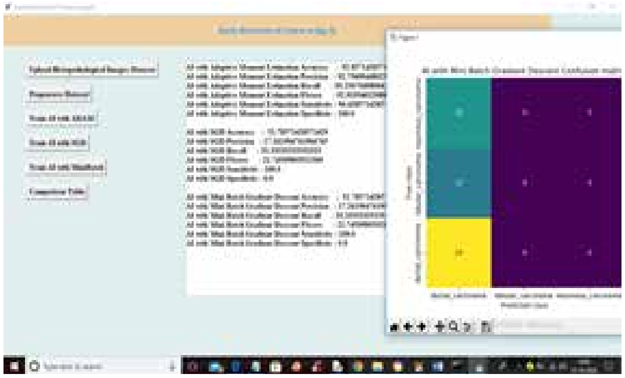
The dataset contains 555 photos, as shown in the above screen. Then, the train and test data dimensions are revealed. The x-axis of the chart represents the type of cancer cells, and the y-axis represents the number of pictures of that type. To train the AI, close the photo and click on the "Train AI with ADAM" button. The output will be below.



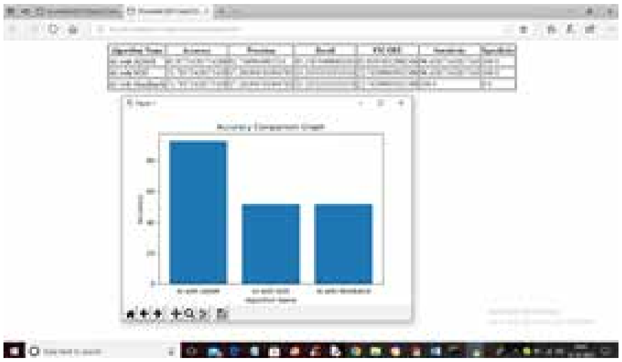
In the following presentation, AI with ADAM achieved an accuracy of 92%. The x-axis shows predicted labels, the y-axis shows true labels, and various coloured boxes reflect the count of correct predictions, while the identical blue boxes represent the count of inaccurate predictions. This is shown in a confusion matrix chart. To learn SGD, just close this window and then click the "Train with SGD" button; the results will be shown below.



To teach AI, shut the above chart, and then click the "Train with MiniBatch" button. We achieved 51% accuracy using SGD. The results are below.



In above screen with Mini Batch also we got 51% accuracy and now close above graph and then click on 'Comparison Table' button to get below output.



In above screen we can see all algorithm performance in tabular and graphical format and in all algorithms 'AI with ADAM' got high performance or accuracy

CONCLUSION

In addition to their many applications in modern medicine, CA D Solutions are useful for the detection of cancerous lesions. A circle fit formula with maximum spread is used to identify a blemish with the desired area, eliminating the requirement for erroneous nodules. We get more exact results with each model. A system-supplied accuracy of 95.6% resulted from this. The system's sensitivity is 93.1% and its uniqueness is 100%. If a CT scan of the lung reveals a nodule, this method can tell you whether it's benign or cancerous. One of this system's future uses is to aid in the diagnosis of cancer in various human organs. This system's methods may be used to either decrease the formation of abnormal cells or distribute them to other places of the body. Improving this approach for MRI and ultrasound images is possible. Compared to support vector machines (SVMs), NN classifiers provide more accurate and exact results; nevertheless, they need a much larger range of data inputs.

REFERENCES

1. World Health Organisation. Cancer: fact Sheet no. 297. 2015 July 8. http://www.who.int/mediacentre/factsheets/fs297/en/.Jemal A, Siegel R, Xu J, et al. Cancer statistics, 2015.. CA Cancer J Clin.2015; 60(5):277–300.

2. The diagnosis of lung cancer (update) Published by the National Collaborating Centre for Cancer (2nd Floor, Front Suite, Park House, Greyfriars Road, Cardiff, CF10 3AF) at Velindre NHS T rust, Cardiff, Wales[2011] Database from: The Cancer Imaging Archive. http://doi.org/10.7937/K9/TCIA.2015.A6V7JIWX

3. Nanusha, "Lung Nodule Detection Using Image Segmentation Methods", International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE) Volume 6, Issue 7, July 2017

4. Awais Mansoor Ph.D et al, "Segmentation and Image Analysis of Abnormal Lungs at CT: Current Approaches, Challenges, and Future Trends", Radio Graphics 2015; 35:1056–1076 Published online 10.1148/rg.2015140232.

5. Ozge Gunaydin et al "Comparision of Lung Cancer Detection Algorithm" 2019 , 978-1-7281-1013-4/19/ $31.00 © 2019 IEEE.

6. Madhura J , Dr .Ramesh Babu D R "A Survey on Noise Reduction Techniques for Lung Cancer Detection" International Conference on Innovative Mechanisms for Industry Applications (ICIMIA 2017), 978-1-5090-5960-7/17/$31.00 ©2017 IEEE

7. Ashwin S, Kumar SA, Ramesh J, Gunavathi K: Efficient and reliable lung nodule detection using a neural network based computer aided diagnosis system. In Emerging Trends in Electrical Engineering and Energy Management (ICETEEEM), 2012 International Conference 2012:135–142.

8. Rachid Sammouda " Segmentation and Analysis of CT Chest Images for Early Lung Cancer Detection" 2016 Global Summit on Computer & Information Technology 978-1-5090-2659-3/17 $31.00 © 2017 IEEE [8] M. Egmonet-Petersen et al "Image Processing with Neural Networks- a Review", Patteren Recognation, Volum-35, No. 10, PP. 2279-2301, 2002.

9. V. Vapnik, "An overview of Statastical Learning Theory", IEEE transactions on Neural Networks, Vol-10 No. 5, PP. 988-999, 1999.

10. A. Kanakatte, N. Mani, B. Srinivasan, and J. Gubbi, "Pulmonary Tumor Volume Detection from Positron Emission Tomography Images," 2008 International Conference on BioMedical Engineering and Informatics, 2008.

11. A. Hashemi, A. H. Pilevar, and R. Rafeh, "Mass Detection in Lung CT Images Using Region Growing Segmentation and Decision Making Based on Fuzzy Inference System and Artificial Neural Network," International Journal of Image, Graphics and Signal Processing, vol. 5, no. 6, pp. 16–24, 2013.

12. Nidhi S. Nadkarni, "Detection of Lung Cancer in CT Images using Image Processing" Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019).

Image Forgery Detection based on Fusion of Lightweight Deep Learning Models

Namana Tarun, Kotthapalli Abhijeeth
Atluri Venkata Mega Syam
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

Jonnadula Narasimharao
SVSV Prasad Sanaboina
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ jonnadula.narasimharao@gmail.com

ABSTRACT

Taking images has really become more popular in the last several years, primarily because of the ubiquitous availability of electronic cameras. We rely on photos every day for a wealth of information, and it's not uncommon to need to edit photos to get more details. While there are many technologies available to improve image quality, they are also often abused to alter images in a way that spreads misleading information. This makes picture imitators much more common and dangerous, which is a major issue right now. There are a lot of tried-and-true methods for spotting photo frauds. Convolutional neural networks (CNNs) have been in the spotlight recently, and they have had an impact on the field of picture fraud detection as a whole. Although there are other CNN-based picture imitation algorithms in the literature, the most of them are only able to identify particular types of fraud, including copy-move or image splicing. As a result, a technique that can precisely identify whether a photo contains a hidden fake is needed. In this study, we offer a long-term deep discovery based technique for identifying image forgeries in the context of double photo compression. We use the difference between an image's original and compressed versions to train our model. The suggested model outperforms state-of-the-art approaches in terms of speed and efficiency while being relatively lightweight. An overall recognition accuracy of 92.23% is rather promising according to the experimental results.

KEYWORDS: *CNN, Image forgery, High efficiency.*

INTRODUCTION

The widespread and inexpensive availability of electronic equipment is a direct outcome of both technical progress and globalisation. Electronic cameras have so become more popular. The world is filled with camera sensing units, and we take use of them to capture a multitude of photographs. Every day, people publish a plethora of photos on social media, and many online filing systems need digital versions of these images. Amazingly, even someone without a formal education can look at photos and get information from them. Images therefore serve a vital purpose in the digital world, both in terms of storing and sharing information. You may rapidly alter the photographs with a variety of tools available [1,2]. The intention behind creating these tools was to enhance the appearance of the photos. Instead of enhancing the image, some individuals, however, abuse their powers to propagate incorrect information and distort photographs [3, 4]. This presents a serious risk since erroneous images can cause serious, often permanent harm.

- Below, we will go over the two main types of image bogus: picture splicing and copy-move.
- In image splicing, a donor's image is cut and pasted into another image. The final fake picture may also be built from a series of donor photographs.
- This scenario just contains one picture, so we can copy and move it. The picture has a portion that has been cut out and pasted. Many more things may be

hidden in this way as well. There are no borrowed features in the final rendered image.

In both instances of photo forging, the main goal is to disseminate false information by replacing the original image content with something else [5,6]. Photos from the past were a great way to share knowledge, but now people exploit them to propagate false information due to picture piracy. Since photo manipulation may be subtle or even unnoticeable to the human eye, this is impacting the public's reliance on photographs. The dissemination of misinformation and the restoration of public trust in images need the identification of picture fake. To do this, one may use different photo handling procedures to identify the many artefacts that are left behind when an image imitation is created.

A number of methods for detecting the presence of photo knockoffs have been proposed by scientists. (7) through (9) To identify fake photos, conventional techniques look for signs of manipulation in the image's illumination, contrast, compression, sensing unit noise, and shadow, among other artefacts. Several computer vision tasks, such as image object recognition, semantic division, and picture categorization, have seen increased interest in CNNs in recent years. The success of CNN in computer vision is enhanced by two important features. As a first step, CNN makes use of the strong relationship between nearby pixels. Due to this, CNN prefers locally-organised linkages rather than one-to-one relationships between every pixel. Secondly, by using shared weights in a convolution process, each output feature map is created. Further, CNN may generalise itself to identify undiscovered forgeries by using found characteristics from training photographs, as opposed to the conventional method that depends on built functions to detect specific imitation. Thanks to these features, CNN is a promising method for detecting instances of imitation in images. A convolutional neural network (CNN) model may be trained to recognise the many artefacts included in a constructed image [10--13]. As a consequence, we suggest a very lightweight convolutional neural network (CNN) that can learn to identify manipulated images by comparing their original and altered attributes.

Major costs associated with the suggested approach are as follows:

Finding good picture replication is the goal of a lightweight CNN-based design. The suggested approach uses picture re-compression to examine disparities in image resources and tests for many artefacts left over from the image manipulation procedure.

- Unlike most current formulae, our approach has achieved great accuracy in picture fake identification and can identify both image splicing and copy-move imitations. - The proposed method can detect the existence of picture fraud in a fraction of the time it takes current methods. Its speed and precision make it perfect for practical use, and it works well even on sluggish devices.

LITERATURE SURVEY

The management of picture forgeries has been proposed in many ways in the literature. Below, we outline new techniques that use convolutional neural networks (CNNs) and deep learning, in contrast to the majority of conventional approaches that rely on artefacts left behind from picture copying. We will go over the various conventional approaches first, and then move on to techniques that rely on deep learning.

The authors of [1] proposed using error level evaluation (ELA) to identify instances of picture mimicry. Forgery in a picture may be detected in [2] by analysing lighting issues with objects. By comparing the lighting instructions of the constructed and real parts of a picture, it attempts to locate the copy. Various traditional methods for detecting fake images have been evaluated in [3]. For the purpose of imitation discovery, Habibi et al. [4] get the side pixels by means of the contourlet transform. A method based on JPEG compression was introduced by Dua et al. in [5]. All of the individual DCT coefficients are checked for every single block of a picture that has been divided into non-overlapping 8 × 8 pixel blocks. When a JPEG compressed picture perturbed, the statistical properties of the a/c components of the block DCT coefficients change. In order to distinguish between real and fake images, the SVM uses the obtained function vector. Ehret et al.'s [6] method for fake detection makes advantage of SIFT, which offers thin keypoints with size, turning, and illumination invariant descriptors. It is suggested to use deep Boltzmann machines (DBM) for picture evaluation of high-level properties in order to detect fingerprint forgeries. Balsa et al. [7] compared the discrete Fourier transform (DFT), Walsh-Hadamard transform (WHT), Haar wavelet transform (DWT), and DCT for analogue picture transmission, adjusting compression, and assessing quality. By comparing the images from different domains, they may be used for photo fake detection. In their proposal for a hybrid approach to photo splicing, Thanh et al. [8] sought to recover the original photographs from which the composite was made, in the event that a certain photo was identified as the composite. Combining Zernike minute and SIFT properties, they provide a hybrid picture access approach.

In their work on image forgeries, Bunk et al. [9] used resampling functions and deep understanding to create a system. Bondi et al. proposed a technique for detecting manipulation of images by the clustering of CNN features derived from cameras. In order to simultaneously obtain forensic aspects of compression artefacts on the DCT and RGB domains, Myung-Joon invented CAT-Net in [2]. The HR-Net (high resolution) network is their main focus. They used the method suggested, which explains how to train a CNN using the DCT coefficient; providing the CNN with DCT coefficients in their raw form is inefficient. Ashraful et al. [10] proposed DOA-GAN, a GAN with dual attention for detecting and localizing copy-move fake in pictures. While the generator's first-order focus is intended to gather copy-move area data, more discriminating residential or commercial features are used in the second-order focus for patch co-occurrence. Location-aware and co-occurrence features are incorporated into the primary detection and localization branches of the network. These features are derived from the focus maps that are retrieved from the fondness matrix.

In their proposal for the detection of copy-move photo forgeries, Yue et al. [12] put out BusterNet. Its two branches meet in the middle, which is a mix component. Visual artefacts are used by both branches to identify possible control positions and copy move areas are located by aesthetic similarities. Using a convolutional neural network (CNN), In [3], Yue et al. developed ManTra-Net, a fully convolutional network capable of handling any dimensional image and many imitation types, including copy-move, augmentation, splicing, removal, and unknown fake types. In their recommendation of PSCC-Net, Liu et al. [9] outlined a two-pronged approach to image analysis: first, a top-down route that retrieves global and regional functions; second, a bottom-up route that detects manipulation and predicts four layers of masks, with each level being constrained on the one before it.

PROPOSED SYSTEM

Designed to function as non-linear linked nerve cells, CNNs take their cues from the human aesthetic system. Photo segmentation and object identification are only two of the many computer vision applications where they have already shown exceptional capability. Photo forensics is only one of many potential additional uses for them. Because it is so dangerous, detecting picture fakes is essential, and it is really rather easy to achieve with the equipment available today. Because each image has its own unique origin, a wide variety of artefacts manifest themselves while moving image fragments from one location to another. These aberrations may be invisible to the naked eye, but convolutional neural networks (CNNs) can identify them in doctored photos. The reason the produced region gets different compression boosts when we recompress historical photographs and the source of the formed area is different is that the two types of photos have different compression ratios. We train a CNN-based design to determine the authenticity or fakery of a picture using this notion in the suggested approach.

The original area's distribution of DCT coefficients will likely be statistically different from the merged region after image splicing. Repeated patterns on the pie chart are the result of the real area being squeezed twice: once in the camera and again in the counterfeit. [2] in When the secondary quantization table is applied, the spliced part mimics the behaviour of an alone pressed area.

Recompressing an image with a fake in it causes the constructed piece to compress differently from the rest of the picture. This is because the original photo's source and the created section's resource are different. This puts the fake component front and centre when comparing the real shot to its compressed version. This is why we use it to train our convolutional neural network (CNN) model that can identify picture manipulation.

The proposed method, as seen in Algorithm 1, is really explained here. We compress the tampered-with picture A (pictures in Figure 1b) and name it Arecompressed (pictures in Figure 1c), then we take the forged picture A and use it as our starting point. Figure 1e shows the difference between Figure 1b and Figure 1c, and we'll refer to this as Adi f. The original and re compressed photographs are now separated. Currently, the altered component becomes more prominent in Adi f (as shown in Number 1d, e in particular) because to the discrepancy between the photo's original and forged parts. We use Adi f as an input function to train a convolutional neural network (CNN) to distinguish between authentic and fake photos, and we do so by identifying the picture as a highlighted one. Number 2 provides a visual representation of how the suggested method generally operates.

We use JPEG compression to generate A recompressed from A. As seen in Figure 3, Image An undergoes JPEG compression, resulting in A recompressed. When there is only one compression, the created section of the image shows the pattern that can be seen in Figure 4. The demagnetized coefficients pie chart likewise shows this pattern. As noted in Number 5, Figure 6 illustrates how this pattern is displayed by the actual portion of the image when there is a subsequent type of double compression. Between the demagnetized coefficients, there is an opening.

Lines 5–13 of Algorithm 1 include our proposed model, which is a lightweight CNN model with few requirements. The following is an explanation of the design we developed, which includes three convolutional layers followed by a dense fully attached layer:

32 3-by-3 filters with a single stride dimension and an activation feature called "relu" make up the first convolutional layer.

- The second convolutional layer has an activation function called "relu" and 32 3-by-3 filters. It also has a stride dimension of one.

- The third convolutional layer has 32 7-by-7 filters, a one-dimensional stride, and the "relu" activation feature. It is followed by a 2-by-2 max-pooling layer.

- The thick layer is finally revealed. It is made up of 256 neurons that are triggered by the "relu" function and connected to two output neurons that are triggered by the "sigmoid" function.

RESULTS

Double click the "run.bat" file to launch the project and see the output below.

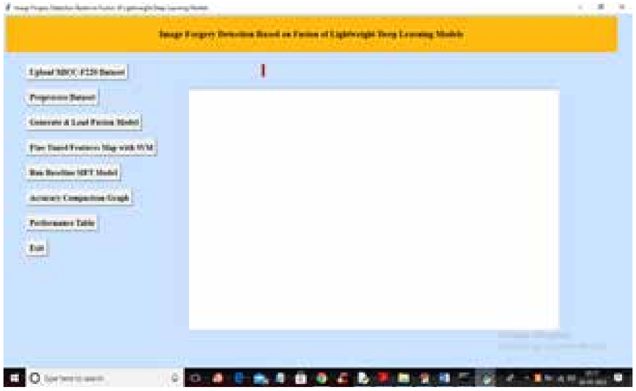


Image Forgery Detection Based on Fusion of Lightweight Deep Learning Models
Upload MICC-F220 Dataset
Preprocess Dataset
Generate & Load Fusion Model
Fine Tuned Features Map with SVM
Run Baseline SIFT Model
Accuracy Comparison Graph
Performance Table
Exit

The following is the result you will see after clicking the "Upload MICC-F220 Dataset" button on the previous page.

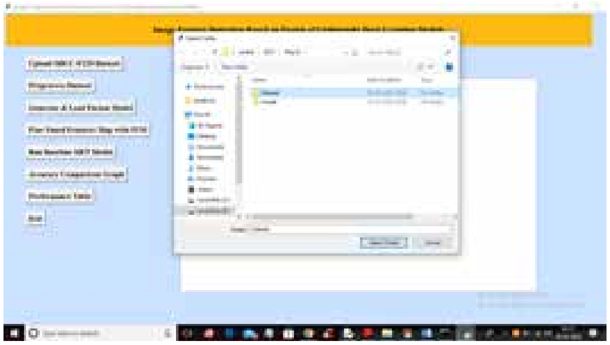


Image Forgery Detection Based on Fusion of Lightweight Deep Learning Models
Upload MICC-F220 Dataset
Preprocess Dataset
Generate & Load Fusion Model
Fine Tuned Features Map with SVM
Run Baseline SIFT Model
Accuracy Comparison Graph
Performance Table
Exit
Dataset

To load the dataset, go to the previous page, choose the "Dataset" folder, and then click the "Select Folder" button. Then, you should see the following output:

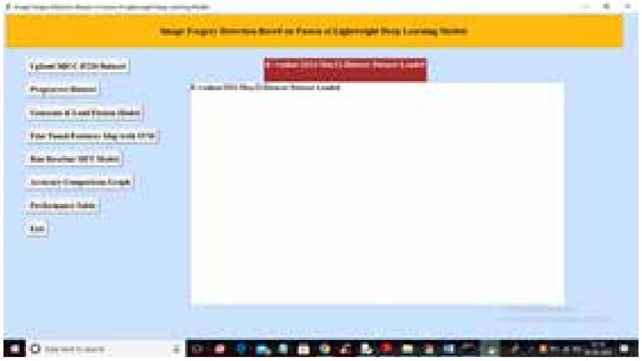
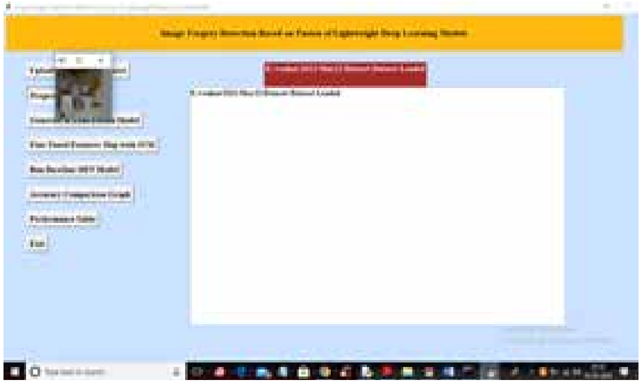
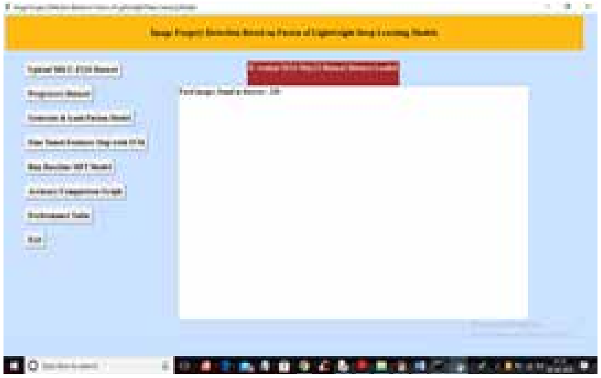


Image Forgery Detection Based on Fusion of Lightweight Deep Learning Models
Upload MICC-F220 Dataset
Preprocess Dataset
Generate & Load Fusion Model
Fine Tuned Features Map with SVM
Run Baseline SIFT Model
Accuracy Comparison Graph
Performance Table
Exit

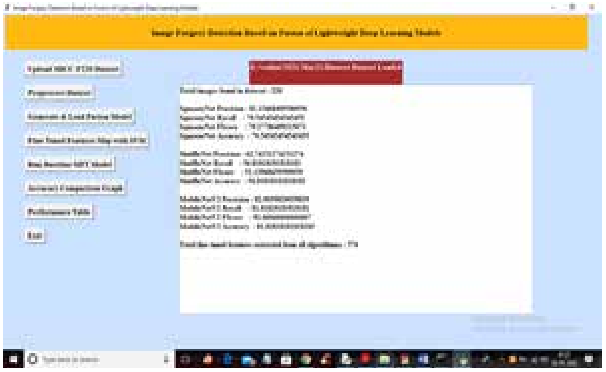
Once you have the dataset loaded in the above page, you may read and normalise all of the photos by clicking the “Preprocess Dataset” button. The result will be shown below.



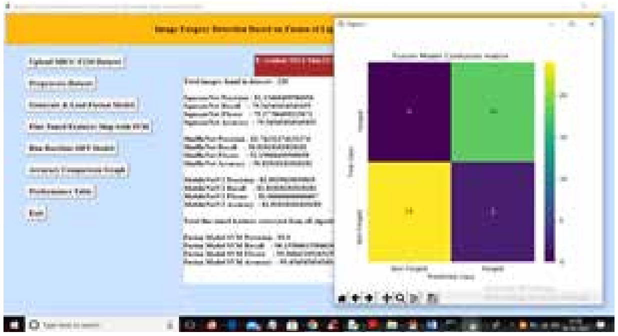
Here you can see all the processed photos. I’ve loaded one example image to make sure everything is working correctly. Now you can close the image to see the outcome.



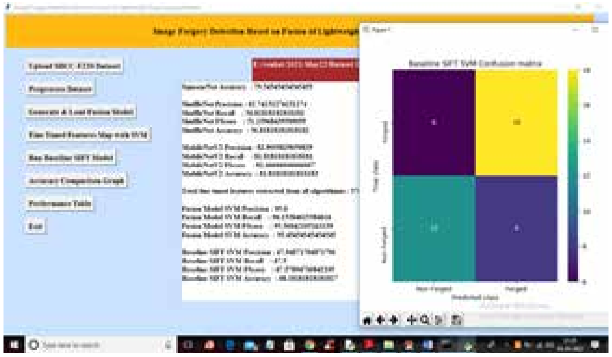
The 220 processed photos make up the dataset; to train the algorithms, extract features, and determine their correctness, click the “Generate & Load Fusion Model” button.



Press “Fine Tuned Features Map with SVM” to use the extracted features to train SVM and get the accuracy of the fusion model. On the previous page, you can see the accuracy of all three techniques. On the final line, you can see that the programme extracted 576 features from all three approaches.



The model obtained 95% accuracy with the refined SVM blend model, as demonstrated by the confusion matrix, where the true labels are displayed on the y-axis and the predicted labels are displayed on the x-axis. and both the X and Y boxes show that there are even more courses that were correctly predicted. Using the ‘Run Standard SIFT Version’ button, we can train SVM using SIFT’s existing features and acquire its accuracy. After that, we can see in all the formulae that tweaking features with SVM has truly produced high precision and a presently closed complexity matrix graph.



In the confusion matrix chart, we can see that current SIFT incorrectly predicted 6 and 8 scenarios, while in the above presentation, we achieved 68% accuracy with existing SIFT SVM. Now that we’ve established that the current SIFT characteristics are inaccurate in the prediction, we can exit the above chart and get the graph below by clicking the “Accuracy Comparison Chart” button.

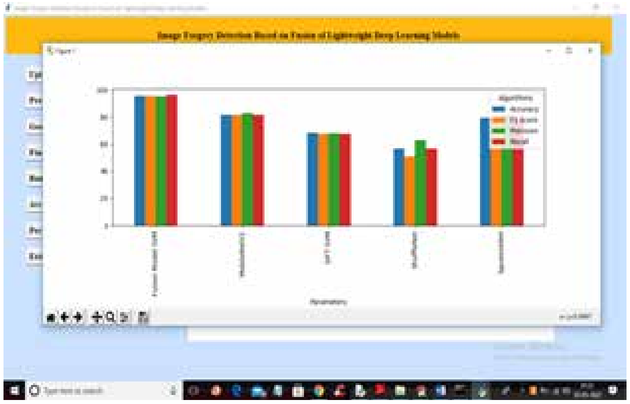
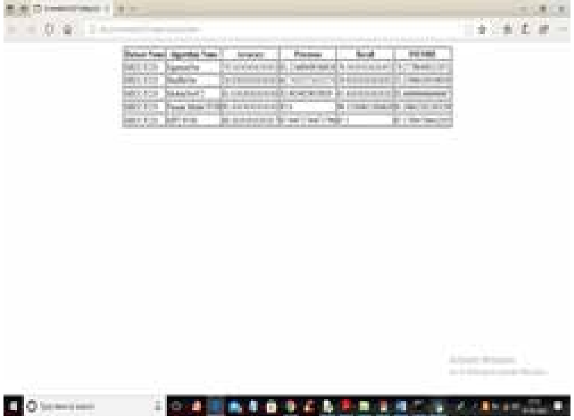


Image Forgery Detection Based on Fusion of Lightweight Deep Learning Models
100
80
60
40
20
0
Algorithms
Accuracy
Parameters

The names of the formulae are shown on the x-axis of the above graph, while the y-axis displays numerous metrics, including accuracy and recall, with different coloured bars representing these metrics. After you've closed the above chart, click the "Efficiency Table" button to get the results in tabular form.



[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

The above-displayed recommendation fusion model SVM with modify characteristics outperformed all other algorithms with a 95% accuracy rate.

CONCLUSION

More people are taking pictures now than ever before because to the widespread availability of electronic cameras. Pictures are vital to our daily lives and have grown in importance as a means of communication because people understand them so easily. There are a number of picture editing programmed available, and although their primary purpose is to improve photographs, people often use these current technologies to create fake images and use them to propagate false information. Image imitation has therefore become a major concern and source of problems. Through the use of convolutional neural networks (CNNs) and deep learning, we provide a novel image imitation discovery method in this article. The suggested approach employs a CNN architecture that utilises variations in picture compression to get desirable outcomes. We train the model using the difference between the original and compressed photos. With the described approach, image splicing and copy-move picture imitations can be identified. The experimental results indicate a very promising overall validation accuracy of 92.23% with a specified iteration constraint.

We aim to further develop our photo imitation localization technology in the future. Furthermore, we will combine the proposed method with other popular image localization techniques to minimize their complexity and increase their speed and accuracy. We will also improve the suggested approach to deal with spoofing. We will enhance the suggested method so it works well with little photographs, since the current methodology requires a photo resolution of at least 128 × 128. In addition, we will be training deep learning networks to identify picture forgeries by creating a demanding large-scale photo imitation dataset.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "IMAGE FORGERY DETECTION BASED ON FUSION OF LIGHTWEIGHT DEEP LEARNING MODELS", which gave us excellent resources and assistance to complete our assignment. I want to express my gratitude to our chairman, director, deans, head of the department, computer science and engineering department, guide, and teaching and non-teaching faculty members for their insightful advice and direction in all facets of our work.

REFERENCES

1. Xiao, B.; Wei, Y.; Bi, X.; Li, W.; Ma, J. Image splicing forgery detection combining coarse to refined convolutional neural network and adaptive clustering. Inf. Sci. 2020, 511, 172–191. [CrossRef]

2. Kwon, M.J.; Yu, I.J.; Nam, S.H.; Lee, H.K. CAT-Net: Compression Artifact Tracing Network for Detection and Localization of Image Splicing. In Proceedings of the 2021 IEEE Winter Conference on Applications of

Computer Vision (WACV), Waikoloa, HI, USA, 5–9 January 2021; pp. 375–384.

3. Wu, Y.; AbdAlmageed, W.; Natarajan, P. ManTra-Net: Manipulation Tracing Network for Detection and Localization of Image Forgeries With Anomalous Features. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9535–9544.

4. Ali, S.S.; Baghel, V.S.; Ganapathi, I.I.; Prakash, S. Robust biometric authentication system with a secure user template. Image Vis. Comput. 2020, 104, 104004. [CrossRef]

5. Castillo Camacho, I.; Wang, K. A Comprehensive Review of Deep-Learning-Based Methods for Image Forensics. J. Imaging 2021, 7, 69. [CrossRef] [PubMed]

6. Zheng, L.; Zhang, Y.; Thing, V.L. A survey on image tampering and its detection in real-world photos. J. Vis. Commun. Image Represent. 2019, 58, 380–399. [CrossRef]

7. Jing, L.; Tian, Y. Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey. IEEE Trans. Pattern Anal. Mach. Intell. 2020, 43, 1. [CrossRef]

8. Meena, K.B.; Tyagi, V. Image Forgery Detection: Survey and Future Directions. In Data, Engineering and Applications: Volume 2; Shukla, R.K., Agrawal, J., Sharma, S., Singh Tomer, G., Eds.; Springer: Singapore, 2019; pp. 163–194.

9. Mirsky, Y.; Lee, W. The Creation and Detection of Deepfakes: A Survey. ACM Comput. Surv. 2021, 54, 1–41. [CrossRef]

10. Rony, J.; Belharbi, S.; Dolz, J.; Ayed, I.B.; McCaffrey, L.; Granger, E. Deep weakly-supervised learning methods for classification and localization in histology images: A survey. arXiv 2019, arXiv:abs/1909.03354.

Hospital Exigency Forecast

Paka Vishwanth, Athram Sandhya Rani,
Mohammed Abdul Wahab
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

Nuthanakanti Bhaskar
Associate Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ bhaskar4n@gmail.com

ABSTRACT

Clients might have significant negative impacts due to overcrowding in emergency departments (EDs). Emergency departments should thus consider using innovative methods to improve patient circulation and alleviate congestion. Data mining utilising AI methods to forecast emergency department admissions is one potential approach. In order to compare and evaluate different artificial intelligence algorithms for predicting the risk of admission from the emergency department, this research uses regularly collected administrative data (120 600 records) from two large acute healthcare institutions in Northern Ireland. The anticipated designs are developed using three algorithms: first, logistic regression; second, decision trees; and third, GBMs. Compared to the logistic regression model (accuracy D79:94%, AUC-ROC D0:849) and the choice tree (precision D80:06%, AUC-ROC D0:824), the GBM performed better. Thanks to logistic regression, we can identify a plethora of factors linked to hospital admissions, such as health centre website, age, arrival setting, triage categorization, treatment team, and prior admissions in the last month and year. The potential usefulness of three popular maker learning algorithms for admissions forecasting is discussed in this article. An image of expected emergency department admissions at a given time would be provided by the designs presented in this paper if they were to be implemented in decision support tools. This would allow for advance source planning, the avoidance of client circulation bottlenecks, and the comparison of actual and predicted admission rates. Although GBMs will undoubtedly be helpful in situations where precision is paramount, EDs should think about using logistic regression methods where translation ability is a crucial aspect.

KEYWORDS: *GBM, ED, AUC ROC D, DNN.*

INTRODUCTION

Deep neural networks (DNNs) revolutionised AI by allowing for the use of massive datasets and function spaces to generate accurate predictions. Demonstrating their efficacy as discovery algorithms, DNNs have achieved modern efficiency in a wide variety of tasks [1]. The medical field has taken note of their feature estimate power; several publications have used them to create useful predictions for different healthcare scenarios [3].

One problem with DNNs is that they aren't perfect optimisation problems; in other words, the optimal performance that the formula promises may be out of reach [4]. The development of methods for providing structured data to the network for training purposes has therefore received a lot of attention [5]. This is now widely used for training DNNs and has the official moniker of an educational programme.

Utilising the idea of curriculum training, this effort aims to train a design that can predict the exact location of a patient's admission from very early data collected by the triage nurse in the emergency department. Our goal is to show that there is a predictable flow of patients from emergency departments to seven distinct kinds of hospital wards. To ensure the patient gets treatment and therapy as soon as feasible, this would allow for the appropriation of a bed and resources long before arrival [6]. In addition, we want to show that this prediction is possible using data collected from patients at the emergency department entrance, which will enhance the movement of patients from the ED to the rest of the hospital. Client admissions to the best health centre ward tend to be more problematic during peak demand periods, such as when seasonal flu cases are at their highest. This is why we monitor our design's efficacy all year round [7].

EXISTING SYSTEM

Using two years of administrative data that is collected frequently, Sunlight et al. [8] created a logistic regression model to predict the likelihood of admission during triage. Age, ethnicity, method of arrival, person acuity score, chronic diseases, and previous emergency department visits or admissions within the last three months were all associated with the threat of hospitalisation. The previous version did not account for sex, even though data showed that more females than men were admitted [9].

Similarly, health centres in Glasgow provided logistic regression data for two years, which Cameron et al. used to predict the probability of admissions during triage. 'Triage group, age, National Early Caution Rating, arrival by rescue, referral source, and admission within in 2014' (pp. 1) was one of the most important predictors in their version, with an area under the curve of the receiver operating characteristic (AUC-ROC) of 0.877. In order to predict emergency admissions, Kim et al. used routine management data in conjunction with a logistic regression model. Their best model, however, had a less exact design, coming in at 76% [10].

PROPOSED SYSTEM

The suggested system uses data mining to identify patients who are at high risk of requiring inpatient admission, and then takes the necessary precautions to avoid system bottlenecks, all with the goal of reducing emergency department capacity and improving client care. For instance, inpatient monitoring, team preparation, and the promotion of specialised job streams within the emergency department might all benefit from a version that accurately predicts hospital admissions. By providing early notification that admission is imminent, Cameron et al. argued that the system's deployment might help to improve patient satisfaction. Data mining techniques might be used to build such a model. These approaches include inspecting and analysing data in order to extract useful features and knowledge that can influence choices. Pattern recognition, data definition, and prediction based on pattern recognition are common components of this process. Using machine learning formulae to create designs that can predict hospital admissions from the emergency room and comparing the efficacy of various approaches to design creation is the main emphasis of this study. We trained and validated the models on data extracted from the HR databases of two high-volume Northern Irish healthcare institutions.

MODULES

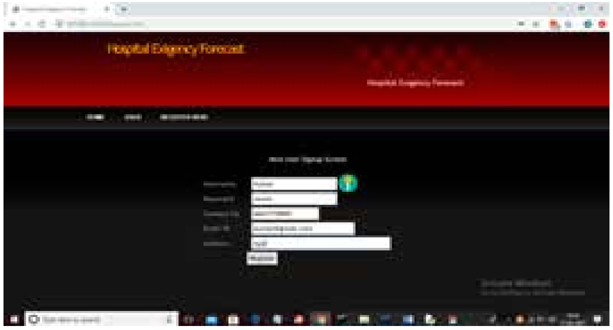
User

Within this component, there is a need for a computerised system that can assess a patient's vital signs and inform them if they require emergency admission or not. Patients who do not require admission can be discharged from the facility. To accomplish this, we are utilising three artificial intelligence formulas: Choice Tree, Logistic Regression, and Slope Boosting. Out of these, Choice Tree and Slope Boosting yield the best results.

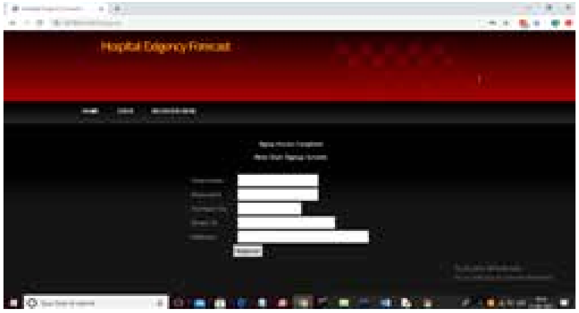
The dataset below contains client vitals; the final column has values of 0 or 1, where 0 indicates that the individual does not need admittance and 1 indicates that the patient does. We will be using this information to implement our project.

Hospital Exigency Forecast
Hospital Exigency Forecast
About Hospital Exigency Forecast

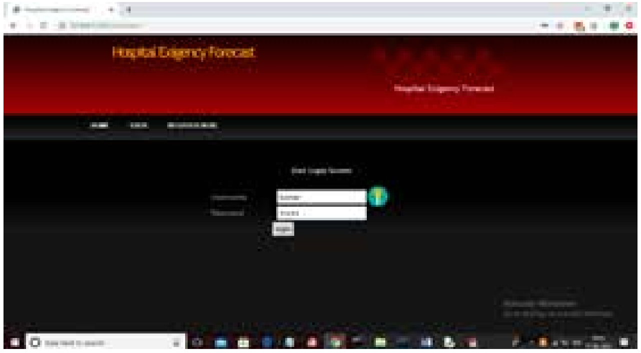
In above screen click on 'Register Here' link to get below signup screen



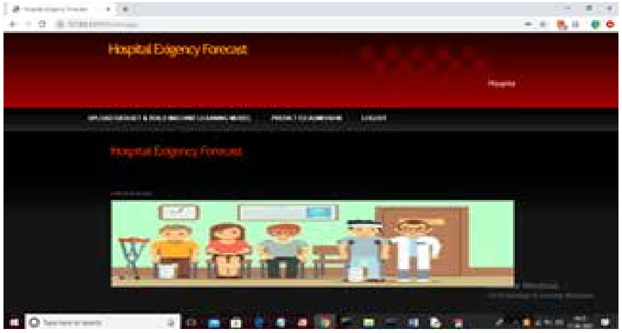
By entering their data on the first page and clicking the "Register" button, users may access the second screen.



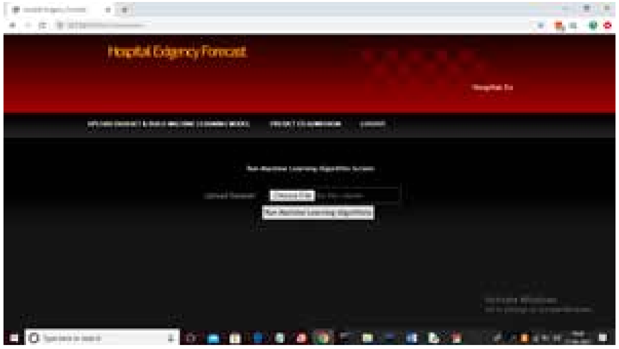
After completing the register process in the previous page, click the "User" link to access the login screen below.



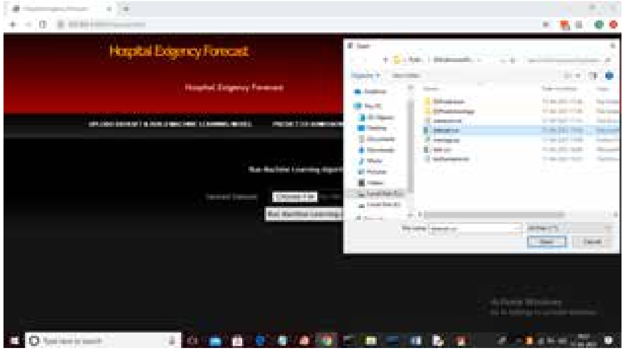
Once the user logs in, they will see the screen below.



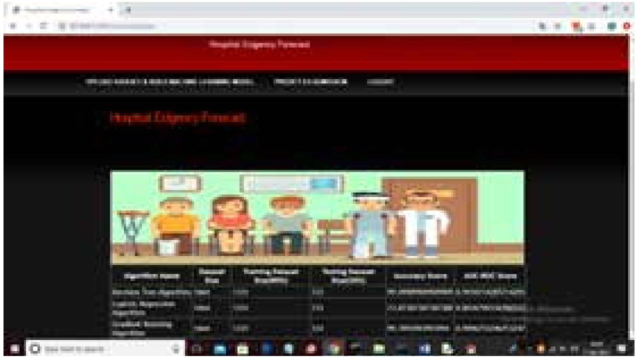
To access the next screen, go to the first screen and click on the "Upload Dataset & Build Machine Learning Model" link.



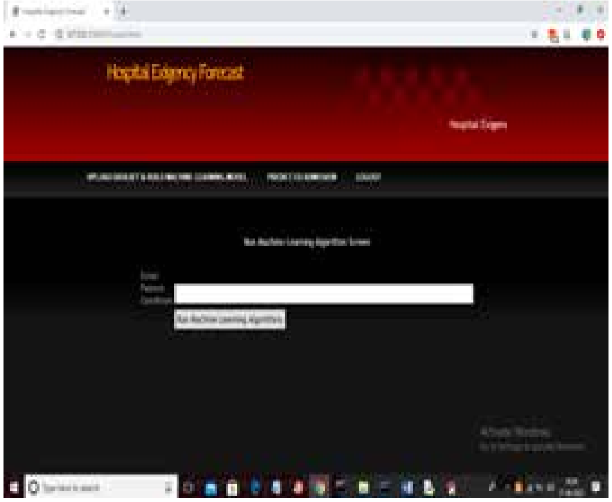
Click the "Choose File" button on the previous page. Then, upload the "dataset.csv" file. Finally, click the "Run Machine Learning Algorithms" button to create the model. The result will be shown below.



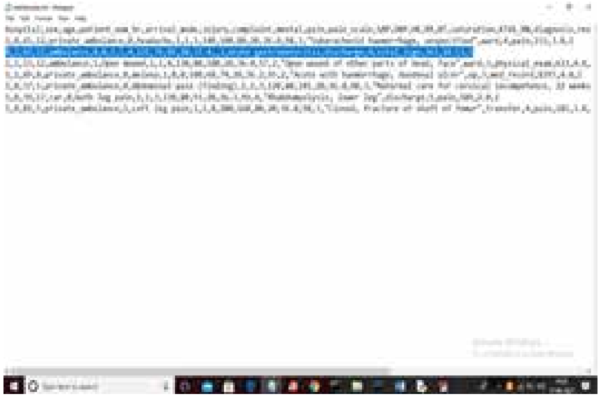
The following results will be shown after selecting and uploading the 'dataset.csv' data on the previous screen, clicking the 'Open' button, and finally clicking the Run button:



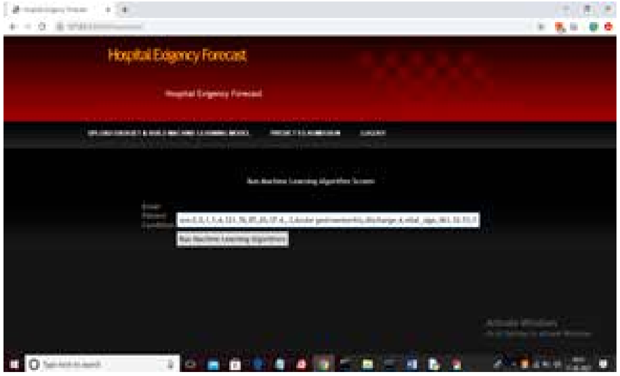
Now that the method train model is complete, click the "Predict ED Admission" link to acquire the following display; we ran three formulae by separating the dataset into train and test sets, and gradient boosting gave superior results in all algorithms.



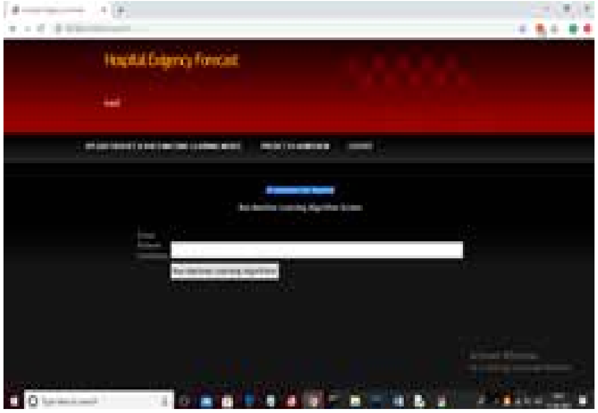
on order to get an admissions forecast, we may copy and paste certain vitals from the testSamples.txt files on the previous screen.



The above screen capture is from the testsamples.txt files. The following screen captures the process of me copying and pasting a document:



At this moment, you may get the forecast result below by clicking the "Run Artificial Intelligence Algorithm" button.



The predicted outcome is that admittance is not required at this time, and the examination is proceed with an additional document, as seen on the blue screen of the test.

CONCLUSION

To predict which OUH Depend on medical institution would accept a patient who is sent to the emergency department, we have presented a new method of training and standardising deep learning models in this paper. This prognosis will be useful in planning the patient's and other patients' rapid treatment and care while they are still in the emergency department. For each sort of individual ward, our version gets an AUC value between 0.60 and 0.78. To help the client implement more crucial elements for future individual wards, our version also provides a description of the predictions' source. Timely admittance to a medical institution and reduction of the moment to treatment are two goals that the authors hope this may help achieve. With much reduced congestion, the quality of care for patients remaining in the emergency department will undoubtedly improve. More generally, this approach could be useful for hospital resource forecasting and optimisation.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "HOSPITAL EXIGENCY FORECAST", which provided good facilities and support to accomplish our work. I sincerely thank our Chairman,

Director, Deans, Head of the Department, Department Of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. O. Karan, C. Bayraktar, H. Gümü¸skaya, and B. Karlık, "Diagnosing diabetes using neural networks on small mobile devices," Expert Syst. Appl., vol. 39, no. 1, pp. 54–60, 2012.

2. F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao, "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," in Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2017, pp. 1903–1911.

3. Z. Liang, G. Zhang, J. X. Huang, and Q. V. Hu, "Deep learning for healthcare decision making with emrs," in Proc. IEEE Int. Conf. Bioinformatics Biomed., 2014, pp. 556–559.

4. C. M. Bishop, Pattern Recognition and Machine Learning. Berlin, Germany: Springer, 2006.

5. Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in Proc. 26th Annu. Int. Conf. Mach. Learn, 2009, pp. 41–48.

6. V. A. Convertino et al., "Use of advanced machine-learning techniques for noninvasive monitoring of hemorrhage," J. Trauma Acute Care Surgery, vol. 71, no. 1, pp. S25–S32, 2011.

7. P. Gueth et al., "Machine learning-based patient specific prompt-gamma dose monitoring in proton therapy," Phys. Med. Biol., vol. 58, no. 13, p. 4563, 2013.

8. J. Labarère, P. Schuetz, B. Renaud, Y.-E. Claessens, W. Albrich, and B. Mueller, "Validation of a clinical prediction model for early admission to the intensive care unit of patients with pneumonia," Academic Emergency Med., vol. 19, no. 9, pp. 993–1003, 2012.

9. O. Hasan et al., "Hospital readmission in general medicine patients: A prediction model," J. General Internal Med., vol. 25, no. 3, pp. 211–219, 2010.

[10] A. K. Diehl, M. D. Morris, and S. A. Mannis, "Use of calendar and weather data to predict walk-in attendance." Southern Med. J., vol. 74, no. 6, pp. 709–712, 1981.

Human Action Recognition from Depth Maps and Postures Using Deep Learning

Palnati Sravani, Godugu Venkateshwarlu, Pendli Karthikeya
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

K Srujan Raju
Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ ksrujanraju@gmail.com

ABSTRACT

One of the active research areas in computer system vision for several settings, including safety and security monitoring, healthcare, and human computer interface, is human activity recognition. Several methods for human action recognition using depth, RGB (red, green, and blue), and skeletal system datasets have been published in recent years. Most of the methods that have been developed for activity categorization utilising skeletal system datasets have limitations when it comes to representing characteristics, complexity, and performance. The challenge of providing a trustworthy approach to human action discrimination using a skeleton dataset remains, however. For optimal feature extraction for precise activity categorization, we use depth images as input one and a suggested moving joints descriptor (MJD) as input two. The MJD depicts the change in position of the body's joints over time. In order to train CNN networks with various inputs, we are getting ready to deploy neural networks for rating combination. We proposed running the code on publicly available datasets such as MSRAction3D.

KEYWORDS: *MJD, MSRAction3D, CNN, Human action.*

INTRODUCTION

Not only is human activity recognition a challenging research problem, but it has also been a popular topic for quite some time due to its broad use in many different applications, such as smart security systems, human-robot communication, and home care systems [1]. The advancement of deep learning in recent years has led to widespread use of Convolutional Neural Networks (CNNs), which have achieved remarkable efficiency on monitoring, detection, classification, and detection tasks. CNNs are especially useful in computer vision and pattern recognition [2]. There are a few things to keep in mind while thinking about action acknowledgment. In order to keep up with technological advances and make it possible for people with disabilities to communicate with creators and understand their tasks through computer systems, standard methods of interaction are being developed as human-machine interaction grows in importance as a field of study in multimedia processing [3]. A lot of studies have attempted to use motion assessment to model and then identify people's behaviour. We focus on human behaviour assessment from video clips in this work. It's worth mentioning that a lot of information is hidden under gestures, fast movements, and walking speed. [1, 2] The expressive qualities provided by the two types of data have encouraged researchers in the field of human action recognition to focus on depth maps and body postures as representations of the action. Third The effectiveness of an activity recognition system relies on a detailed representation that provides unique characteristics of each task for categorization [5]. Because two movements could seem same from the front but distinct from the side, employing deepness map data for activity detection is still difficult for certain activities, leading to incorrect categorization [4]. When large obstructions are present, the depth maps captured by the depth cameras might be rather noisy, and the tracked joints' 3D settings can be completely off, leading to an increase in the actions' intraclass variances [5]. They propose "Skepxels," a spatial-temporal framework for skeletal series that makes full advantage of "regional" connections between joints by using CNN's 2D convolution bits. After using Skepxels to convert skeletal motion movies into salable images, they build a convolutional neural network (CNN) architecture for accurate human action identification using the produced images [9].

LITERATURE SURVEY

(Wang, Zhao-Xuan) [7] In order to validate the proposed technique, the following datasets were utilised: KTH and TJU, two well-known deepness datasets (MSR action 3-D and MSR daily task 3-D), and MV-TJU, a groundbreaking multiview multimodal dataset. Extensive experimental evidence in RGB and depth modalities demonstrates that this method outperforms popular, less expensive alternatives based on 2-D/3-D component models for a wide variety of human actions. I am C. Krishna Mohan: [8] Using a deep fully convolutional style, they suggested using action financial institution incorporates to identify human activities in videos. The similarities between the video and action bank movies are described by direct patterns called action bank characteristics. They are computed using an activity ban, which is a set of prescribed photographs. To Hiroshi Miki: [9] They provide an approach to recognition that examines the connection between human behaviour and items functions; the objective is to enhance our method by including human activities into dynamic item segmentation. To quote Ziaeefard and Maryam: [10] A state-of-the-art method for human action recognition based on normalized-polar histograms is proposed in this study. The process of amassing skeletal pictures was clarified.

It is a pattern of movement that uses range and angle. The colour cyan is used to emphases one of the most important aspects of this task. The symbol is formed by encircling the core with all of the skeletal system model frames. A two-level multi-class support vector machine (SVM) was used to classify people's behaviours; the model was trained using generic functions first, and then with salient features. Mejdi DALLEL: [11] They revealed a large-scale RGB+S keleton action acknowledgment dataset called "Industrial Human Being Action Recognition Dataset (InHARD)". We have 4804 unique samples of industrial activity in our collection, spread out among 38 films representing 14 various categories. In order to evaluate our dataset using the proposed metrics, they finish building an end-to-end regression classification LSTM network. Si Yang: [12] By combining a pattern recognition semantic network with an autoencoder, they were able to construct a novel semantic network. human activities recognition using deep neural networks. Confirmation of the design they suggested came from The benefits of the model were discovered via experiments. By comparing results, they produced a readily accessible model. Many noteworthy achievements: The researchers uncovered a novel approach of merging many frames of data into a single image. This method has several applications. a method for automatically extracting features of human actions using a deep neural network Hello, Chen Xu, Lei Zong, and HongLin Yuan! [12] The current state of RF fingerprint identification is flawed because it uses a preset resolution formula, which has a small range of potential uses and requires a lot of previous information. A convolutional semantic network (CNN) RF fingerprint identification approach would be ideal for handling these issues. Three main aspects of the research are RF fingerprint extraction, convolutional neural network architecture, and wireless transmitter identification and verification. There is no fraudulent usage of fingerprint info [13]. Advanced convolutional neural networks (DCNNs) outperform traditional approaches that rely on hand-crafted functions. A new FLD approach called an upgraded DCNN with photo scaling is available, although many CNN designs suffer from taken-care-of-scale photos. The complexity matrix is used as an efficiency indicator in FLD for the first time. The amounts of the hypothetical results using the LivDet 2011 and LivDet 2013 datasets provide further evidence that our technique is more efficient at discovery than others. [14] A highly accurate computerized latent fingerprint recognition system is required to compare concealed fingerprints found at crime scenes to a large database of referral prints and provide a list of potential friends. Their Convolutional Neural Network (ConvNet)-based automatic unexposed fingerprint recognition method achieves 64.7% accuracy with the NIST SD27 dataset and 75.3% accuracy with the WVU dataset when tested against a reference data source of 100,000 rolled prints.

EXISTING SYSTEM

Occasionally, it may be more convenient for customers to browse merchandise rather than wait in queue to pay, as the latter takes up even more of their time. Now we're building this system that may serve the customer in every way and also the shop owner by drawing inspiration from this situation that was common in all the shops. Consequently, we devise a technique that allows the customer to comprehend their expenses as they add things to the basket. If a consumer buys anything from this grocery shop basket, they can quickly charge it, which is the greatest and most practical example.

PROPOSED SYSTEM

This method introduces innovative advancements compared to the current purchasing mechanism. Providing a web-based, centralised invoicing system is the main objective of this project. Not only does it have automatic billing, but it also has certain unique capabilities. The word "Supermarket Basket" is new to us.

MODULES EXPLANATION

1. Data Collection: Collecting multivariate time series data from the sensing units of the watch and phone is the first stage. Testing is done on the sensors at a constant frequency of 30 Hz. Then, segmentation is done using the sliding window method, in which the instant collection is divided into consecutive windows of the cared-for period without inter-window gaps (Banos et al., 2014). Since the sliding window approach does not require the time series to be pre-processed, it is best suited for real-time applications.

2. Preprocessing: After that, filtering is carried out to eliminate noisy values and outliers from the accelerometer time series data in order to make sure that it would be suitable for the feature extraction stage. In this action, average filters (Sharma et al., 2008) or average filters (Thiemjarus, 2010) are the two main types of filters that are typically used. The type of sound that is produced here is similar to the sound of salt and pepper that can be found in pictures; that is, it is a sharp acceleration worth that occurs in a few isolated shots spaced throughout the time collection. In order to eliminate this kind of noise, an average filter of order 3 (window dimension) is applied.

3. Function Extraction: Below, one attribute vector per segment, or a fixed amount of attributes, will sum up each resultant part. Both the time and frequency domains yield the required properties. Considering this, a lot of activities are repetitive in nature; for example, walking and running are examples of a collection of motions that are performed occasionally. Since this rep frequency—also known as dominant regularity—is a detailed characteristic, it has been considered.

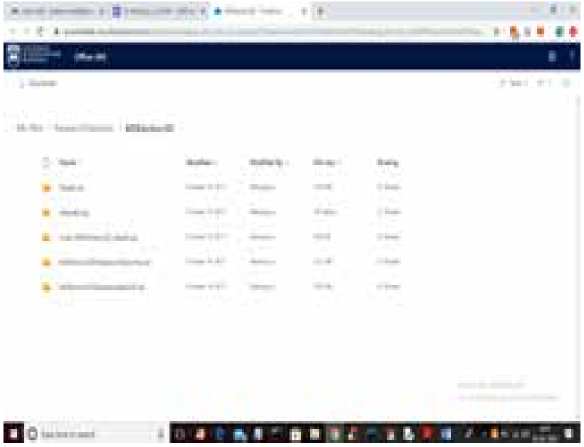
4. Standardization: Since some category formulas employ distance metrics, all characteristics should have the same range in order to allow for a fair comparison between them. This is because the moment domain name features are measured in (m/s 2), whilst the regular ones are measured in (Hz). Z-Score normalization, which is defined as xnew = (x − μ)/ σ, where μ and σ are the quality's mean and standard deviation, respectively, is used in this action to alter the credit to have absolutely no mean and unit variance (Gyllensten, 2010).

Human Activity Recognition from deepness maps and Poses utilizing Deep Discovering In this paper, author is using the CNN (Convolution Neural Networks) algorithm to identify human action as this formula will remove crucial attributes by filtering the exact same information multiple times in order to make the most of possibilities of precise activity category, CNN networks are educated with different inputs features which will certainly not happen in existing RGB Deepness algorithm which will get train on two features such as pictures and skeletal system information.

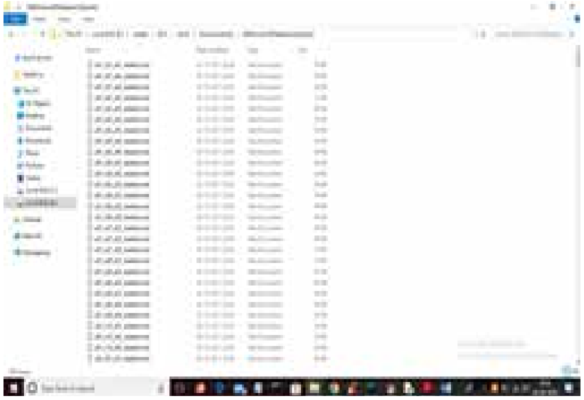
As existing formulas are not reliable, so writersmake use of the CNN formula which currently verifies its success in numerous fields such as image classification, weather and stock forecast etc

The MSRAction3D skeleton dataset, which contains 20 different actions like "high arm wave," "straight arm wave," "hammer," "hand catch," "ahead punch," "high throw," "draw x," "attract tick," "attract circle," "hand clap," "2 hand wave," "side-boxing," "bend," "ahead kick," "side kick," "jogging," "tennis swing," "tennis serve," "golf swing," "grab & throw."

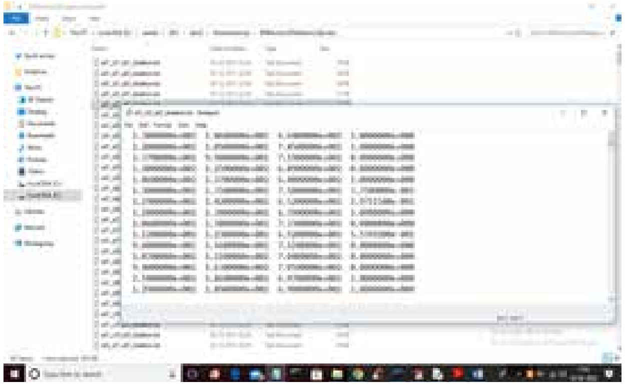
All this activities data is extracted from the MSRAction3D dataset and below are the display shots of that dataset.



Below is a screenshot of the dataset files. The dataset, named "MSRAction3DSkeleton(20joints)," was obtained from the aforementioned URL. Since it was recorded using DEPTH cameras, it will only record skeleton values.



All of the files in the dataset display basic information; for example, "a01" denotes activity 1 out of a possible twenty, "s01" is the subject ID, and "e01" is the circumstances ID. Following training, whenever we publish any kind of article, CNN will be able to use the aforementioned data to make predictions about future actions. As you can see in the screen capture below, every document will have skeleton values.



The following components were developed in order to carry out this assignment:

1) Following MSRAction3D For example, we may use this module to upload our activity dataset to an app.

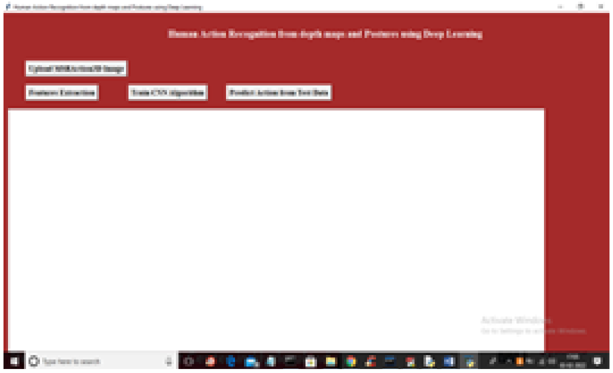
2) Includes Removal: This module will examine all files, delete functions (dataset values), and then visualise the results in a chart fashion. The action value will be considered the course title.

Third, we have the train convolutional neural network (CNN) algorithm, which takes in the extracted functions, trains it, and then uses test data from the experienced version to determine the correctness and complexity of the resulting matrix graph.

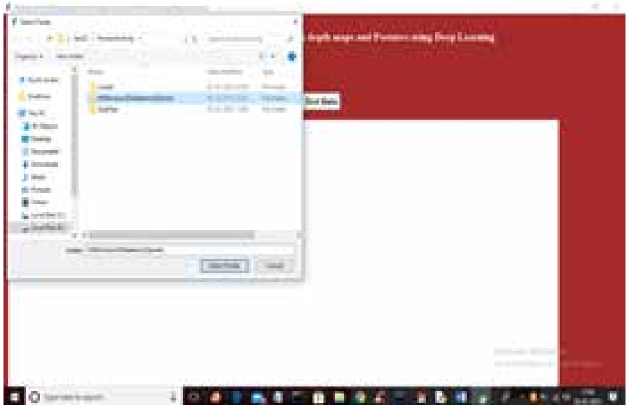
4. Anticipate Activity from Test Data: This component allows users to submit test data, which is subsequently processed by CNN. The network then checks the functionalities in the test documents and determines the activity based on those results.

RESULTS EXPLANATION

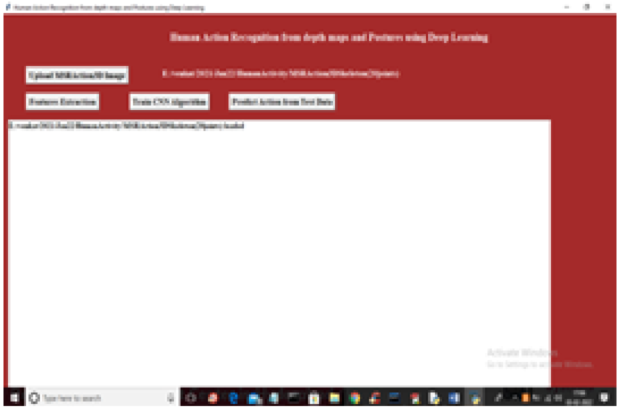
To run project double click on 'run.bat' file to get below screen


Human Action Recognition from depth maps and Postures using Deep Learning
Upload MSRAction3D Image
Features Extraction
Train CNN Algorithm
Predict Action from Test Data

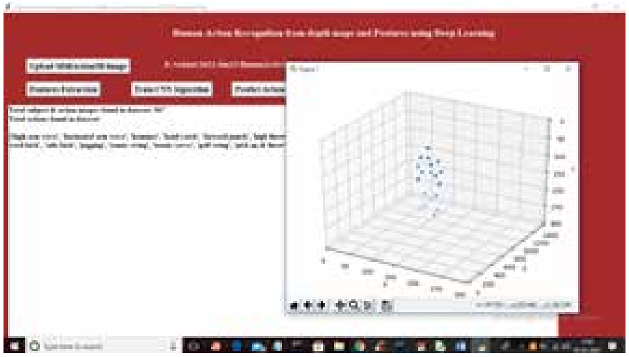
To upload the dataset, go to the first screen and click on the “Upload MSRAction3D Image” button. Then, you’ll see the second page.

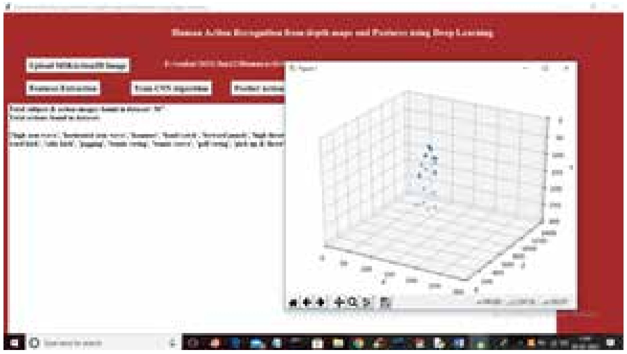


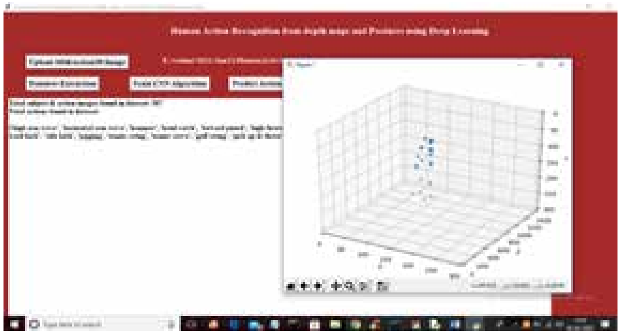
In the previous page, choose and submit the MSRACTION dataset. Then, to load the dataset, click on the “Select folder” button. This will bring up the following screen.


Human Action Recognition from depth maps and Postures using Deep Learning
Upload MSRAction3D Image
Features Extraction
Train CNN Algorithm
Predict Action from Test Data

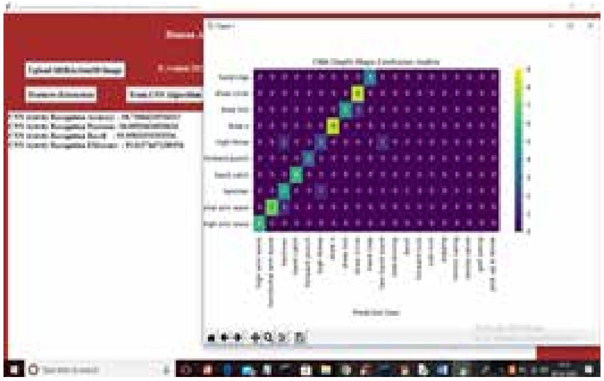
After reviewing all of the papers, building a functions array, and finally visualising one skeletal system image, I clicked the “Attributes Extraction” button in the aforementioned display dataset.


Human Action Recognition from depth maps and Postures using Deep Learning
Upload MSRAction3D Image
Features Extraction
Train CNN Algorithm


Human Action Recognition from depth maps and Postures using Deep Learning
Upload MSRAction3D Image
Features Extraction
Train CNN Algorithm

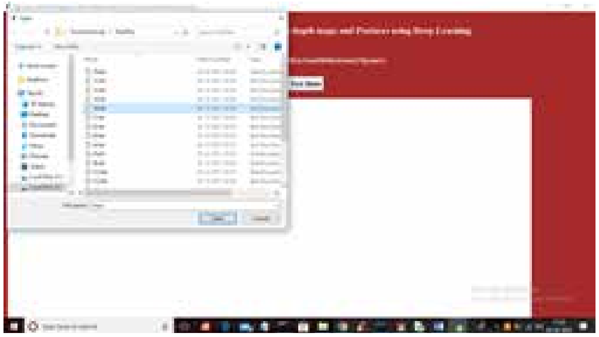

Human Action Recognition from depth maps and Postures using Deep Learning
Upload MSRAction3D Image
Features Extraction
Train CNN Algorithm

The dataset contains 567 documents, and I’ve shown 20 different motions, such “high arm wave” and “horizontal arm wave,” on the screen above. The skeletal system is moving in the above graphic, which represents the activity of a human in the dataset. You may see the skeletal system activity on the graph after you post. After you’ve examined the graph, click the “Train CNN Algorithm” button to begin training the CNN and get the results shown below.

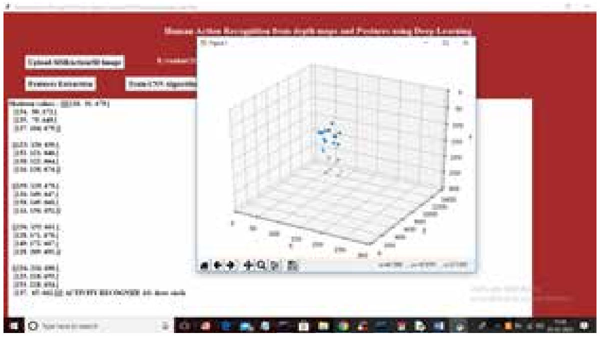


In the above display, we can see that CNN achieved an action recognition accuracy of 94%. The x-axis represents expected activity classes, and the y-axis represents initial classes. All of the class prediction values shown in the diagonal boxes are correct predictions, and very few values are out of the diagonal, indicating that CNN is very efficient and also achieves a 94% precision. Now closed above the chart

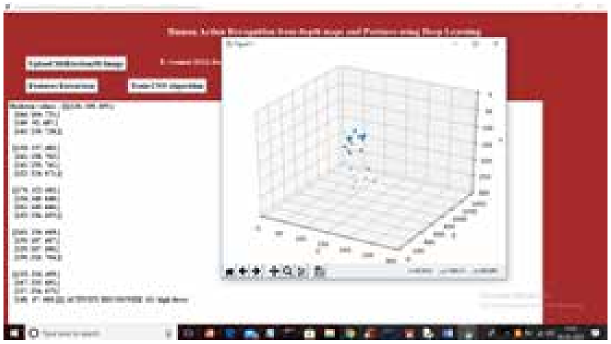
To upload test data documents, click the "Predict Activity from Examination Information" button. CNN will then acknowledge activity based on that test file information.



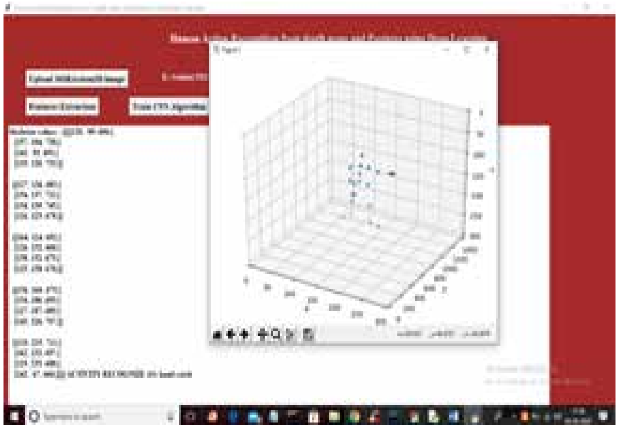
You may receive the following activity acknowledgment result by selecting the "14. txt" file in the previous screen, clicking the "Open" button, and then packing the examination data.



All of the values in square brackets in the previous presentation are skeleton values; the outcome is "Activity Acknowledged as 'draw circle'" on the final line, and the skeleton's activity is seen in the chart.



In above screen,action is recognized as 'high throw'.



The aforementioned on-screen gesture is called a "hand catch," and it works just like any other file upload and testing tool.

CONCLUSION

It has been proposed to use deep convolutional neural networks to recognise human actions based on depth maps and posture data. By combining the outputs of the three convolutional neural network (CNN) networks, we were able to optimise attribute extraction utilising two activity representations and three channels from convolutional semantic networks. The method has been tested on three publicly available, industry-standard datasets. When compared to state-of-the-art methods that rely on deepness or posture data, the three datasets' category accuracy is light years ahead. The claim made in this work is that various representations of actions provide different clues. Unlike the other depictions, one of them has action functions.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "HUMAN ACTION RECOGNITION FROM DEPTH MAPS AND POSTURES USING DEEP LEARNING", which gave us excellent resources and assistance to complete our assignment. I want to express my gratitude to our chairman, director, deans, head of the department, computer science and engineering department, guide, and teaching and non-teaching faculty members for their insightful advice and direction in all facets of our work.

REFERENCES

1. C. Schuldt, I. Laptev, and B. Caputo, “Recognizing human actions: A local SVM approach,” in Proc. IEEE 17th Int. Conf. Pattern Recognit., vol. 3. Cambridge, U.K., Aug. 2004, pp. 32–36.
2. J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li, “Hierarchical spatio-temporal context modeling for action recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Miami, FL, USA, Jun. 2009, pp. 2004–2011.
3. I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Anchorage, AK, USA, Jun. 2008, pp. 1–8.
4. W. Li, Z. Zhang, and Z. Liu, “Action recognition based on a bag of 3D points,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops, San Francisco, CA, USA, Jun. 2010, pp. 9–14.
5. J. Wang, Z. Liu, Y. Wu, and J. Yuan, “Mining actionlet ensemble for action recognition with depth cameras,” in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 1290–1297, IEEE, 2012.
6. Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, “Skeleton optical spectra based action recognition using convolutional neural networks,” arXiv preprint arXiv:1703.03492, 2016.
7. Liu, An-An; Su, Yu-Ting; Jia, Ping-Ping; Gao, Zan; Hao, Tong; Yang, Zhao-Xuan (2015). Multipe/Single-View Human Action Recognition via Part-Induced Multitask Structural Learning. IEEE Transactions on Cybernetics, 45(6), 1194–1208. doi:10.1109/tcyb.2014.2347057
8. Ijjina, Earnest Paul; Mohan, C. Krishna (2014). [IEEE 2014 13th International Conference on Machine Learning and Applications (ICMLA) - Detroit, MI, USA (2014.12.3-2014.12.6)] 2014 13th International Conference on Machine Learning and Applications - Human Action Recognition Based on Recognition of Linear Patterns in Action Bank Features Using Convolutional Neural Networks. , (), 178–182. doi:10.1109/icmla.2014.33
9. Miki, Hiroshi; Kojima, Atsuhiro; Kise, Koichi (2008). [IEEE 2008 Second International Conference on Future Generation Communication and Networking (FGCN) - Hainan, China (2008.12.13-2008.12.15)] 2008 Second International Conference on Future Generation Communication and Networking - Environment Recognition Based on Human Actions Using Probability Networks. , (), 441–446. doi:10.1109/fgcn.2008.62.
10. Y. Kim, J. Chen, M.-C. Chang, X. Wang, E. M. Provost, and S. Lyu, “Modeling transition patterns between events for temporal human action segmentation and classification,” in Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on, vol. 1. IEEE, 2015, pp. 1–8.
11. C. Chen, R. Jafari, and N. Kehtarnavaz, “Action recognition from depth sequences using depth motion mapsbased local binary patterns,” in Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on. IEEE, 2015, pp. 1092–1099.
12. J. Koushik, “Understanding convolutional neural networks,” arXiv preprint arXiv:1605.09081, 2016.
13. J.Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, and G. Wang, “Recent advances in convolutional neural networks,” arXiv preprint arXiv:1512.07108, 2015.
14. E. Park, X. Han, T. L. Berg, and A. C. Berg, “Combining multiple sources of knowledge in deep cnns for action recognition,” in Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on. IEEE, 2016, pp. 1–8.
15. J. Zhang, W. Li, P. O. Ogunbona, P. Wang, and C. Tang, “Rgb-d-based action recognition datasets: A survey,” Pattern Recognition, vol. 60, pp. 86–105, 2016.

Text and Image Plagiarism Detection

Kondabala Bindhu, Chakali Virupakshi
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

B Pooja
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
🖂 pooja.cse@cmrtc.ac.in

ABSTRACT

Researchers are scrutinizing instances of plagiarism at an unprecedented rate. Due to online circumstances and the capacity to execute complex and intelligent searches quickly, research has really suffered serious harm. Visuals are ignored by plagiarism detection tools that concentrate on text. In contrast, images play an essential role in conveying the vast amounts of data presented in academic papers and other types of academic writing. Due to the large amount of information included in flowcharts and the large number of photographs used in computer-generated texts, plagiarism might potentially arise. Our goal is to find out how many instances of picture plagiarism there are in a work by using the Pie Chart Design.

KEYWORDS: *Text, Plagiarism, Image.*

INTRODUCTION

In academic circles, the issue of plagiarism is often debated. Plagiarism is the act of passing off the ideas or work of another as one's own without giving proper credit [1]. It is, at its core, a repackaging of data that is up-to-date. Plagiarism, according to S. Hannabuss, "is the act of copying or manipulating someone else's innovation or concept without consent and providing it as one's very own" [2]. Thanks to the proliferation of internet access, a wealth of previously unavailable information is now freely accessible online. The Internet has evolved into a vast repository for information [3]. Since individuals can easily access the information they need from the internet, there is no need for them to produce their own text records. The ease with which a plagiarist may find an acceptable text fragment to reproduce is making plagiarism detection a more pertinent issue [4]. However, it gets more difficult to precisely identify plagiarised passages as the number of diverse sources increases. [5] Plagiarism is a common occurrence in many fields, including national politics, academia, the media, and science. Because document-to-document comparison formulae cannot be utilised when there is no referral collection available or when all the possible duplicate sources are not supplied, this method of plagiarism detection is especially useful in such instances. Other kinds of plagiarism include text manipulation and others [6]. Similarly, there are a variety of methods that may be used to detect instances of plagiarism. At this time, text-control-based system applications aren't good enough for practical use. So, we came up with a fresh and easy method to detect plagiarism in text sets by using a maker learning strategy [7]. To find the plagiarised text series, we use our plagiarism detection threshold value, which is a percentage based on the number of words that are similar between the two papers [8].

PROBLEM DEFINITION

There has never been a controlled analytic environment dedicated to the detection of plagiarism before the corpus and the methodologies were developed. Publishing a collection of real plagiarism instances for evaluation purposes is not practical due to the ineffective unionization of these cases, unlike other tasks in natural language processing and information retrieval. Hence, the analyses presented in the literature are not only unmatched, but sometimes even impossible to replicate. Here, we provide a recently built massive library of synthetic plagiarism as well as novel discovery performance stages tailored to the examination of plagiarism detection algorithms.

Job Objective

We set out to create a corpus that would reveal the kinds of plagiarism that students commit in an academic context, as accurately as possible, so that we could improve and test plagiarism detection algorithms.

EXISTING SYSTEM

If the source and assumed pictures haven't been rotated by a wide margin, the current technique may be enough to detect photo copying. However, if there have been rotational modifications, the method will not work. Even if the picture is rotated, the proposed method will still identify instances of plagiarism or attacks, including rotational modification. On top of that, the current algorithm isn't good at detecting plagiarism for various image formats. By making use of adaptive limit settings, the suggested approach will guarantee that. With each iteration of improvement, the algorithm significantly reduces the search field, ensuring that the photo matching time is much smaller.

PROPOSED SYSTEM

To determine if a user-submitted image is plagiarised, the Suggested Text and Image of Pictures plagiarism detector will look at the user-submitted image. Next, we'd use the corpus approach to generate the image's Phash value. At the moment, the supplied photo will be checked for plagiarism using the pictures in the local database. Images are stored in the data source together with their associated hash values. One of the steps that the plagiarism detection engine will take to identify instances of plagiarism is to compare the hash values of the input photos with those in the database. Finally, results will be shown according to what the detecting engine has managed to do. A text file was also found using the corpus formula.

MODULES

The Registration Process for New Users

To begin, the user must sign up for the application.A username and password are useful for accessing the application.

Second, the user will enter their username and password to access the application.

You may load all the files from the corpus folder by creating an Upload Source File Folder and then clicking the Upload Source Files link.

Upload materials that may be harmful

Please load the suspicious file and provide the outcome. When the user selects "Upload Suspicious Files," the programme will run the uploaded file. The LCS score is 1.0, which indicates a perfect match with the corpus file, allowing for the detection of plagiarism. Moreover, you may insert any text file and get the results.

Put the Original Image Here

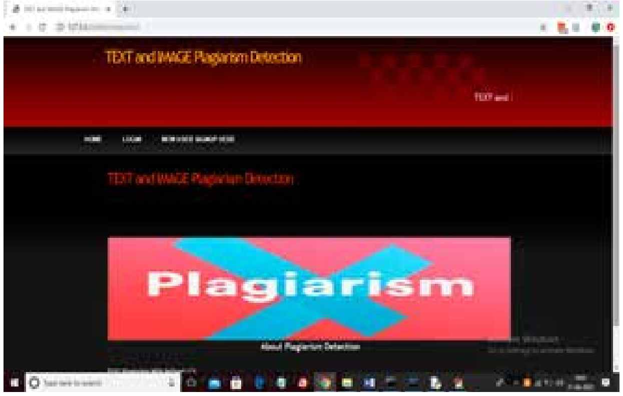
In this module, we will compute and store the histograms of all the photos in the database in an array. Whenever we upload fresh test images, we will compare the two histograms.

Upload Questionable Image

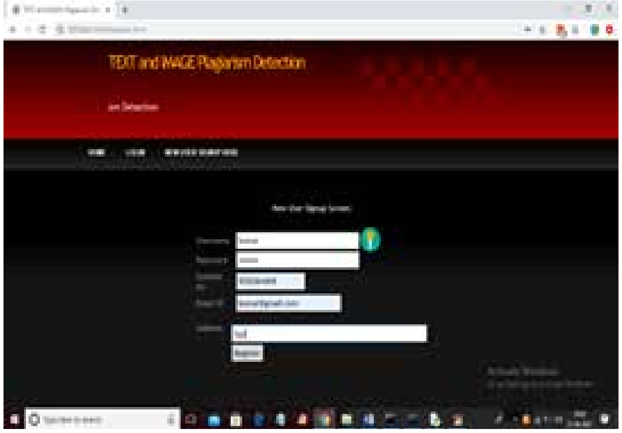
We created a histogram for both the database image and the uploaded image; because there is no match, we can conclude that no plagiarism has occurred. Now you can upload the picture from the "images" folder and view the outcome; the histogram pixel matching score is 15173 out of 40000 pixels, meaning the image is not plagiarised. The histograms of the original and uploaded images are identical, indicating copying. The result is below. The above result is considered plagiarism since the histogram matching score is 40,000, indicating that all pixels matched.

OPERATION

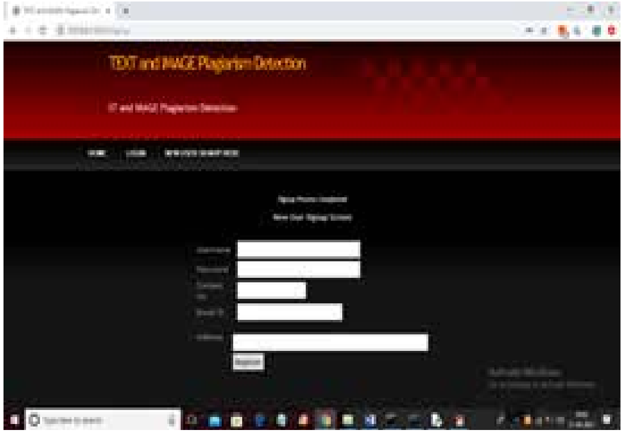
Installing python 3.7, DJANGO server, and deploying code to the server are the necessary steps to execute the project. Once done, run the code from your browser to see the screen below.


TEXT and IMAGE Plagiarism Detection
TEXT and IMAGE Plagiarism Detection
Plagiarism
About Plagiarism Detection

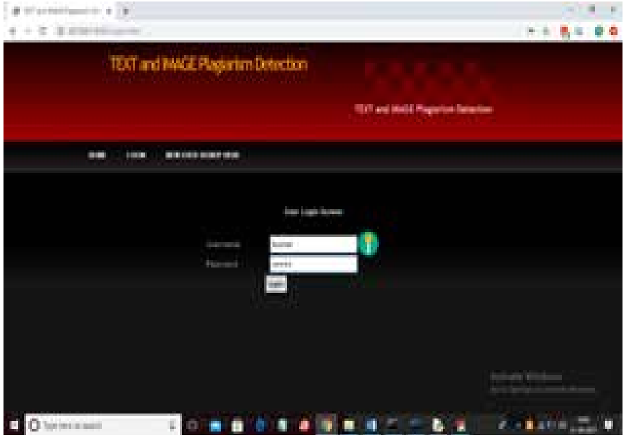
Click the “New User Signup Here” link on the previous page to access the screen below.



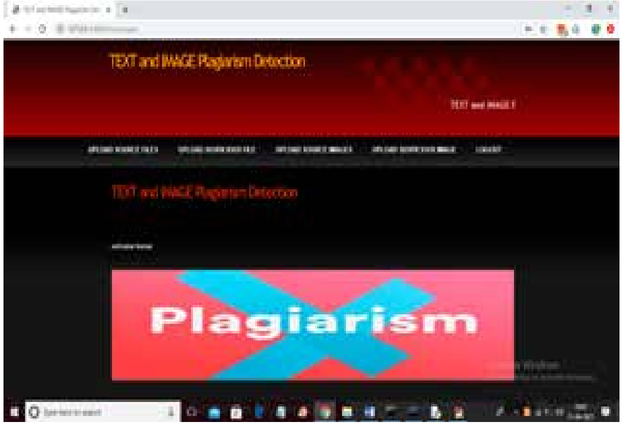
After entering their details, users may go to the next page by clicking the “Register” button on the previous screen.



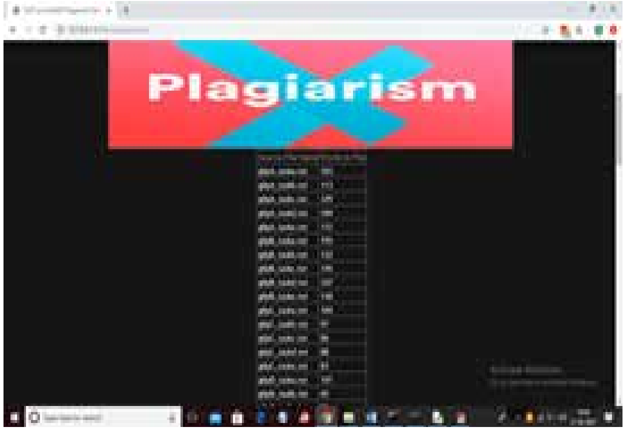
After completing the register procedure, customers may see the following display by clicking the “Login” option.


TEXT and IMAGE Plagiarism Detection
TEXT and IMAGE Plagiarism Detection
User Login Screen
Login

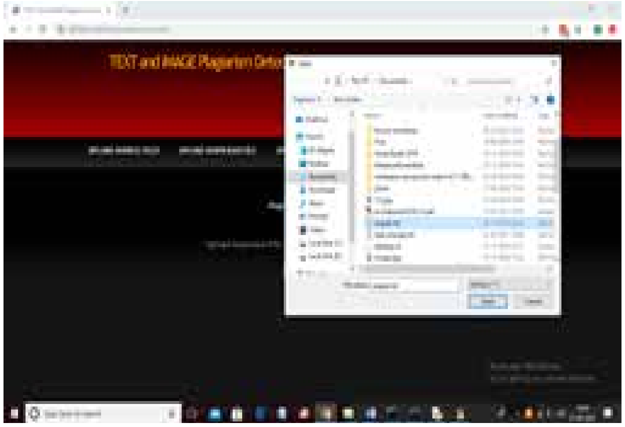
The user logs in on the top screen, and then they click the button to go to the bottom screen.


TEXT and IMAGE Plagiarism Detection
TEXT and IMAGE Plagiarism Detection
Plagiarism

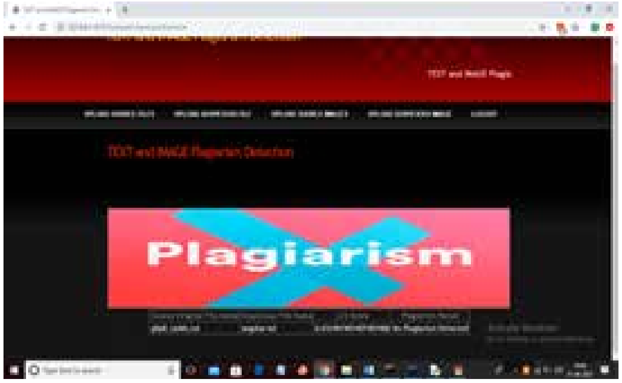
To pack all files from the corpus folder, click the “Upload Source Files” online link in the preceding page.



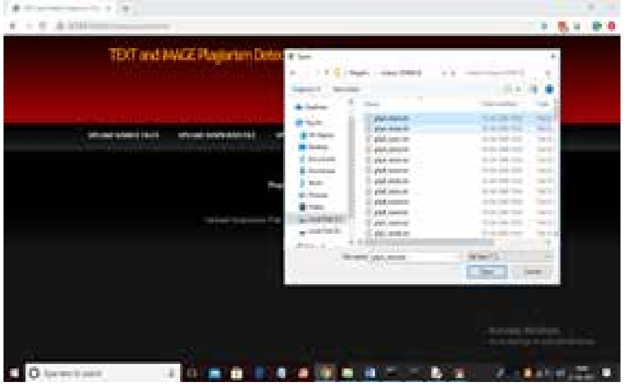
After you’ve finished filling out the form, click the “Upload Suspicious Documents” button. go to a large number of questionable files and get the outcome.



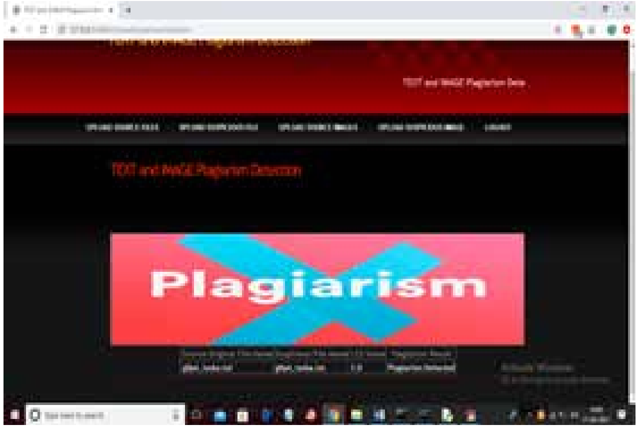
To retrieve the results shown below, choose the "angular.txt" file in the previous screen, click the "Open" button, and then click the "Examine Plagiarism" button.



No plagiarism was found when we compared the angular.txt texts to the g) pB_taskb.txt corpus file; the similarity rating was just 0.03. You may now submit any data type to the corpus and view the results.



To get the results shown below, I am choosing and uploading the first file in the preceding screen, and then I click the switch.



A perfect match with the corpus data (a 1.0 LCS score) means that plagiarism has been identified; moreover, you may input any kind of text file and get the results. Presently, to upload all the photographs from the 'pictures' folder, click on the 'Upload Resource Photos' web page.

CONCLUSION

The corpus was used with great success in the First International Competition on Plagiarism Detection and is the first standardised corpus devoted to the assessment of automated plagiarism detection. Our hope is that future studies on plagiarism detection may benefit from our corpus and performance metrics. An upgraded version of the database is now in development.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "TEXT AND IMAGE PLAGIARISM DETECTION", which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department Of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. Imam Much Ibnu Subroto and Ali Selamat, "Plagiarism Detection through Internet using Hybrid Artificial Neural Network and Support Vectors Machine," TELKOMNIKA, Vol.12, No.1, March 2014, pp. 209-218.

2. Upul Bandara and Gamini Wijayrathna ,"Detection of Source Code Plagiarism Using Machine Learning Approach," International Journal of Computer Theory and Engineering, Vol. 4, No. 5, October 2012, pp.674-678.

3. Salha Alzahrani, Naomie Salim, Ajith Abraham, and Vasile Palade," iPlag: Intelligent Plagiarism Reasoner in Scientific Publications," IEEE World Congress on Information and Communication Technologies, 2011.

4. Barrón Cedeño, A., & Rosso, "On automatic plagiarism detection based on n-grams comparison," In Advances in Information Retrieval, Vol. 5478. Lecture Notes in Computer Science, pp. 696– 700, Springer.

5. Ahmad Gull Liaqat and Aijaz Ahmad, "Plagiarism Detection in Java Code," Degree Project, Linnaeus University, June 2011, pp. 1-7.

6. A Selamat, IMI Subroto and Choon-Ching Ng, "Arabic Script Web Page Language Identification Using HybridKNN Method," International Journal of Computational Intelligence and Applications, 2009, pp. 315- 343.

7. Michael Tschuggnall and Gunther Specht , "Detecting Plagiarism in Text Documents through Grammar Analysis of Authors," pp. 241-255.

8. Bill B. Wang, R I. (Bob) McKay, Hussein A. Abbass and Michael Barlow, "Learning Text Classifier using the Domain Concept Hierarchy," ACT 2600, pp. 1-5

9. Francisco R., Antonio G., Santiago R., Jose L., Pedraza M., and Manuel N., ―Detection of Plagiarism in Programming Assignments,‖ IEEE Transactions on Education, vol. 51, No.2, pp.174-183, 2008.

Machine Learning for Web Vulnerability Detection

Alampally Santhosha, Yapakayala Rajendar, Bannuru Sahithi
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

Najeema Afrin
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ Najeema.afrin@gmail.com

ABSTRACT

We provide a method that uses artificial intelligence to find internet application vulnerabilities in this assignment. The wide use of customised development approaches and the variety of web apps make them particularly difficult to assess. ML's ability to use manually categorised data to incorporate human knowledge of web application semiotics into automated analytical devices makes it very helpful for web application safety and security. Mitch, the first machine learning solution for the concealed detection of Cross-Site Request Imitation (CSRF) vulnerabilities, incorporates our methodology. We were able to locate 35 new CSRFs on 20 significant locations and 3 new CSRFs on manufacturing software because of Mitch.

KEYWORDS: *Web application, ML, Widespread, CSRF.*

INTRODUCTION

The majority of people's first point of contact with possibly harmful facts and capabilities these days is by utilising programmes on the internet. We use them for a lot of common things, such filing our taxes, seeing the results of our doctors' exams, making money transfers, and communicating with our loved ones [1]. Worst case scenario: this shows that internet apps are attractive targets for bad actors that want to steal money, acquire unauthorised access to personal information, or embarrass their intended victims. The security of web applications is often thought of as a challenging [2] in

A number of factors contribute to this, including the increasing diversity and complexity of the online system and the proliferation of unrestricted scripting languages that provide unclear protection guarantees and are difficult to rectify. In such a situation, many individuals use black-box susceptibility detection techniques. Black-box techniques function at the level of HTTP website traffic, which includes both requests and activities, as opposed to white-box techniques, which depend on having access to the source code of an online service [3]. A language-agnostic vulnerability detection approach is provided by this narrow viewpoint, which may miss out on certain essential insights. This method is effective because it provides a normal user interface for the most extensive collection of online apps without getting into the subtleties of scripting languages. [4] At first glance, this seems to be competent; nevertheless, previous studies have shown that this kind of analysis is everything from unimportant. A major obstacle to effective vulnerability finding is teaching automated devices the semantics of web applications. In visual form: One common kind of web attack is known as cross-site request forgery (CSRF), and it involves tricking an authorised user into making malicious HTTP requests to an unprotected web app. The basic premise of cross-site request forgery (CSRF) is that an attacker may trick an internet app into thinking they are sending a genuine request by using the user's browser to send fake requests that seem to be permitted inquiries. [5]

An example of a typical CSRF attack might look like this:

1) Alice logs in to her favourite social media site or any other harmless yet vulnerable web app. You should picture the following scenario: 1) While Alice is surfing

the internet, she comes across an ad for a website that could pose a threat due to the information it carries.

2) The advertisement asks the user to “like” a political event or send out a cross-site demand to social networks using HTML or JavaScript.

(3) Any subsequent request to the online application is immediately associated with the session cookie by the web browser.

Due to the need of Alice’s cookies, it has been fine-tuned for usage in the context of verifying her social media accounts. Net survey results could be impacted if the harmful commercial convinces Alice to “like” the intended political event in this manner. Web developers must take specific precautions to prevent cross-site request forgery (CSRF), although it is not necessary for the adversary to be eavesdropping or to tailor responses to individual customers. [7] As long as adding extra human participation does not significantly impact functionality, it is feasible to prevent cross-site requests from being undetected by requiring re-authentication or by using single passwords/CAPTCHAs. Still, most people prefer automated defences; for example, new web applications are required to utilise the SameSite cookie feature to prevent cookie device on cross-site requests, which eliminates the cross-site demand imitation (CSRF) resource [8]. Unfortunately, this protection is not yet widely implemented, however the following measures are often used by current internet apps to decrease cross-site demand:

first, determining the value of the Referrer and Beginning HTTP request headers, which reveal the URL of the originating website; 2) testing the accessibility of non-standard HTTP request headers (such as X-Requested-With) that cannot be defined in a cross-site context; 3) exploring sensitive types for unusual anti-CSRF symbols that the web server may have provided [9]

A recent study weighed the benefits and drawbacks of these different services. While there are benefits to each of the three choices, they all share the necessity for thorough safety assessments. To provide complete security while maintaining the user experience, it is necessary to pair tokens with just the security-sensitive HTTP queries. [10]

The aforementioned attack can be avoided by using a token to protect a “like” button, although having a token on the social network’s homepage is not advised. This is due to the possibility that it may reject valid cross-site requests, such as those sent by network-indexing search engines. Net programmers frequently handle the difficult problem of gradually determining the “ideal” location of anti-CSRF defenses. Although there is automatic assistance in current frameworks for building web applications for this, even highly rated websites may be attacked via cross-site request forgery (CSRF). As a result, reliable CSRF finding technologies are in high demand. How can we provide automated device support for CSRF detection, however, in the absence of a mechanism for automatically identifying which HTTP requests are most sensitive to security? beside one another, not apart.

EXISTING SYSTEM

It is well-known that safeguarding web applications inside the current system is no easy feat. Reasons for this include the internet’s inherent complexity and variety as well as the widespread use of scripting languages that lack discipline, provide dubious security guarantees, and are not amenable to fixed evaluation. This limited viewpoint may not yield all the necessary insights, but it does provide the critical advantage of using a susceptibility finding technique independent of language. This method offers a uniform user interface for the greatest variety of internet apps and does away with the requirement to comprehend scripting languages.

PROPOSED SYSTEM

Cross-Site Request Forgery (CSRF) is a common tactic used by attackers to fool users into sending malicious HTTP requests to untrusted web applications while they are authorized. A critical security flaw in cross-site request forgery (CSRF) attacks arises when a user’s web browser is used to send malicious queries to an online application; these requests might be similar to the customer’s intended benign ones. Under the CSRF, the attacker is not required to change or intercept the victim’s requests or feedback all that is needed is for the target to visit the attacker’s website, where the assault is then launched. Thus, any rogue website on the Internet may take advantage of CSRF vulnerabilities.

MODULES DESCRIPTION

User

Anyone may sign up for the initial. He needed a working customer email and phone number for future correspondence when he signed up. The client may be activated by the admin after the customer has registered. After the administrator has enabled the client, the user will be able to access our system. Client is able to do data pre-processing. Running site name was the first request. The client may see the csrfs by using that website. The user may get all associated csrfs and created algorithm names with the help of the bolt device. The outcome will most definitely be saved as JSON data. The results of the Mitch dataset will be available to the person later on. In order to find out how to get a blog post method, the Mitch dataset was examined. The browser will show the outcome.

Administrator: Administrator may log in using his credentials. The customers may be activated the moment he logs in. The customer-only login feature in our apps has been activated. With the Mitch Dataset, the administrator may build up the project's training and testing datasets. The user may access all the links related to the csrf token on the admin's website. The administrator also has the option to see data related to the GET method and the blog post method that was run from the dataset.

False Negatives and inaccurate Positives: When Mitch returns an immutable prospect CSRF, it creates an inaccurate favourable. Although the process is tedious and time-consuming, this is something that can be easily detected by manual screening. In most cases, knowing all the CSRF vulnerabilities on the tested websites is impractical, hence it is not possible to properly detect when Mitch generates an inaccurate negative. We monitor all the sensitive queries made by the ML classifier that was placed in Mitch and focus our manual testing efforts on these cases in order to get close to this crucial component. This is a wise choice to streamline the investigation, as we have previously demonstrated that the classifier operates effectively when employing basic validity actions.

Using AI as a Classifier:

Utilizing a dataset of about 6,000 HTTP requests from active websites that were gathered and categorized by two human experts, Mitch trained his ML classifier. The classifier's attribute area X records 49 different dimensions, each of which corresponds to a different business or residential property that HTTP wants. These may be neatly organised into the following categories.

This set of mathematical operations:

Total number of parameters: the range of possible criteria;

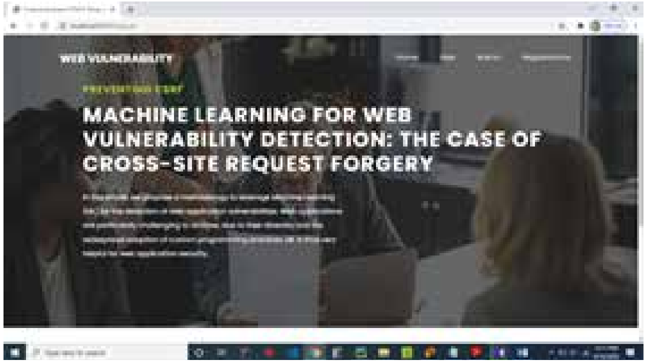
Number of Bools: the set of all demand parameters that may be expressed as a boolean;

The NumOfIds is the number of demand parameters associated with a common identifier in our dataset, which is a hexadecimal string.

Num Of Blobs: the total amount of request parameters (a non-identifying string) connected to a blob;

ReqLen is the overall character count of the request, inclusive of specification names and values.

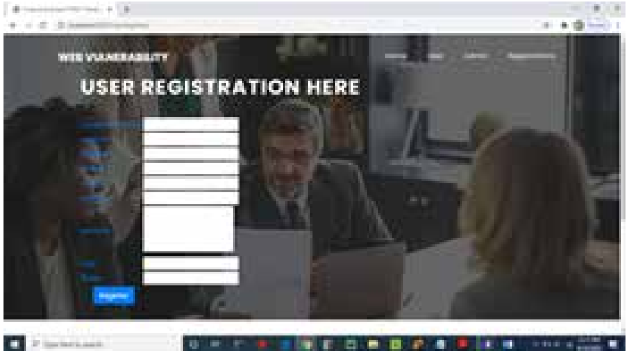
Home page



WEB VULNERABILITY

MACHINE LEARNING FOR WEB VULNERABILITY DETECTION: THE CASE OF CROSS-SITE REQUEST FORGERY

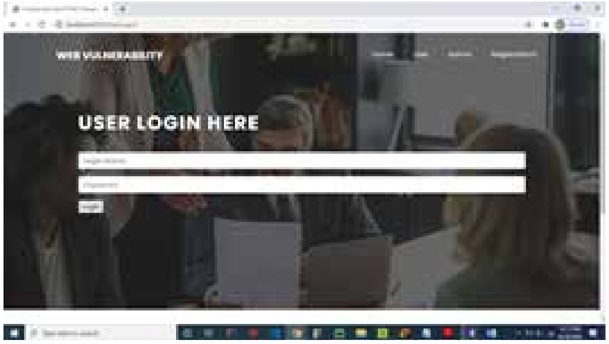
User Registration Form



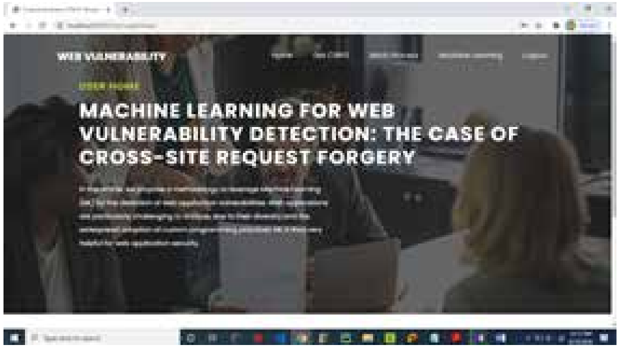
WEB VULNERABILITY

USER REGISTRATION HERE

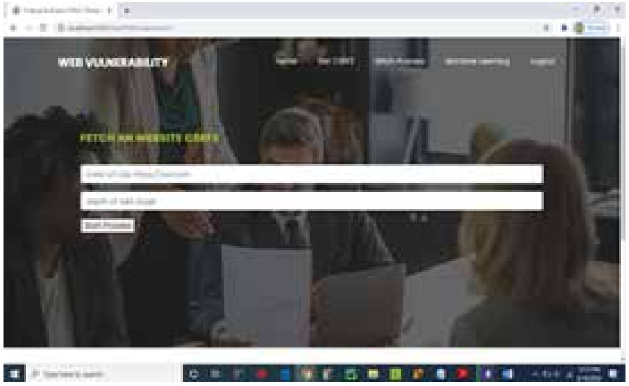
User Login Form



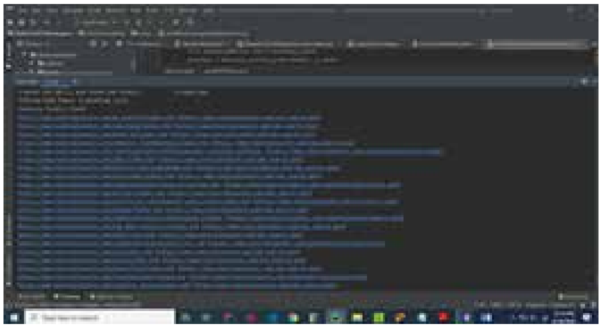
User Home


WEB VULNERABILITY
USER HOME
MACHINE LEARNING FOR WEB VULNERABILITY DETECTION: THE CASE OF CROSS-SITE REQUEST FORGERY

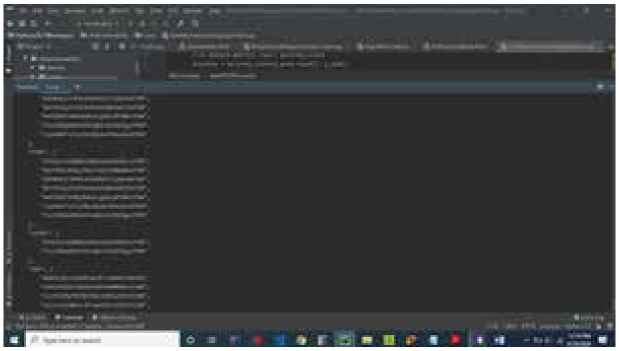
Getting website csrfs



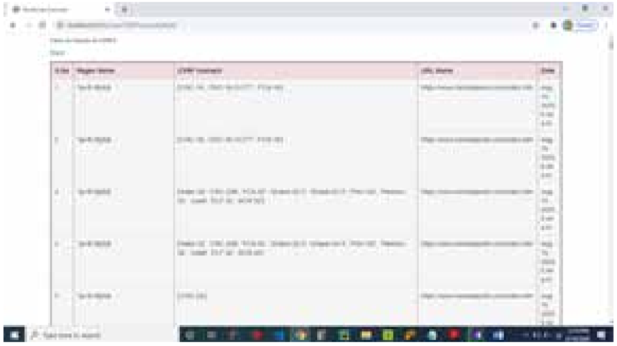
Scanning urls



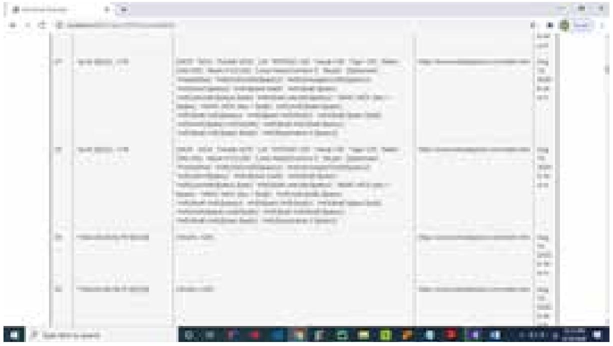
CSRF token



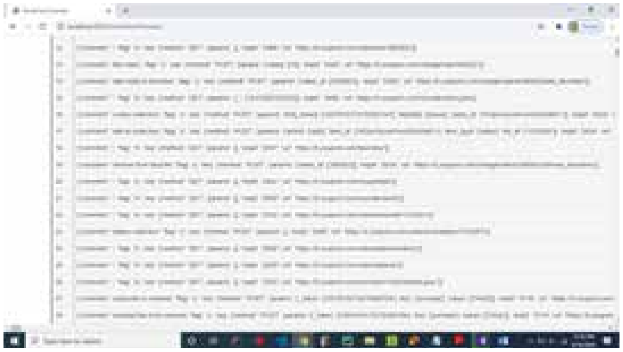
Given website csrf results



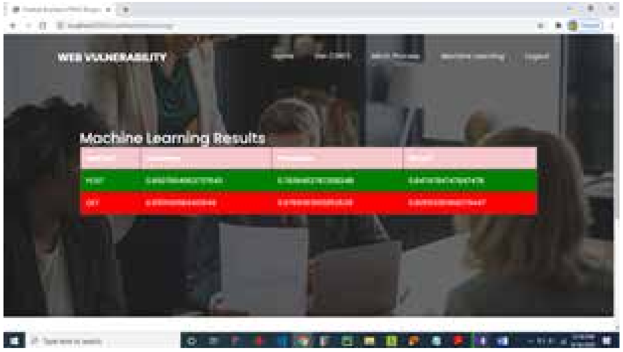
MD5 Token



Mitch Detected sites



Machine Learning Results



Admin Login

CONCLUSION

The variety of web apps and the extensive use of customised programming approaches make them especially challenging to assess. ML's ability to use manually-identified data to convey human knowledge of web application semantics to automated assessment tools makes it very applicable to the online environment. We were able to confirm this claim by developing and testing Mitch, the first machine learning solution for the covert identification of cross-site request forgeries. We sincerely hope that other researchers will find our methodology helpful in their efforts to find different kinds of vulnerabilities in online applications.

ACKNOWLEDGMENT

We appreciate CMR Technical Campus's sponsorship of this paper, which is named. "MACHINE LEARNING FOR WEB VULNERABILITY DETECTION", which gave us excellent resources and assistance to complete our assignment. I want to express my gratitude to our chairman, director, deans, head of the department, computer science and engineering department, guide, and teaching and non-teaching faculty members for their insightful advice and direction in all facets of our work.

REFERENCES

1. Stefano Calzavara, Riccardo Focardi, Marco Squarcina, and Mauro Tempesta. Surviving the web: A journey into web session security. ACM Comput. Surv., 50(1):13:1–13:34, 2017.

2. Avinash Sudhodanan, Roberto Carbone, Luca Compagna, Nicolas Dolgin, Alessandro Armando, and Umberto Morelli. Large-scale analysis & detection of authentication cross-site request forgeries. In 2017 IEEE European Symposium on Security and Privacy, EuroS&P 2017, Paris, France, April 26-28, 2017, pages 350–365, 2017.

3. Stefano Calzavara, AlviseRabitti, AlessioRagazzo, and Michele Bugliesi. Testing for integrity flaws in web sessions. In Computer Security - 24rd European Symposium on Research in Computer Security, ESORICS 2019, Luxembourg, Luxembourg, September 23-27, 2019, pages 606–624, 2019.

4. OWASP. OWASP Testing Guide. https://www.owasp.org/index.php/ OWASP Testing Guide v4 Table of Contents, 2016.

5. Jason Bau, ElieBursztein, Divij Gupta, and John C. Mitchell. State of the art: Automated black-box web application vulnerability testing. In 31st IEEE Symposium on Security and Privacy, S&P 2010, 16-19 May 2010, Berkeley/Oakland, California, USA, pages 332–345, 2010.

6. Adam Doup´e, Marco Cova, and Giovanni Vigna. Why johnny can't pentest: An analysis of black-box web vulnerability scanners. In Detection of Intrusions and Malware, and Vulnerability Assessment, 7th International Conference, DIMVA 2010, Bonn, Germany, July 8-9, 2010. Proceedings, pages 111–131, 2010.

7. Adam Barth, Collin Jackson, and John C. Mitchell. Robust defenses for cross-site request forgery. In Proceedings of the 2008 ACM Conference on Computer and Communications Security, CCS 2008, Alexandria, Virginia, USA, October 27-31, 2008, pages 75–88, 2008.

8. Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of Machine Learning. The MIT Press, 2012.

9. Michael W. Kattan, Dennis A. Adams, and Michael S. Parks. A comparison of machine learning with human judgment. Journal of Management Information Systems, 9(4):37–57, March 1993.

10. D. A. Ferrucci. Introduction to “This is Watson”. IBM Journal of Research and Development, 56(3):235–249, May 2012.

11. Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, IoannisAntonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, NalKalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, KorayKavukcuoglu, Deep neural networks and tree search are used to master the game of Go. Jan 2016, Nature 529(7587):484–489.

12. WilayatKhan, Michele Bugliesi, Stefano Calzavara, Riccardo FocardiCookiext is a browser patch that protects against session hijacking attacks. Journal of Computer Security, Vol. 23, No. 4, 2015, pp. 509–537.

13. Salvatore Orlando, Stefano Calzavara, Gabriele Tolomei, Andrea Casini, Michele Bugliesi, and Gabriele Tolomei On the web, a supervised learning strategy to safeguard client authentication. TWEB, 9(3), 2015, 15:1– 15:30.

14. Gabriele Tolomei, Stefano Calzavara, Mauro Conti, Riccardo Focardi, AlviseRabitti Mitch: A machine learning technique to detecting CSRF vulnerabilities in the blackbox. EuroS&P 2019, Stockholm, Sweden, June 17-19, 2019, pages 528–543, in IEEE European Symposium on Security and Privacy.

15. Martin Johns, Simon Koch, Michael Backes, and Christian Rossow. Giancarlo Pellegrino, Martin Johns, Simon Koch, Michael Backes, and Christian Rossow. Deemon: Using dynamic analysis and property graphs to detect CSRF. CCS 2017, Dallas, TX, USA, October 30 - November 03, 2017, pages 1757–1771, in Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security.

Movie Recommendation System Using Sentiment Analysis from Micro Blogging Data

Badrakanti Deekshitha, Ladesani Gayathri, Ediga Jayanth Goud
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

K Maheswari
Associate Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ mahi.kirubakaran@gmail.com

ABSTRACT

The potential of referral systems (RSs) in online marketplaces and media consumption has attracted a great deal of attention. Both content-based filtering (CBF) and collaborative filtering systems (CFS) are used in RSs, although they have their limitations, such as relying on past customer data and routines to provide suggestions. This article suggests a hybrid RS for films that takes the best ideas from CF and CBF and adds belief assessment of small blogging site tweets to lessen the impact of such a limitation. You may learn about current trends, popular opinion, and individual reactions to the film by analysing film tweets. Results from experiments conducted on the publicly available dataset are promising.

KEYWORDS: *CF, CBF, RS, Hybrid RS, Micro blogging data.*

INTRODUCTION

The internet has become an integral component of modern human existence. The issue of too much information being readily accessible is one that customers often face. Customers may manage this information deluge with the help of newly launched recommendation systems (RS) [1]. Websites for tourism, leisure, and online shopping are some examples of e-commerce applications and expertise monitoring systems that make heavy use of RS [2]. The primary focus of this study is RS as films play a significant role in our daily lives as a kind of entertainment. People rely on internet sites for movie suggestions [3]. Genres such as humour, suspense, animation, and action allow for easy categorization of motion movies. Metadata, such as release year, language, supervisor, or cast, provides an additional layer of potential film classification [4]. The majority of online video-streaming services use a user's viewing or rating history to recommend other, similar films. The use of motion picture suggestion tools [3, 4, 5, 6, 7] not only makes it easier to find good films, butit also helps us find more of the films we enjoy. The most important thing about a movie recommendation system is that it should be trustworthy and provide the user with recommendations for films that are comparable to their tastes. Decisions in many areas of daily life may now be greatly aided by RS, thanks to the exponential growth of internet data. Commonly, RS are categorised into two groups: content-based filtering systems and collective filtering systems. The CF principle originated from the human propensity to base judgements on facts, set regulations, and known information that is easily accessible online. To aid users in finding relevant short articles among a large number of postings, Resnick et al. [6] introduced the concept of CF in netnews. CF trains people to think about the feelings and experiences of others before making decisions. Whereas in CBF [7] things are suggested based on similarity among the goods' content information, two people are deemed comparable when their item rankings are similar. People may now broadcast their everyday emotional states online because to the proliferation of social media sites like Quora, Facebook, and Twitter. Since its inception in 2006, Twitter has grown to become one of the most popular social media platforms, allowing users to express themselves concisely and with little character [8].

Existing users gain access to a variety of user-generated content in addition to information based on their social network ties; this is Twitter's unique selling point. Tweets, which are short messages that keep users informed about their favourite subjects, people, and films, are the main source of information on Twitter. Using the Flick Tweetings data source, we present a flick recommendation structure that combines crossbreeding with sentiment evaluations. Here are the main points of the paper:

1. We propose a hybrid recommendation system that combines content-based filtering with joint filtering.
2. This recommendation system is enhanced with the usage of view analysis.
3. There is a thorough evaluation of the proposed referral mechanism based on substantial experimental evidence. Lastly, it is shown that there is a quantitative and qualitative disparity with other versions of the standards.

LITERATURE SURVEY

Evaluating personal models on Twitter for tailored news recommendations

K. Tao, F. Abel, Q. Gao, and G.-J. Houben are the authors

How exactly can Twitter's microblogging features be put to use for the purpose of user modelling and customisation? This study delves into this issue and introduces a framework for Twitter customer modelling that enhances the semantics of tweets and identifies items (such as people, events, and goods) mentioned in them. We consider the methodologies for building customer accounts based on hashtags, entities, or topics and how they benefit from semantic enrichment. We also examine the accounts' temporal features. Within the framework of a tailored information referral system, we go even further in identifying and contrasting the efficacy of the various modelling approaches. We found that semantic enrichment increases the variety and quality of the generated customer accounts. We also find that taking temporal profile trends into account may increase referral quality, and we see how various user modelling techniques impact customisation.

Consumer behaviour modelling for information referrals via Twitter

The authors of this work are K. Tao, G.-J. Houben, Q. Gao, and F. Abel.

How can we make the most of Twitter's microblogging features to better understand and cater to our users? In this study, we delve into this question and provide a model for personal modelling on Twitter that improves the meaning of tweets and finds the people, places, and things that are addressed in them. We assess the methods that use semantic enrichment to create consumer profiles based on hashtags, entities, or topics, and we look at how these accounts change over time. Within the framework of a personalised news recommendation system, we are better able to evaluate and compare the efficacy of the various modelling approaches. Our results show that semantic enrichment improves the quality and selection of the established user accounts. Additionally, we see that taking temporal profile trends into account might improve referral quality and that various individual modelling approaches impact customisation.

A review of the current and potential extensions towards the next generation of recommendation systems.

G. Adomavicius and A. Tuzhilin are the authors.

This article introduces the field of recommended systems and details the latest generation of recommendation techniques, which are often categorised into content-based, collaborative, and hybrid tactics. Additionally, this article outlines several limitations of current recommendation techniques and discusses potential extensions that might enhance suggestion capabilities and broaden the range of applications for recommended systems. Some of the features that have been added to this version include better user and object comprehension, the ability to incorporate contextual facts in referrals, support for ratings based on multiple criteria, and recommendations that are both more flexible and less intrusive.

Making ensemble approaches for deep knowledge sentiment analysis in social applications even better.

Sánchez-Rada, O. Araque, I. Corcuera-Platas, and C. A. Iglesias were the writers.

Belief evaluation approaches based on deep learning have become quite popular. In comparison to conventional function-based procedures (i.e., surface area approaches), they provide automated feature extraction in addition to improved efficiency and more robust representation capabilities. The extraction procedure is crucial to attribute driven approaches, which are based on traditional surface methods that rely on hand-drawn characteristics. By combining their predicting capacities with the new deep-knowing methodologies, these long-standing tactics may provide robust norms. Our goal in writing this study is to improve the efficiency of deep learning algorithms by combining them with conventional surface techniques that rely on manually deleted features. There are six main points made by this study. To begin, we use a direct machine-discovery approach and a word-embedding architecture to set up a view classifier that relies on deep knowledge. The outcomes that follow will be measured against this classifier. Additionally, we propose two ensemble methods that combine our basic classifier with a number of different surface area classifiers that are often used in sentiment evaluation. Third, to combine data from several sources, we provide two models that use deep and surface functions together. In the fourth place, we provide a taxonomy that may be used to recognise the many models found in literature and the ones we propose. Fifth, to evaluate how these variants stack up against the gold standard in deep learning, we run a battery of trials. We do this by making use of seven publicly available datasets that were extracted from the domains of microblogging and movie reviews. Finally, a statistical investigation confirms that these version recommendations outperform our original F1-Score benchmark.

METHODOLOGY

Collaboration filtering (CF) and content-based filtering (CBF) are two examples of traditional RS techniques. Both of these approaches have their limitations, such as the fact that they need knowledge of the user's past actions and preferences in order to provide recommendations.

PROBLEM DEFINITION

Users often encounter the challenge of extremely provided information. The information explosion is being tackled by deploying recommendation systems (RSs) to aid users. Netflix, Prime Video, and IMDB are digital entertainment platforms that primarily use RS. E-commerce platforms like Amazon, Flipkart, and eBay all make extensive use of RS. An essential tool for leisure and home pleasure, RS for films is the subject of this brief essay. Online portals are the backbone of movie recommendations for people. Genres including "comedy," "thriller," "computer animation," and "activity" make it easy to classify movies. Some other workable way to sort the movies according to their metadata, such release year, language, director, or cast. In order to provide a more tailored experience for its users, some online video-streaming providers compile user data, including what they've seen and rated content.

OBJECTIVE OF PROJECT

Understanding the current trends, public opinion, and user reaction to the film may be achieved by analysing movie tweets. The results of experiments conducted on the publicly available data set are encouraging.

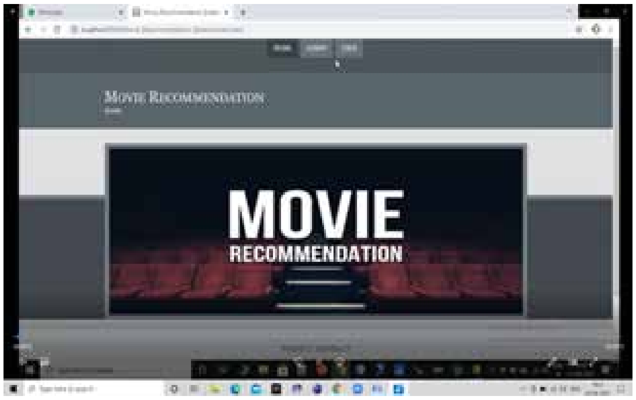
In Figure 1 we can see the proposed sentiment-based RS display. Here we detail the various parts of the suggested RS. A. Summary of the Data Set There are two types of data sources needed by the proposed system. Both the individual tweets from Twitter and a user-rated film database that include ratings for related films are available.

1) Sources of Public Data: numerous people use the numerous popular public databases that are available to them to get recommendations for films and other forms of entertainment material. We extracted movie-related tweets from Twitter and compared them to films already in our database so that we could include view analysis into the suggested structure. Due to the lack of micro blogging information, experiments performed utilising several public data sources, such as Movielens 100K,2 Movielens 20M,3 the Internet Film Database (IMDb,4) and the Netflix database,5, were not found to be perfect for our work. The MovieTweetings database [12] was

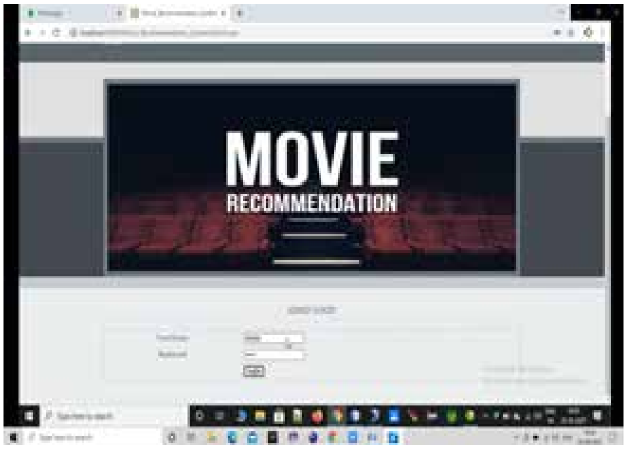
finally selected for the proposed system after a thorough review of the aforementioned data sources. Many see MovieTweetings as an updated take on the MovieLens database. Providing a current movie rating that includes more useful information for sentiment analysis is the goal of this data source. Table I provides the necessary information about the Flick Tweetings database.

2) Source for Customised Twitter Data for Motion Pictures: The suggested effort involves tailoring the Movie Tweeting data source to run the RS. Using sentiment analysis of user-generated tweets in the RS movie prediction was the primary motivation for updating the database. From 1894 to 2017, the films covered by the Motion Picture Tweetings database have been released. We extracted a subset of the database that met our purpose by only considering films that were launched in or after 2014 due to the lack of tweets for older films.

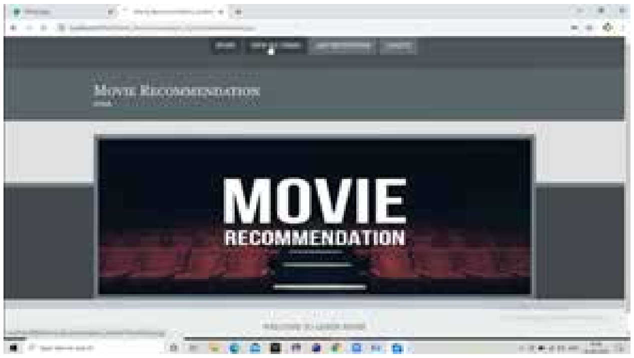
Home Page



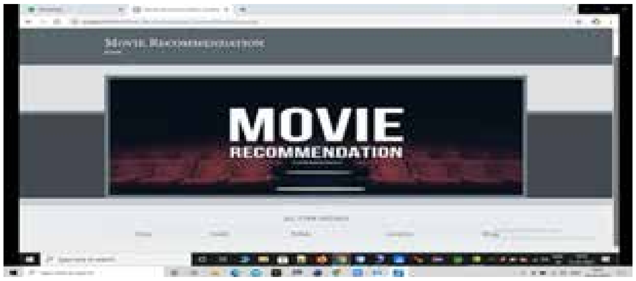
Admin Login



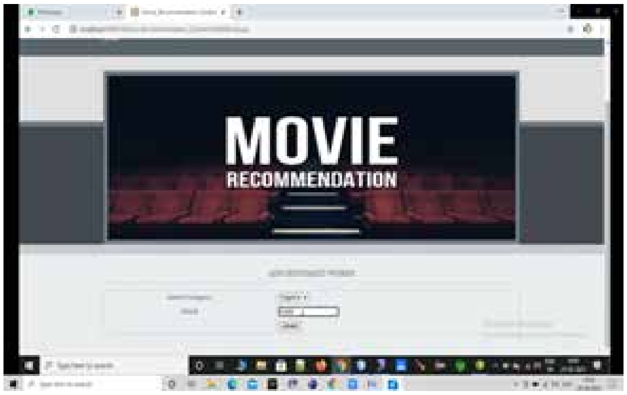
Admin Page



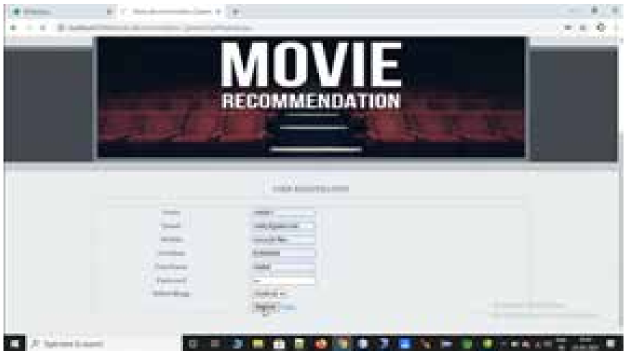
View All Users



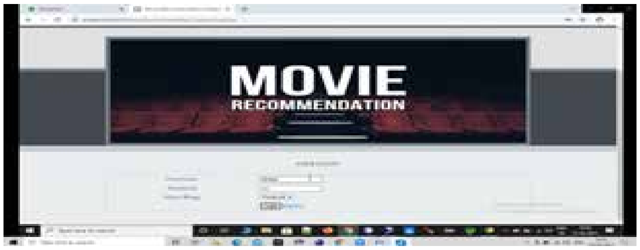
Add Sentiment Words



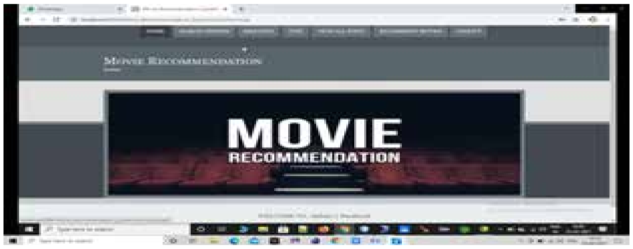
User Registration



User login



User HomePage



CONCLUSION

With the availability of vast amounts of data, RSs have become an integral part of information filtering systems in the current day. This article presents a movie recommendation system (RS) that makes use of Twitter sentiment assessment data, film information, and a social chart to suggest films. With the use of view analysis, we can learn how people are responding to a film and how valuable that feedback is. In order to enhance the recommendations, the suggested approach made extensive use of score fusion. Our test results show that the mean accuracy in the top five for believed resemblance is 0.54 and 1.04, for crossbreed it is 1.86 and 3.31, and for the proposed model it is 2.54 and 4.97. Compared to the previous models, we found that the proposed model provides more targeted recommendations. We want to take into account additional information about the individual's psychological tone from other social media sites and languages other than English in the future in order to improve the RS even more.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "MOVIE RECOMMENDATION SYSTEM USING SENTIMENT ANALYSIS FROM MICRO BLOGGING DATA", which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. F. Abel, Q. Gao, G.-J. Houben, and K. Tao, "Analyzing user modeling on Twitter for personalized news recommendations," in Proc. 19th Int. Conf. Modeling, Adaption, Pers. (UMAP). Berlin, Germany: Springer Verlag, 2011, pp. 1–12.
2. F. Abel, Q. Gao, G.-J. Houben, and K. Tao, "Twitter-based user modeling for news recommendations," in Proc. Int. Joint Conf. Artif. Intell., vol. 13, 2013, pp. 2962–2966.
3. G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," IEEE Trans. Knowl. Data Eng., vol. 17, no. 6, pp. 734–749, Jun. 2005.
4. O. Araque, I. Corcuera-Platas, J. F. Sánchez-Rada, and C. A. Iglesias, "Enhancing deep learning sentiment analysis with ensemble techniques in social applications," Expert Syst. Appl., vol. 77, pp. 236–246, Jul. 2017.
5. E. Aslanian, M. Radmanesh, and M. Jalili, "Hybrid recommender systems based on content feature relationship," IEEE Trans. Ind. Informat., early access, Nov. 21, 2016, doi: 10.1109/TII.2016.2631138.
6. J. Bobadilla, F. Ortega, A. Hernando, and J. Alcalá, "Improving collaborative filtering recommender system results and performance using genetic algorithms," Knowl.-Based Syst., vol. 24, no. 8, pp. 1310–1316, Dec. 2011.
7. R. Burke, "Hybrid recommender systems: Survey and experiments," User Model. User-Adapted Interact., vol. 12, no. 4, pp. 331–370, 2002.
8. E. Cambria, "Affective computing and sentiment analysis," IEEE Intell. Syst., vol. 31, no. 2, pp. 102–107, Mar./Apr. 2016.
9. I. Cantador, A. Bellogín, and D. Vallet, "Content-based recommendation in social tagging systems," in Proc. 4th ACM Conf. Rec. Syst. (RecSys), 2010, pp. 237–240.
10. P. Cremonesi, Y. Koren, and R. Turrin, "Performance of recommender algorithms on top-N recommendation tasks," in Proc. 4th ACM Conf. Rec. Syst. (RecSys), 2010, pp. 39–46.

Packet Inspection to Identify Network Layer Attacks using Machine Learning

Shaik Mohammed Altaf,
Ammannagari Vidhathri, Maddela Anuroop
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

Antharam Ganapathi
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ antharamganapathi@gmail.com

ABSTRACT

One dependable way of network security is intrusion detection, which may identify unknown attacks from network website traffic. Traditional maker detecting models like KNN, SVM, etc., are the basis of most current approaches to network anomaly detection. While these methods do achieve some impressive results, they rely heavily on human-made traffic function arrangements, which is now completely out of date in this era of big data, and thus achieve very poor accuracy. A traffic anomaly detection version BAT is suggested to fix the problems with lower accuracy and function design in invasion detection. Bidirectional Long Temporary Memory (BLSTM) and a focus device are both incorporated into the BAT model. The BLSTM model generates packet vectors that make up the network circulation vector; the focus device assesses these vectors to determine the network web traffic category's essential functions. Also, in order to get the specific details of the traffic data, we use a number of convolutional layers. Since BAT architecture makes use of several convolutional layers to handle data samples, we call it BAT-MC. Network website traffic category is when the soft max classifier comes into play. Automatic discovery of the power structure's key features is possible with the proposed end-to-end model, which does not require any function design skills. Anomaly detection capabilities may be improved, and network web traffic behaviour can be described adequately. Experimental results demonstrate that our version of the model outperforms other comparison approaches when tested on a publicly available criterion dataset.

KEYWORDS: *BAT, MAC, CNN, DL, ML.*

INTRODUCTION

When it comes to protecting sensitive data on a network, breach detection is a must. When it comes to intrusion detection, machine learning techniques for identifying malicious online traffic have been widely deployed. Aspect engineering and selection are typical tenets of these methods, which are superficial learning approaches. Low acknowledgment accuracy and a high dud rate are results of their inability to adequately address the massive breach information categorization problem and issues with characteristics [1]. A slew of intrusion detection methods based on deep discoveries have been proposed in recent years. Below, we will go over some of the most common types of attacks that hit networks. More and more threats to the security of the Internet and computer networks are cropping up. Constantly emerging new types of assaults make it difficult to design security-oriented techniques that can adapt to changing circumstances [2]. To protect specific systems and networks from harmful activities, anomaly-based network breach detection algorithms are a helpful piece of current technology. In order to help train the dataset and forecast the test dataset, we examine the website traffic and create a dataset for types of strikes [3]. Pattern recognition and the development of detection systems based on the dataset are important goals of IDS aided machine learning research [4]. Since there are many different types of network attacks, we know we need to develop methods for identifying web traffic packets in order to assess these attacks [5].

Packet analysis is a method for evaluating the contents of data packets passing through a network interface, such a firewall or router. In order to find threats to the network's security that operate at the network layer, this method is used. An essential tool for ensuring the integrity of data sent across a network and protecting the network itself is packet inspection [6]. Although packet evaluation is useful for identifying attacks at the network layer, keeping up with the ever-changing threat environment may be difficult. With the use of machine learning (ML), systems may instantly respond to new risks by learning from past data and improving the efficiency of package assessment [5], [7]. Traditional techniques of evaluating software depend on established standards and signatures to identify common types of attacks; nevertheless, these approaches may fail to detect novel or unanticipated threats [8]. Machine learning algorithms may learn to see trends in network website traffic and identify irregularities that might indicate an attack, all without using pre-established criteria. Systems can adapt to changing threats and learn from new attack patterns by analysing network data via device learning. Improved packet inspection accuracy, fewer false positives, and quicker detection and action time frames are all possible outcomes of this [9]. Additionally, machine learning may aid in the detection of previously undiscovered or "zero-day" attacks that have not been officially categorised or recognised. Since these attacks may go undetected by conventional signature-based methods, they pose a particularly serious threat [10].

PROBLEM DEFINITION

Most of the current approaches to detecting anomalies in networks use old-school machine learning models like KNN, SVM, etc. Despite the impressive characteristics that may be obtained, these approaches suffer from poor accuracy and mainly depend on human-designed traffic features, which is no longer relevant in the era of big data [5].

OBJECTIVE OF PROJECT

To capture the local aspects of traffic data, we use many convolutional layers. We call the BAT model BAT-MC since it uses several convolutional layers to handle data samples. For the purpose of categorising network traffic, the soft max classifier is used.

LITERATURE SURVEY

Research on Web of Things Security Breach Detection Techniques

Nishtha Kesswani and Sarika Choudhary wrote it.

The Net of Points is the newest lingo in internet technology. The Internet of Things (IoT) is an expanding network that will enable everyday items to become intelligent digital counterparts. The Internet of Things (IoT) is a diverse network that allows devices with different functions to communicate with one another and share data. These days, the actions of the enlightened are crucial to human survival. Because of this, establishing safe connections inside the IoT network is challenging. An Invasion Detection System is necessary to secure the IoT network from cyber-attacks, since verification and file encryption are not enough. The Internet of Things (IoT), its architecture, contemporary technology, attacks, and intrusion detection systems (IDS) are the areas of emphasis in this research piece. An introduction to the Internet of Things (IoT), various methods for discovering security breaches, and threats to the IoT are the primary aims of this article.

Detection of network intrusions

K.N. Levitt, L.T. Heberlein, and B. Mukherjee wrote the paper.

An innovative, retrofit approach, intrusion detection allows preexisting computer systems and data networks to continue operating in their current "open" state while providing a feeling of security. The goal of intrusion detection is to identify instances of unauthorised access, manipulation, or exploitation of computer systems by both internal and external users. The increased interconnection of computer system systems provides more access to outsiders and makes it much easier for trespassers to avoid detection, making the invasion finding problem a more difficult one to tackle as diverse local area networks proliferate. Based on the idea that a trespasser's behaviour would differ considerably from a legitimate customer's, breach detection systems (IDSs) are able to identify several prohibited behaviours. In order to detect breaches, intrusion detection systems often employ rule-based misuse models and analytical anomaly detection. There are a number of model intrusion detection systems (IDSs) that have been developed at various institutions, with some of them implemented in experimental systems. This research assesses and identifies the characteristics of several intrusion detection systems (IDSs), both host-based and network-based. When it comes to detecting malicious tasks, host-based systems rely on the operating system's audit trails as their main source of information. On the other hand, network-based intrusion detection systems (IDSs) rely on monitored web traffic as its discovery device, with some also using host audit routes. Also included is a synopsis of an analytical anomaly finding formula that is used in a standard intrusion detection system.

Research on a system that uses machine learning to identify network breaches based on SDN

THE WRITERS: W. Peng, R. Alhadad, N. Sultana, and N. Chilamkurti

Thanks to the advent of programmable features, software-defined networking (SDN) offers the potential to rapidly discover and monitor network security concerns. Recently, SDN-based NIDS have begun to use Machine Learning (ML) approaches to better protect computer networks and address network security issues. The SDN environment is seeing the emergence of a new wave of advanced device discovery methods, known as deep knowing technology (DL). Several existing deal with artificial intelligence (ML) approaches that use SDN to establish NIDS were assessed in this research. More specifically, we examined the methods used by deep learning to build SDN-based NIDS. Meantime, we discussed gadgets in this research that may be used to develop NIDS versions in an SDN environment. Lastly, this research wraps up with a discussion of future employment opportunities and ongoing difficulties in implementing NIDS using ML/DL.

EXISTING SYSTEM

There have been other algorithms that were proposed for use before. The authors of [6] summarise the following pattern matching formulae used in the Invasion Detection System: KMP, BM, BMH, BMHS, a/c, and AC-BM. The results of the experiments show that the enhanced algorithm is able to increase the matching rate and performs well. There is a comparison of the three most successful pattern/intrusion finding algorithms in [7]: the Ignorant technique, the Knuth-Morris-Pratt formula, and the Rabin-Karp formula. In order to determine the formula's efficiency, we have used Pcap documents as datasets and taken their execution times into account.

PROBLEMS WITH THE CURRENT SYSTEM:

1. Diverse safety risks are also something we have to cope with. Network safety and security has shifted the focus of cultural and government divisions due to the growth in network infections, eavesdropping, and damaging attacks.
2. Identifying various harmful network website traffic, especially unexpected harmful network web traffic, is an essential and insurmountable difficulty.

PROPOSED SETUP

The BAT-MC network can achieve an accuracy of 84.25%, which is 4.12% higher than the current CNN version and 2.96% higher than the RNN version. Our work has many important contributions and discoveries, including the following:

Using a combination of BLSTM and an attention system, we propose the BAT-MC end-to-end deep knowledge design. Using BAT-MC, we can find a new way to study invasion detection and fix the problem of breach finding.

The second step is to provide the attention mechanism to the BLSTM model so that it can focus on the important input. Consecutive data consisting of information package vectors are used to conduct feature knowing using the focus method. The collected feature data is accurate and fair.

We evaluate BAT-MC in comparison to more conventional deep learning methods; the BAT-MC architecture is able to extract data from all of the packages. The BAT-MC version improves function recording by fully utilising structural information of network web traffic.

We test our suggested link using a real-life NSL-KDD dataset. Based on the results of the speculation, BAT-MC outperforms the conventional methods.

The system's proposed benefits are:

1. From bottom to top, the BAT-MC model consists of five layers: input, several convolutional layers, BSLTM, interest, and result.
2. The BAT-MC paradigm transforms every byte of online traffic into a one-hot data format at the input layer. An n-dimensional vector is used to inscribe each traffic byte. We carry out normalisation procedures once the traffic byte is swapped for a mathematical type.

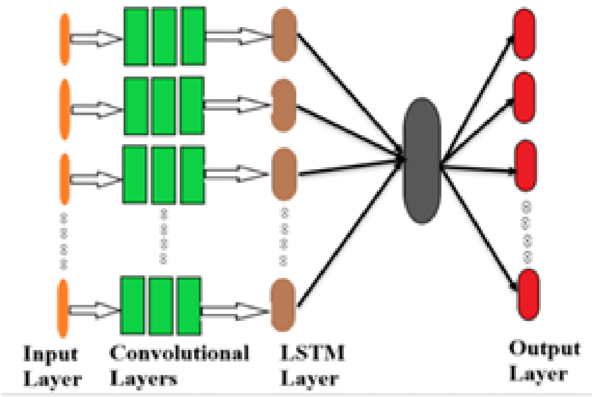
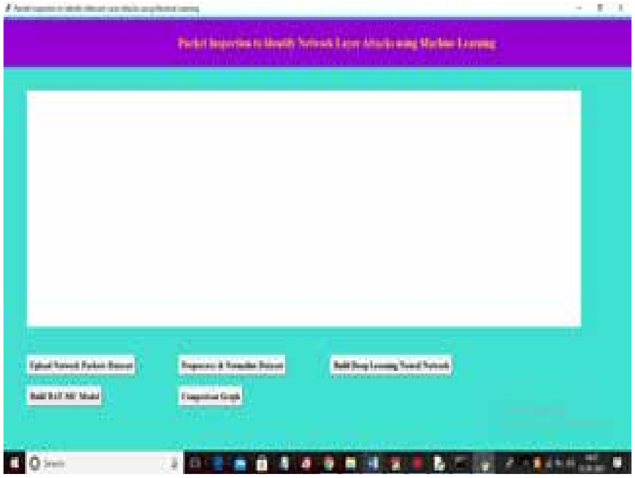


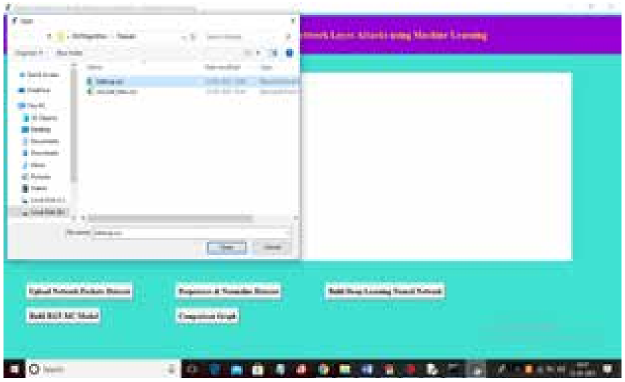
Fig.1. System Architecture

METHODOLOGY

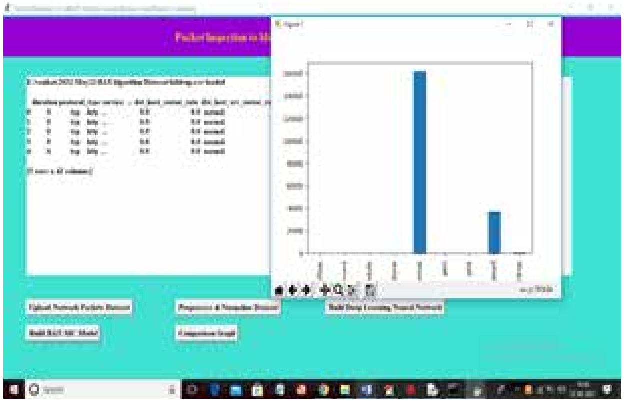
To run project, double click on 'run.bat' file to get below screen.


Packet Inspection to Identify Network Layer Attacks using Machine Learning
Upload Network Packets Dataset
Preprocess & Normalize Dataset
Build Deep Learning Neural Network
Build BAT-MC Model
Comparison Graph

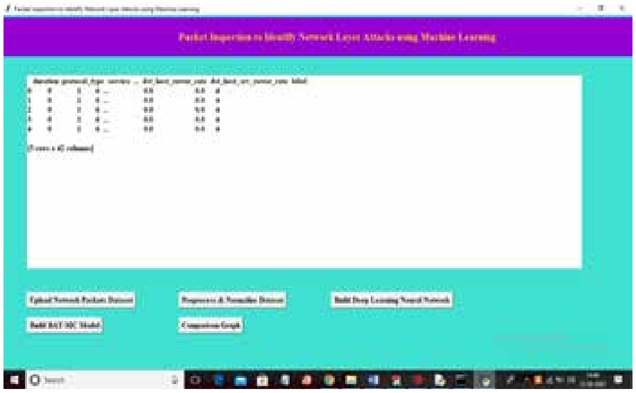
To upload the dataset, go to the first page and click on the "Upload Network Packets Dataset" button. Now you should see the second screen.


Open
Network Layer Attacks using Machine Learning
Upload Network Packets Dataset
Preprocess & Normalize Dataset
Build Deep Learning Neural Network
Build BAT-MC Model
Comparison Graph
Open
Cancel

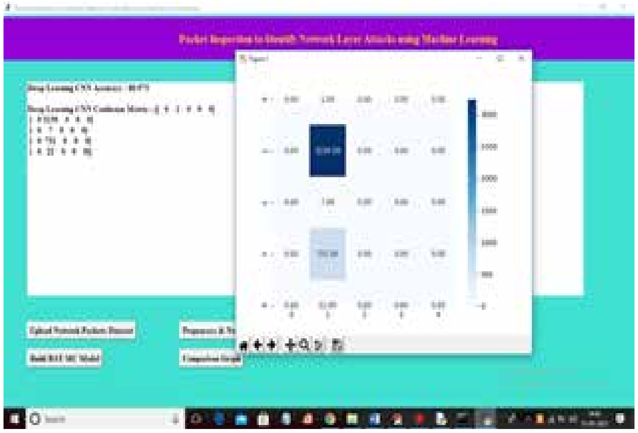
To load the dataset, choose the "kddcup.csv" file and click the "Open" button on the previous page. This will bring up the following screen.


Packet Inspection to Id
Figure 1
Upload Network Packets Dataset
Preprocess & Normalize Dataset
Build Deep Learning Neural Network
Build BAT-MC Model
Comparison Graph

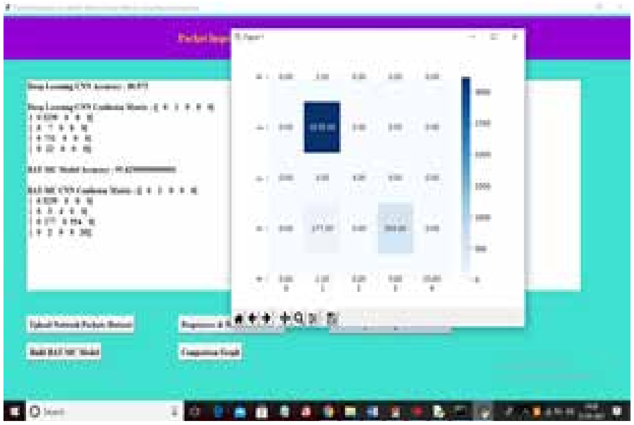
The dataset is loaded in the message area of the above screen. The data is numerical and contains alpha values; since ML formulas only work with numerical values, we need to pre process and stabilise them. The chart shows the various attack types on the y-axis and the names of the strikes on the x-axis. To normalise the information, close the graph and then click the "Preprocess & Normalise Dataset" button.


Packet Inspection to Identify Network Layer Attacks using Machine Learning
Upload Network Packets Dataset
Preprocess & Normalize Dataset
Build Deep Learning Neural Network
Build BAT-MC Model
Comparison Graph

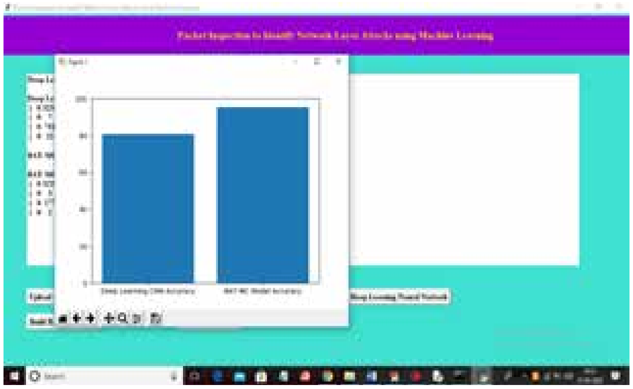
Click the "Build Dee Knowing Semantic network" button to train a convolutional neural network (CNN) using the prepared dataset; after that, check the accuracy of the predictions. The dataset was transformed to numerical values by assigning IDs to each unique piece of non-numerical data.


Packet Inspection to Identify Network Layer Attacks using Machine Learning
Upload Network Packets Dataset
Preprocess & N
Build BAT-MC Model
Comparison Graph

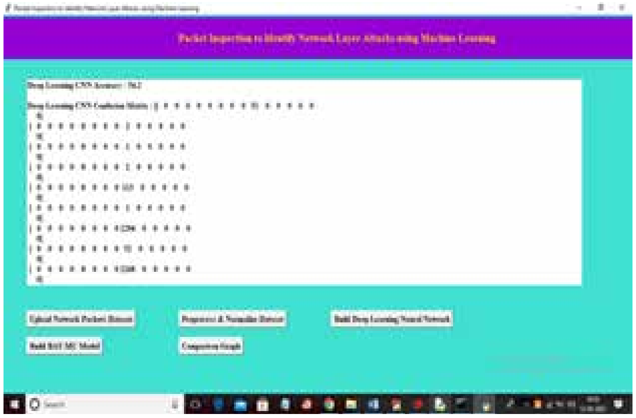
The CNN algorithm achieved an accuracy rate of 80%, as seen in the above screen. The confusion matrix shows that there are five distinct attacks, and it also shows how many times each attack was predicted. Strike 2 was expected to occur 32,390 times in the whole test data. Click "Construct BAT-MC Version" after you close the above chart to train the BLSTM algorithm on the dataset and then check the accuracy of your predictions using the test data.



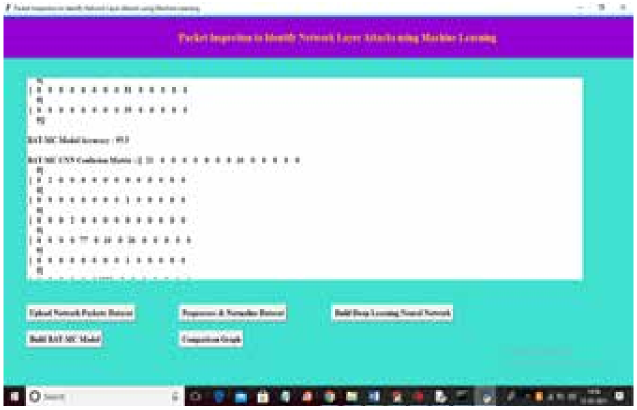
After you've closed the graph up top and clicked the "Comparison Graph" button, you'll see the BAT-MC model below, which has a prediction accuracy of 95.


Packet Inspection to Identify Network Layer Attacks using Machine Learning
Deep Learning CNN Accuracy
BAT-MC Model Accuracy
Build Deep Learning Neural Network

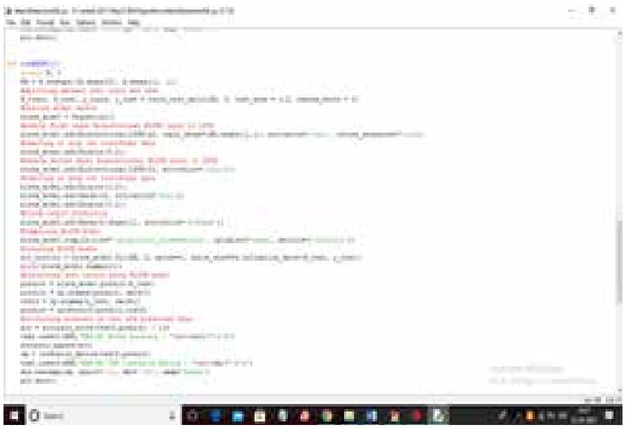
On the one hand, we have the BAT-MC model, which provides superior accuracy, and on the other, we have the algorithm names shown on the x-axis of the aforementioned graph. The process of building CNN and BAT-MC models is similar; you can just input different datasets. The following screen displays the correctness of the NSL dataset:


Packet Inspection to Identify Network Layer Attacks using Machine Learning
Upload Network Packets Dataset
Preprocess & Normalize Dataset
Build Deep Learning Neural Network
Build BAT-MC Model
Comparison Graph

After scrolling down to the message position, you may acquire the BAT-MC accuracy; in the above display, CNN achieved 54% precision.


Packet Inspection to Identify Network Layer Attacks using Machine Learning
Upload Network Packets Dataset
Preprocess & Normalize Dataset
Build Deep Learning Neural Network
Build BAT-MC Model
Comparison Graph

The code screen below shows how we are developing a BLSTM model to construct an attack detection model, while the top screen shows that BAT-MC achieved 95% accuracy.



If you want to know how the BAT-MC model has evolved, look at the red remarks on the screen above.

CONCLUSION

When it comes to classifying network web traffic, the existing deep learning methods aren't making full advantage of the structured data. We propose a novel model, BAT-MC, based on the two-stage learning of BLSTM, and we zero in on time series characteristics for intrusion detection using the NSL-KDD dataset, drawing on deep learning application techniques in the domain of natural language processing. Attributes on each package's traffic bytes are removed using the BLSTM layer, which links the forward and backward LSTM. A package vector may be generated by any data package. A network circulation vector is formed by arranging these package vectors. Attribute learning is performed on the network circulation vector, which consists of package vectors, using the focus layer. Deep semantic networks, which lack a traditional function design, may now do the aforementioned feature learning operation in a flash. This update deftly sidesteps the issue of features that need user interaction. The BAT-MC method is evaluated using the KDDTest+ and KDDTest-21 datasets. The BAT-MC design achieves rather good accuracy, according to the speculative results on the NSL-KDD dataset. The results of the BAT-MC versions are quite promising when compared to other current deep learning-based approaches, as seen by these comparisons with a common classifier. Our group has come to the conclusion that the suggested method is a powerful instrument for dealing with the intrusion detection problem.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "PACKET INSPECTION TO IDENTIFY NETWORK LAYER ATTACKS USING MACHINE LEARNING", which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. B. B. Zarpelo, R. S Miani, C. T. Kawakani, and S. C. de Alvarenga, "A survey of intrusion detection in Internet of Things," J. Netw. Comput. Appl., vol. 84, pp. 25–37, Apr. 2017.

2. B. Mukherjee, L. T. Heberlein, and K. N. Levitt, "Network intrusion detection," IEEE Netw., vol. 8, no. 3, pp. 26–41, May 1994.

3. S. Kishorwagh, V. K. Pachghare, and S. R. Kolhe, "Survey on intrusion detection system using machine learning techniques," Int. J. Control Automat., vol. 78, no. 16, pp. 30–37, Sep. 2013.

4. N. Sultana, N. Chilamkurti, W. Peng, and R. Alhadad, "Survey on SDN based network intrusion detection system using machine learning approaches," Peer-to-Peer Netw. Appl., vol. 12, no. 2, pp. 493–501, Mar. 2019.

5. M. Panda, A. Abraham, S. Das, and M. R. Patra, "Network intrusion detection system: A machine learning approach," Intell. Decis. Technol., vol. 5, no. 4, pp. 347–356, 2011.

6. W. Li, P. Yi, Y. Wu, L. Pan, and J. Li, "A new intrusion detection system based on KNN classification algorithm in wireless sensor network," J. Electr. Comput. Eng., vol. 2014, pp. 1–8, Jun. 2014.

7. S. Garg and S. Batra, "A novel ensembled technique for anomaly detection," Int. J. Commun. Syst., vol. 30, no. 11, p. e3248, Jul. 2017.

8. F. Kuang, W. Xu, and S. Zhang, "A novel hybrid KPCA and SVM with GA model for intrusion detection," Appl. Soft Comput.., vol. 18, pp. 178–184, May 2014.

9. W. Wang, M. Zhu, X. Zeng, X. Ye, and Y. Sheng, "Malware traffic classification using convolutional neural network for representation learning," in Proc. Int. Conf. Inf. Netw. (ICOIN), 2017, pp. 712–717.

10. P. Torres, C. Catania, S. Garcia, and C. G. Garino, "An analysis of Recurrent Neural Networks for Botnet detection behavior," in Proc. IEEE Biennial Congr. Argentina (ARGENCON), Jun. 2016, pp. 1–6.

Predicting the Rice Leaf Diseases using CNN

Sara, Anumula Abhitej Reddy
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

Voruganti Naresh Kumar
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ nareshkumar99890@gmail.com

ABSTRACT

A variety of diseases strike rice, one of the many widely cultivated plants in India, at various points during its growth process. Manual disease identification is very challenging for farmers with little knowledge. Convolutional Neural Network (CNN)–based automated picture identification systems have lately shown promising results in deep learning research. Due to the difficulty in locating a picture dataset of rice leaf disease, our deep finding model was constructed using Transfer Learning on a tiny dataset. The planned convolutional neural network (CNN) architecture is trained and tested using VGG-16 using net and rice area datasets. A price of 95% accuracy is associated with the proposed design. Some of the keywords included in this index are deep learning, fine-tuning, CNN, and illnesses that affect rice leaves.

KEYWORDS: *CNN, VGG. SVM, ANN.*

INTRODUCTION

For most people, including those in India, rice is the main food staple. A vast assortment of maladies attack it during its advancement training. Medical early detection and treatment of various illnesses is crucial for a high-quality harvest, but this is challenging since individual farmers tend vast tracts of land, which harbour a wide range of diseases, and because a single plant might be infected with several diseases [1]. Trying to find agricultural experts in faraway places is a time-consuming and difficult task. Because of this, automated systems are necessary [2]. Researchers have used support vector machines (SVMs) and synthetic neural networks (ANNs) to aid struggling farmers and improve the accuracy of plant disease detection. However, picking the right characteristics has a major impact on how accurate these algorithms are [3]. Convolutional neural networks have made picture recognition possible, doing away with the need for image pre-processing while providing built-in function choice [4]. Another issue is the scarcity of large datasets pertaining to these types of challenges. When working with small datasets, it's recommended to use designs that have been trained on larger datasets. Removing or fine-tuning the last layer of connections to be more specific to the dataset at hand is an option when building a new version using Transfer Learning. Using their mobile devices, farmers may upload photos of infected leaves to our server [5]. Our semantic network will then determine the disease and provide the farmers with a diagnosis and treatment recommendations. The pervasive scheduling of mobile phones made us anxious about this notion. This research presents an automated system component for sickness categorization [6]. We were able to create a deep understanding method for this task with the aid of convolutional neural network research. In order to accommodate our own datasets into the completely linked levels of the VGG-16 architecture, we have used Transfer Discovering to enhance the fully connected layers. We reviewed our errors and tried to figure out what went wrong in the end [7] [8].

LITERATURE SURVEY

Applying Deep Learning to Discover Plant Diseases via Images

THE AUTHORS: V. Singh and A. Misra

The abstract Although agricultural diseases pose a significant danger to global food security, they are difficult to detect quickly due to a lack of equipment. Disease medical diagnosis by smartphone is becoming a reality because to the enormous usage of smart devices worldwide and current advancements in computer vision allowed by deep understanding. Using 54,306 images of healthy and diseased plant leaves (or their absence) to train a convolutional semantic network, we are able to identify 14 plant kinds and 26 diseases from a publicly accessible dataset. By using a held-out test set, the expert model proves the strategy's validity with an accuracy of 99.35%. Using a different collection of images from the ones used for training, the design achieves an accuracy of only 31.4%. An additional diverse set of training data may still increase the overall accuracy, however. There is a clear way forward for global plant disease detection using publicly accessible image datasets trained on deep learning models.

Detection of grape leaf diseases using SVM classifiers

The authors of this work are P. B. Padol and A. A. Yadav. India is home to some of the world's most beloved grape varieties. Infections that affect the grapevine's fruit, stems, and leaves reduce harvest yields. Viruses, fungi, and bacteria are the most prevalent sources of these toxins. A number of factors, including health problems, limit the amount of fruit that may be produced. Without a correct medical diagnosis of the illness, it is not feasible to implement effective control measures. Photo handling is a commonly used technique for identifying and classifying plant leaf diseases. To aid in the identification and classification of grape leaf diseases, SVM classification is used in this assignment. To pinpoint the sick spot, we use K-means clustering for segmentation, and then we extract texture and colour data. The penultimate step in diagnosing leaf diseases is to use a classification system. The proposed approach has an accuracy of 88.89% when applied to the issue under study.

Extremely Deep Convolutional Networks for Massive Image Analytics

Authors: Andrew Zisserman and Karen Simonyan

The abstract Using a large-scale picture identification challenge, we investigate the effect of convolutional network depth on performance. Using a design with tiny (3x3) convolution filters, we examined deep learning networks and found that increasing the depth to 16-19 weight layers greatly outperformed state-of-the-art configurations. This study's primary contribution is this. This investigation led to our team's first-and second-place finishes in the 2014 ImageNet Challenge's localization and category competitions. The superior outcomes produced by our representations are applicable to a broad range of datasets. We have made available two of our top ConvNet architectures to facilitate further investigation into deep visual representations in computer vision.

Deep learning with hyper-class augmentation and regularisation for granular image categorization

Y. Lin, X. Wang, S. Xie, and T. Yang are the authors

In the field of large things acknowledgment (CNNs), deep convolutional semantic networks have shown impressive results. Generic object identification is much easier than fine-grained image categorization (FGIC) because of the low cost of coarse-grained categorised data (usually requiring domain name competency) and the large intra-class and modest inter-class variation. One of the most common approaches is to pre-train the CNN on an external dataset (such as ImageNet) and then fine-tune it using the small target data to achieve a specific classification objective. This paper introduces two new aspects to the problem of learning a deep CNN: (i) a new way of knowing that is based on the concept of a "multitask knowing" paradigm; and (ii) the identification of easily annotated hyper-classes in the fine-grained data and the collection of images with hyper-classes classified from external, easily accessible sources (like image search engines). The suggested method has been evaluated using a large, automotive-specific dataset in addition to two small, fine-grained datasets (Stanford Dogs and Stanford Cars).

EXCITING SYSTEM

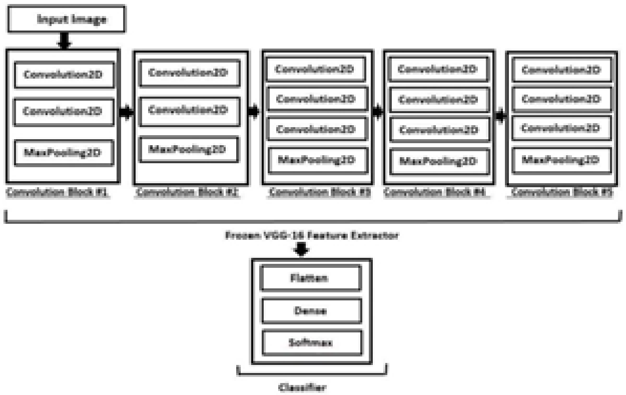
Manual examination of the fallen leaf was the only way to diagnose a sickness in bygone days. Visual examination of plant leaves in conjunction with consulting a reference book revealed the illness. Limitations in accuracy, inability to study every leaf, and time commitment are the three main issues with this technique. More precise ways of detecting these diseases are appearing as a result of ongoing scientific study and technological advancement. Two approaches to consider are deep discovery and photo processing.

In image processing, techniques such as clustering, filtering, and pie chart assessment are used to identify the affected area. Conversely, issues are uncovered by deep finding semantic networks.

PROPOSED SYSTEM

This article trains a collection of rice illnesses using a VGG16 transfer finding semantic network. The trained version may be used to forecast disease from fresh photos. Since the writer was unable to train the VGG16 model using the KAGGLE Rice Fallen Leave dataset, he resorted to the transfer discovering CNN technique, which involves transferring an existing CNN model to a new dataset and then training it using the new information.

Evidence suggests that VGG16 transfer learning improves prediction accuracy in both the standard CNN model and the standard CNN design enhanced with VGG16 transfer learning.



METHODOLOGY

MODULES

Setup for Experiment

The trial ran on a Windows 10 PC with 64-bit architecture. Python 3.7.2, the Tensor Flow 1.14.0 backend, and the Keras 2.2.4 deep learning framework were used in the development of the CNN design.

Picture Processing

Not only were all of the images shot outdoors, but they were also shot entirely digitally. The photographs that make up the dataset summary include leaf explosion, brown place, healthy plants, and leaf curse. Revamping and Enhancing Images Using Image Information Generator in Keras, new images are created at 224 × 224 pixel resolution after applying several enhancement algorithms to the collected images, such as zooming, rotating, and straight and vertical shifting.

The Modelling School at CNN

Training and screening cannot take place without first packing the image information set. For training purposes, class labels and photographs are stored in separate choices. Training uses 70% of the data and testing uses 30% as a result of the train-test split technique. After the first 70% of data is broken up, 30% is used for validation. Each label is represented as a vector rather than an integer, and the class labels are encoded as integers using a one-hot inscribing process. In the end, Keras removes the last fully linked layers. System improvements that are not trainable are implemented. We finished by applying a softmax filter to the squished result from the function extractor before turning to the filter. We used the Adam optimizer to build our system from the ground up, and we used categorical cross decline as our category loss feature. We have really stopped here as the results did not change after 25 dates. Our real category treatment processes are shown in number 3.

Provide validation in support of the selected design.

The process of "transfer learning" allows us to apply what we learn in one context to another. When training neural network models, transfer learning is quite useful as most real-world scenarios do not include many recognised data components. A massive amount of data is required to train a neural network from scratch, yet this data isn't always easily accessible. Because the version has already been pre-trained, a modest quantity of training data may be used to construct a solid device learning design. Instead, we used a VGG Web that had already been trained on our little dataset.

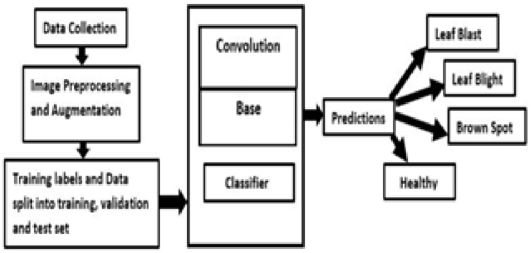

Data Collection
Image Preprocessing and Augmentation
Training labels and Data split into training, validation and test set
Convolution
Base
Classifier
Predictions
Leaf Blast
Leaf Blight
Brown Spot
Healthy

Fig. Overview of the steps of the proposed model

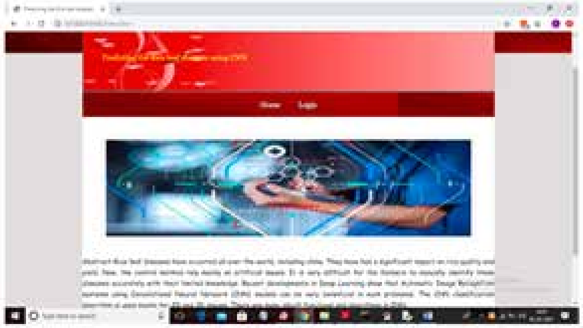
OPERATION

To run project install python 3.7 and tensorflow package 1.14.0 and then install Django==2.1.7

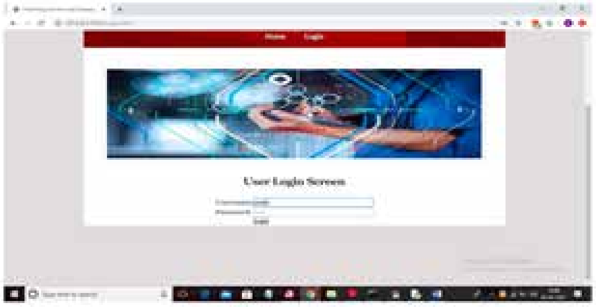
After installation run below command from 'RiceDisease' folder

Python manage.py runserver

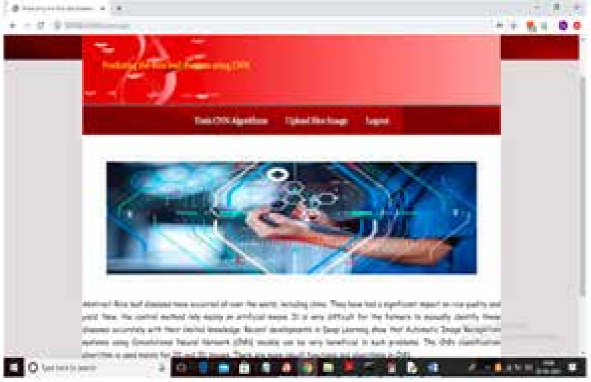
Then open browser and enter URL as http://127.0.0.1:8000/index.html and press enter key to get below screen



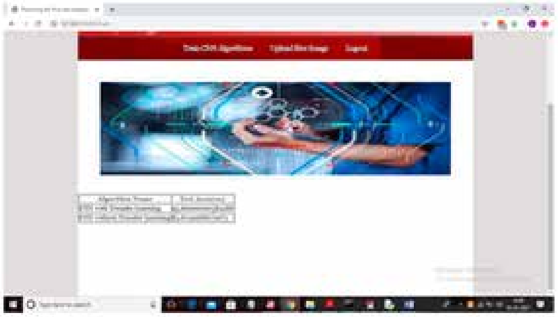
In above screen click on 'Login' link to get below login screen



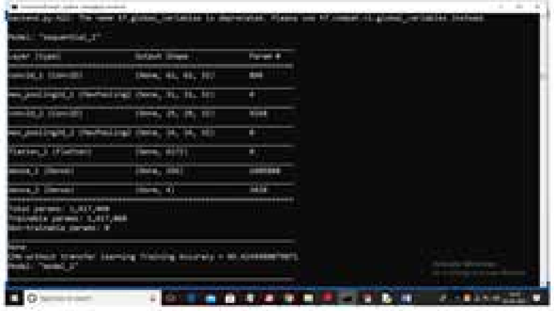
In above screen enter username as 'admin' and password as 'admin' and then click on 'Login' button to get below screen



In above screen click on 'Train CNN Algorithms' link to train both VGG16 and normal CNN without transfer learning on rice dataset and then calculate prediction accuracy

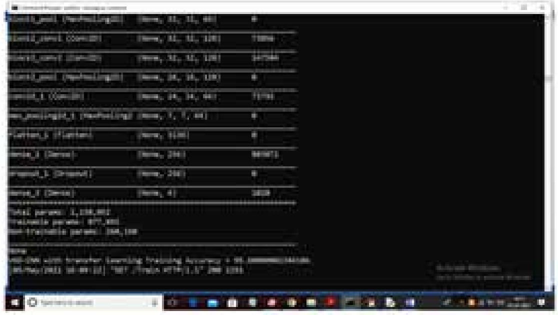


In above screen CNN with transfer learning VGG16 got 95% accuracy and without transfer learning got 85% accuracy so VGG16 is giving better result. In below console you can see layer details of VGG 16 and Normal CNN



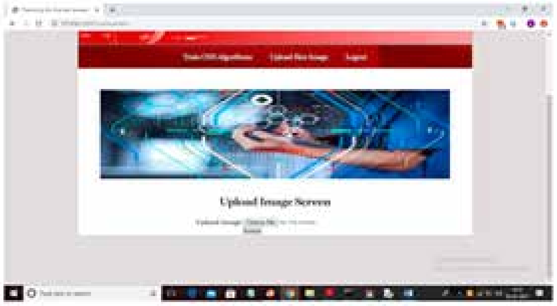
```
backend.py:422: The name tf.global_variables is deprecated. Please use tf.compat.v1.global_variables instead.

Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_1 (Conv2D)            (None, 62, 62, 32)        896
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 31, 31, 32)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 29, 29, 32)        9248
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 14, 14, 32)        0
_________________________________________________________________
flatten_1 (Flatten)          (None, 6272)              0
_________________________________________________________________
dense_1 (Dense)              (None, 256)               1605888
_________________________________________________________________
dense_2 (Dense)              (None, 4)                 1028
=================================================================
Total params: 1,617,060
Trainable params: 1,617,060
Non-trainable params: 0
_________________________________________________________________
None
CNN without transfer learning Training Accuracy = 85.[illegible]
Model: "model_1"
```

In above screen normal CNN created 4 layers and got 85% accuracy and in below screen you can see VGG16 layers

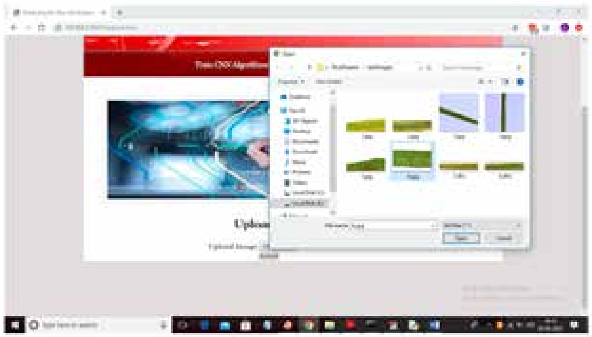


```
block1_pool (MaxPooling2D)   (None, 32, 32, 64)        0
_________________________________________________________________
block2_conv1 (Conv2D)        (None, 32, 32, 128)       73856
_________________________________________________________________
block2_conv2 (Conv2D)        (None, 32, 32, 128)       147584
_________________________________________________________________
block2_pool (MaxPooling2D)   (None, 16, 16, 128)       0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 14, 14, 64)        73792
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 7, 7, 64)          0
_________________________________________________________________
flatten_2 (Flatten)          (None, 3136)              0
_________________________________________________________________
dense_3 (Dense)              (None, 256)               803072
_________________________________________________________________
dropout_1 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_4 (Dense)              (None, 4)                 1028
=================================================================
Total params: 1,138,052
Trainable params: 877,892
Non-trainable params: 260,160
_________________________________________________________________
None
VGG-CNN with transfer learning Training Accuracy = 95.[illegible]
[[illegible]] "GET /Train HTTP/1.1" 200 [illegible]
```

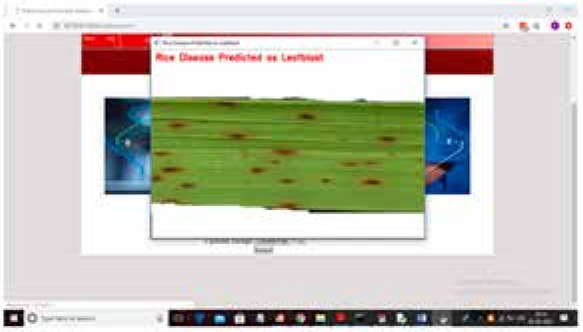
In above screen VGG16 contains so many layers and its accuracy is 95% and now in below screen click on 'Upload Rice Image' link



Now in above screen click on 'Choose File' button to upload leaf test image from 'testImages' folder

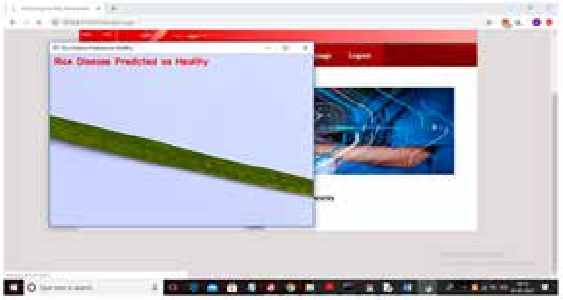


In above screen selecting and uploading '6.jpg' file and then click on 'Open' button then click on 'Submit' button to get below result



Rice Disease Predicted as Leafblast

In above screen in uploaded image disease predicted as 'Leaf Blast' and now test other image



Rice Disease Predicted as Healthy

In above screen leaf predicted as healthy and similarly you can upload other images and test them

CONCLUSION

Our deep learning architecture successfully classifies 95% of the test images after training on 40 images of rice leaves and subsequent screening on 20 additional images. By honing the VGG16 model, we significantly improved the design's performance on this little dataset. After getting evidence showing no progress in accuracy or decreased loss on either the training or recognition sets, we maxed the variety of epochs employed at 20.

In the future, we will need a lot more images taken at farms and by agricultural research groups if we want our findings to be as accurate as possible. To further verify our searches, we want to include a cross-validation technique in the future. It would be helpful if we could compare the results of the completed searches to those of other ongoing initiatives, such as sophisticated deep understanding models. If you use this version, you may find other leaf illnesses in plants, and those plants are big in India.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "PREDICTING THE RICE LEAF DISEASES USING CNN", which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. T. Gupta, "Plant leaf disease analysis using image processing technique with modified SVM-CS classifier," Int. J. Eng. Manag. Technol, no. 5, pp. 11-17, 2017.
2. Y. Es-saady,T. El Massi,M. El Yassa,D. Mammass, and A. Benazoun, "Automatic recognition of plant leaves diseases based on serial combination of two SVM classifiers," International Conference on Electrical and Information Technologies (ICEIT) pp. 561-566, 2016.
3. P. B. Padol and A. A. Yadav, "SVM classifier based grape leaf disease detection," Conference on Advances in Signal Processing (CASP), pp. 175-179, 2016.
4. L. Liu and G. Zhou, "Extraction of the rice leaf disease image based on BP neural network," International Conference on Computational Intelligence and Software Engineering ,pp. 1-3,2009.
5. S. Arivazhagan and S.V. Ligi, "Mango Leaf Diseases Identification Using Convolutional Neural Network," International Journal of Pure and Applied Mathematics, vol. 120,no. 6, pp. 11067-11079,2008.
6. B. Liu,Y. Zhang, D. He and Y. Li, "Identification of Apple Leaf Diseases Based on Deep Convolutional Neural Networks Symmetry," X. X. &Suen, C. Y. A novel hybrid CNN–SVM classifier for recognizing handwritten digits. Pattern Recognition,vol. 45,pp. 1318–1325,2012.

7. . Lu, S. Yi,N. Zeng,Y. Liu, and Y. Zhang, “Identification of Rice Diseases Using Deep Convolutional Neural Networks,” Neurocomputing, 267, pp. 378-384,2017.

8. R. R. Atole, D. Park, “A Multiclass Deep Convolutional Neural Network Classifier for Detection of Common Rice Plant Anomalies,” International Journal Of Advanced Computer Science And Applications,vol. 9,no. 1,pp. 67–70,2018.

9. V. Singh, A. Misra, “Detection of Plant Leaf Diseases Using Image Segmentation and Soft Computing Techniques,” Information Processing in Agriculture,vol. 4 ,no. 1,pp. 41–49,2017.

10. P. Konstantinos Ferentinos, “Deep Learning Models for Plant Disease Detection and Diagnosis,” Computers and Electronics in Agriculture ,vol. 145,pp. 311–318,2018.

Respiratory Analysis Detection of Various Lung Infections using Cough Signal

Paindla Saketh, Gajawada Rahul
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

M Madhusudhan
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ madhu9963@gmail.com

ABSTRACT

Many people, of all ages, succumb to pulmonary chronic lung illnesses every year. One of the most useful diagnostic tools for identifying lung diseases is a lung sound assessment. The previous method of detecting lung diseases, which relied on hands-on discovery, was unreliable for a number of reasons, including limited audibility and differences in how different medical practitioners perceived different sounds. Patients suffering from a variety of lung disorders may now benefit from significantly more effective treatment options made possible by modern computerised analysis that returns data with substantially more precision. Among these illnesses are Pneumonia, Asthma, Bronchitis, Emphysema, and Tuberculosis. Hissing, shortness of breath, rhonchi, and a persistent cough are some of the symptoms. Asthma, pneumonia, bronchitis, and many more disorders may be predicted using the respiratory audio dataset that we are working with in this task. We trained a convolutional semantic network (CNN) formula model using essence functions from both the condition medical diagnostic dataset and the respiratory system sound dataset, and then we were able to complete the job. Following the training design, we are free to provide additional test data of any kind in order to use it for sickness prediction.

KEYWORDS: *Breath, CNN, Asthma, Heart sounds.*

INTRODUCTION

The inability to breathe normally is known as a pulmonary issue. In the past, doctors would only have a rough understanding of the disease from their hands-on exams, leading to severe therapy [1]. Previously, this was an effective kind of exercise. A very precise assessment of the severity of sickness is necessary due to the fact that the extremely high incidence of contamination and people's unhealthy habits has led to the development of far more complex disorders [2]. The only way to achieve this level of precision is to automate the assessment process. The ability to differentiate between the sounds produced by infected lungs and those of normal, healthy lungs was recognised by researchers as a promising tool for the development of more precise diagnostic tools [3]. The standard procedure for doing the assessment has been to record the lung sounds, isolate them from the heart sounds and other background noise, and then analyse the waveform of the filtered lung sound. The lung seems to be filtered and handled in a variety of ways [4]. There are a plethora of filtering systems and LS examination techniques shown by a cursory review of the aforementioned papers. Due to creepy and temporal overlap in both audios, the separation of HS from LS is the most challenging task in the analysis. Methods for filtering out short-term scary components' temporal trajectories are used, such as the Inflection Domain filtering system [5]. The Fourier transform and subsequent partition into successively overlapping frames allows for the investigation of signals. In the adaptive-frequency domain name filtering system mix [6], a very simple approach is detailed, which entails removing heart sounds from a mix of heart and lung sounds [7].

EXISTING SYSTEM

Lungs are an integral part of the respiratory system and play a role in exchanging gases like carbon dioxide and oxygen. When we breathe. The pulmonary system is responsible for transferring oxygen from the air to the blood and carbon dioxide from the blood back into the air [5], [6]. For many diseases affecting the respiratory system, coughing up blood is the first sign of trouble [7]. A wet cough produces some mucus and acts as a defense mechanism to prevent the respiratory system tract from inadvertently absorbing foreign substances or those produced internally by an infection, whereas a dry cough produces no noticeable mucus [8] [9]. Changes in the pattern of the cough sound could indicate lung disease. Cases of pathology may arise as a result of issues such as obstruction, limitation, and embedded patterns [10].

PROPOSED SYSTEM

Asthma, pneumonia, bronchitis, and many other respiratory illnesses may be predicted with the use of the respiratory system sound dataset, which is utilised in this task. In order to carry out this task, we have used a convolutional semantic network (CNN) algorithm version that was trained on condition diagnostic datasets and respiratory system audio datasets after removing features from each dataset. Once the model has been trained, we may use it to predict diseases using any kind of fresh examination data.

MODULES DESCRIPTION

1) Submit the Respiratory System Sound Dataset: This component will be used to submit the respiratory system sound dataset and the dataset for sickness diagnosis.

2) Core Dataset characteristics: Before creating the training dataset, this module will be used to extract features from the two datasets.

3. Construct a Convolutional Neural Network (CNN) Model: We will use the over-train dataset to train a CNN model, which will allow us to create a trained model. This design may be applied to the task of disease prediction using any fresh study sound documents.

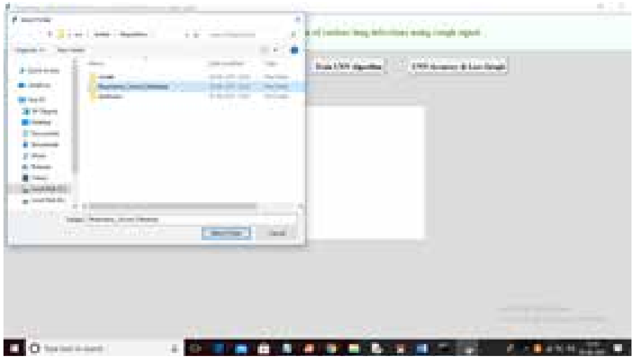
4) The CNN Precision & Loss Graph: This component will be used to provide a comparison chart between the CNN certified version's accuracy and loss.

5) We will use this component to upload test sound samples and then apply a CNN skilled model to those sounds in order to anticipate illness.

Operation

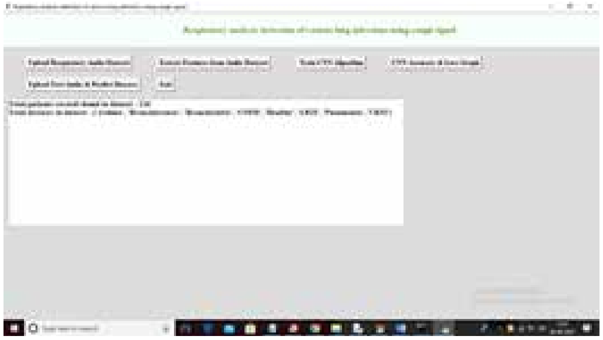
Double click the "run.bat" file to launch the project and see the screen below.

Press the "Upload Respiratory Sound Dataset" button on the previous screen to send in your dataset.

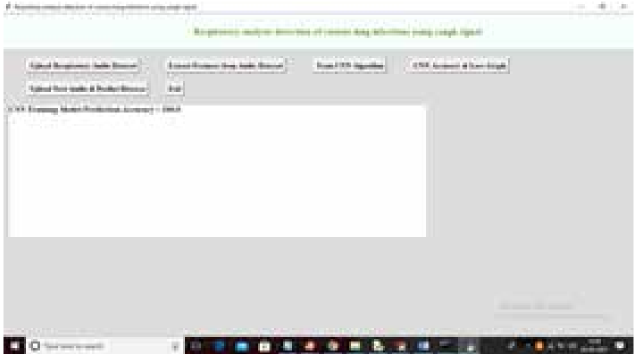


After choosing and uploading the whole respiratory sound folder in the previous page, you may load the dataset on the one below by clicking the "Select Folder" button.

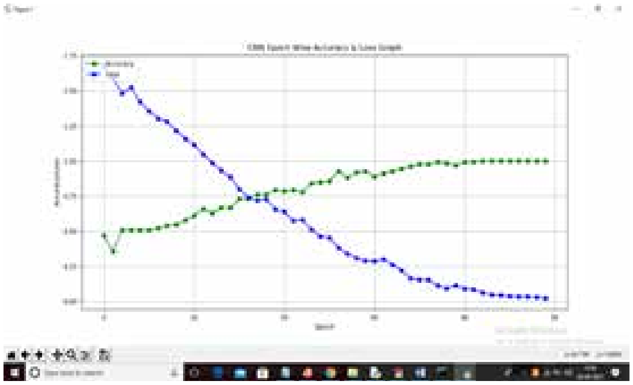
To start extracting features from the audio recordings, click the “Extract Features from Audio Dataset” button after checking the patient’s related disease diagnosis in the previous screen. Next, give each audio recording a class label according to the identified ailment.



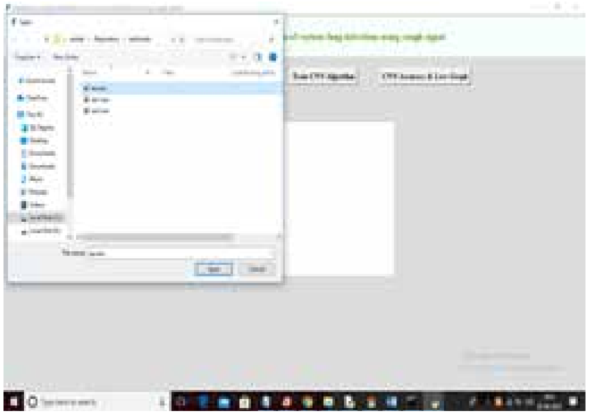
After locating 126 audio samples belonging to patients with 8 distinct illnesses on the aforementioned page, you can prepare your dataset for CNN training by clicking the “Train CNN Algorithm” button.



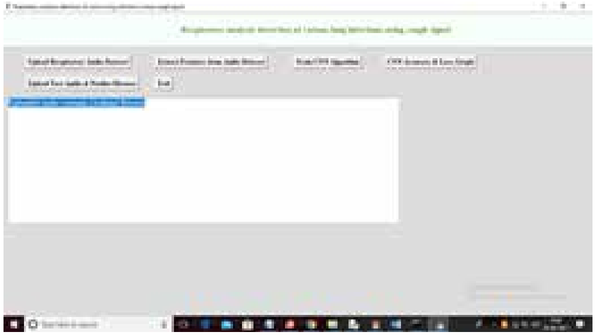
The graph below, which was obtained by clicking the “CNN Accuracy & Loss Graph” button, demonstrates that CNN trained on sound features and attained 100% accuracy.



We used 50 EPOCH to train the CNN design, and as we can see from the graph, accuracy increases with each iteration, loss values decrease to 0, and accuracy reaches 100%. On the x-axis, we have EPOCH/ITERATIONS. We can see the accuracy and loss figures on the y-axis. The blue line denotes loss, and the green line stands for accuracy. Click “Upload Test Audio & Predict Disease” immediately to post test sound documents.



Use the ‘Open’ button to get the following prediction after selecting and uploading ‘aa.wav’ files on the previous screen.



Condition predicted as “BRONCHIAL ASTHMA” type Submitted audio data and evaluation with other files may be seen in the blue text display above.

CONCLUSION

By exchanging gases, the lungs are essential to the respiratory system (oxygen and carbon dioxide). while we inhale. The process of transporting carbon dioxide from the blood back into the atmosphere and oxygen from the air into the blood occurs in the lungs. We have trained a convolutional neural network (CNN) formula design using features extracted from both the condition diagnostic dataset and the respiratory sounds dataset in order to accomplish this task. Any additional test data may be uploaded once the training version is complete in order to use it for condition prediction.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "RESPIRATORY ANALYSIS DETECTION OF VARIOUS LUNG INFECTIONS USING COUGH SIGNAL", which gave us excellent resources and assistance to complete our assignment. I want to express my gratitude to our chairman, director, deans, head of the department, computer science and engineering department, guide, and teaching and non-teaching faculty members for their insightful advice and direction in all facets of our work.

REFERENCES

1. Pulmonary breath Sounds. East Tennessee State University, November 2002.
2. J. J, Ward. R.A.L.E. Lung Sounds Demo. Med. RRT in Respiratory Care, Canada, 2005.
3. Think labs Digital Stethoscope Lung Library.
4. 3M Littmann Stethoscope Lung Sound Library.
5. Tiago H. Falk, Wai-Yip Chan, Ervin Sejdic´ and Tom Chau, "Spectro-Temporal Analysis of Auscultatory Sounds", New Developments in Biomedical Engineering, Intech, 2010.
6. Gadge PB and Rode SV, "Automatic Wheeze Detection System as Symptoms of Asthma Using Spectral Power Analysis", Journal of Bioengineering & Biomedical Science, 2016.
7. Bor-Shing Lin, Huey-Dong Wu and Sao-Jie Chen, "Automatic Wheezing Detection Based on Signal Processing of Spectrogram and Back - Propagation Neural Network", Journal of Healthcare Engineering, Vol. 6, No. 4, pp. 649- 672, 2015.
8. L Pekka Malmberg, Leena Pesu, Anssi R A Sovijarvi, "Significant differences in flow standardised breath sound spectra in patients with chronic obstructive pulmonary disease, stable asthma, and healthy lungs", Thorax, Vol. 50, pp. 1285-1291, 1995.ss
9. Arati Gurung, Carolyn G Scrafford, James M Tielsch, Orin S Levine and William Checkley, "Computerized Lung Sound Analysis as diagnostic aid for the detection of abnormal lung sounds: a systematic review and meta-analysis", Respiratory Medicine, Vol. 105, No. 9, pp. 1396–1403, 2011.
10. Pankaj B. Gadge, Bipin D. Mokal, Uttam R. Bagal, "Respiratory Sound Analysis using MATLAB", International Journal of Scientific & Engineering Research, Vol. 3, Iss. 5, 2012.
11. Rutuja Mhetre and U.R.Bagal, "Respiratory sound analysis for diagnostic information", IOSR Journal of Electrical and Electronics Engineering, Vol. 9, Iss. 5, 2014.
12. Ria Lestari Moedomo, M. Sukrisno Mardiyanto, Munawar Ahmad, Bachti Alisjahbana, Tjahjono Djatmiko, "The Breath Sound Analysis for Diseases Diagnosis and Stress Measurement", Proc. of International Conference on System Engineering and Technology, Bandung, Indonesia, 2012.

Using Deep Neural Networks to Detect Electricity Theft in Smart Grids

Thati Yashwanth, Talagama M Deekshit, Sugamanchi Narender
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

Jonnadula Narasimharao
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ jonnadula.narasimharao@gmail.com

ABSTRACT

Damage to power grids caused by electrical energy theft is a big contributor to nontechnical losses (NTLs) in distribution networks, which in turn affects the quality of power supplies and lowers operational profitability. The purpose of this study is to provide a novel CNN-RF version for automated detection of electricity theft, which will assist utility corporations in dealing with the problems of wasteful electrical energy examination and irregular power use. First, utilizing large amounts of dynamic smart meter data, this version trains a convolutional neural network (CNN) to distinguish between characteristics at different times of the day and over the course of days via convolution and down sampling. Moreover, the back breeding strategy is used during training to improve the network's criteria, and a dropout layer is employed to reduce the risk of overfitting. After that, the characteristics are used to train the arbitrary woodland (RF) to determine whether the customer is stealing electricity. The hybrid version uses the grid search technique to build the RF by determining the optimal criterion. Finally, experiments are carried out with real data on energy intake, and the outcomes show that the suggested detection strategy performs better in terms of accuracy and effectiveness than a wide range of alternative approaches.

KEYWORDS: *NTL, Electricity model, CNN, Hybrid model, RF.*

INTRODUCTION

One major issue that power providers worldwide confront is power loss during electrical power transmission and circulation. Standard practice dictates that energy losses be either technological (TLs) or nontechnical (NTLs) [1]. Transmission losses (TL) are an inevitable aspect of electrical power transmission and are caused by internal processes in power system components like transformers and transmission lines [2]. Non-transmission losses (NTL) are defined as the difference between total losses and TLs and are mostly caused by energy theft. Physical assaults such as line touching, metre breakage, or metre analysis interfering are really the most common methods of power theft [3]. Power providers can see a drop in revenue as a result of these schemes. An example would be the anticipated yearly losses in the US of around $4.5 billion due to electrical power burglaries [4]. Power theft costs utility companies throughout the globe an estimated $20 billion annually [5]. The security of the power system may also be affected by actions related to electrical power theft. For example, electrical power theft may lead to massive electric system fires, which pose a threat to public safety and security. Consequently, the safety and reliability of the power system depend on the prompt and accurate detection of electrical power theft.

By using AMI in smart grids, electricity companies were able to collect massive volumes of data on electrical energy use from smart metres on a regular basis, making it possible to detect power theft [6]. But there are two sides to every coin; new electrical power burglary attacks are possible thanks to the AMI network. Attacks on the AMI may originate from a variety of sources, including cyber-attacks and digital devices. Examining problematic instruments or equipment, comparing destructive metre records with benign ones, and manually looking for illegal line diversions are the main ways to find electrical energy burglaries. On the other hand, these methods need costly and time-consuming verification of every metre in a system. Furthermore, cyber attacks cannot be evaded using these manual tactics. There have been a number of proposals for solutions to the problems outlined above in recent years. The most common ways to categorise these approaches are as follows: state-based, game-theory-based, and AI-based. Important devices like distribution transformers and wireless sensors constitute the basis of state-based detection [7]. While these technologies potentially detect electrical energy theft, they rely on the often-impossible real-time acquisition of system geography and other physical dimensions. A game between an electrical utility and a burglar may be created using game-based discovery approaches. From this, a variety of distributions of normal and aberrant behaviours can be derived. Following the steps outlined in [8], they may reduce power burglary in an economical and practical way. But it's still not easy to figure out the utility function of all the players, including suppliers, regulators, and thieves. Methods that rely on AI include both surface-level AI and more advanced deep learning techniques. As shown in [9], current AI solutions may be further classified into clustering models and category models. Despite the outstanding and state-of-the-art nature of the aforementioned device finding discovery procedures, their performances are still insufficient for practical use. One issue is that these approaches struggle to handle high-dimensional data without requiring manual feature extraction. The standard deviation, minimum, maximum, and mean of the consumption data are unquestionably traditional properties that are hand-designed. In addition to being a tedious and time-consuming process, manually removing functions from smart metre data cannot detect 2D characteristics.

The authors of [10] examine deep finding approaches for power burglary detection and analyze a number of deep discovering architectures, including as convolutional neural networks (CNNs), long-short-term memory (LSTM) recurrent semantic networks (RNNs), and stacked auto encoders. But the detectors' effectiveness is tested using fake data, so we can't be sure how well they function when compared to more surface-level designs. Also, to counteract these cyber-attacks, the authors of [8] proposed a customer-specific detector based on deep semantic networks (DNNs). Convolutional neural networks (CNNs) have a variety of applications, and in recent years, they have been utilized to produce discriminatory and useful features from unprocessed input [9]. These applications motivate the extraction of CNN features for electricity theft detection from high-resolution smart meter data. An extensive and profound version of a convolutional semantic network (CNN) was developed and applied to investigate electrical energy theft in smart grids.

Similar to a general single hidden layer feed forward neural network (SLFN), the softmax classifier layer in a basic convolutional neural network (CNN) is trained by the back propagation approach [4]. While using the back propagation process, an over-trained SLFN is likely to have worse generalisation efficiency. However, local minima of training mistakes are taken into consideration by the back propagation formula, which is grounded on empirical threat reduction. Despite its impressive performance in attribute removal, CNN isn't always the best choice for classification (as we saw before, this is because of the softmax classifier's downside). Consequently, it is vital to discover a superior classifier that can fully use the acquired characteristics and not just has the same capability as the softmax classifier. To overcome the softmax classifier's limitations, many classifiers use the arbitrary forest (RF) classifier, This makes use of arbitrary attribute selection and bagging, two potent AI strategies. The creation of a unique convolutional neural network-random forest (CNN-RF) architecture for the purpose of detecting power theft was motivated by these specific goals. One important factor in the effectiveness of the electrical energy burglary detection model is the recommendation to use CNN for the automated collection of various functions of customers' consumption patterns from smart metre data. By substituting the RF with the softmax classifier, which discovers consumer patterns based on eliminated features, detection efficiency is improved. The data used to train and assess this model comes straight from the electrical energy utility customers in London and Ireland [9. 10].

LITERATURE SURVEY

Since a theft from the electrical power system would undoubtedly cause significant financial loss and disruption in power supply, the author of this study is using a combination of CNN and Random Forest to detect such a crime. Integrating CNNs with Random Forest methods has improved prediction accuracy compared to conventional methods, allowing us to reliably identify power grid theft. In power consumption, a number of 1 indicates energy theft if there is a spike in consumption within a certain time period; a value of 0 indicates normal power use.

Providing utilities with a prioritised list of customers based on the probability that their electrical energy metre contains an anomaly is the primary goal of the method outlined in this article.

Figure 1 shows the three main steps of the electrical burglary detection system, which are as follows:(i) Analyzing the factors that impact energy users' behaviours is the first step in data assessment and preprocessing, which is necessary for explaining the factor of employing a CNN for attribute extraction. Data transformation (normalisation), missing value imputation, and information cleansing (correcting outliers) are some of the pre-processing activities that come to mind.(ii)Creating a train dataset and an examination dataset: the cross-validation technique divides the pre-processed dataset into a train dataset and a test dataset so that the approach described in this study may be tested. We use the train dataset to hone our design criteria and the test dataset to see how effectively it adapts to novel, undiscovered customer scenarios. Given that power theft users far outnumber non fraudulent ones, the dataset's imbalance may significantly hinder the effectiveness of monitoring machine finding methods. The SMOT formula is used to equalise the number of electrical burglaries and non fraudulent users in the train dataset, hence reducing this bias.(iii)CATEGORY USING THE CNN-RF MODEL: The first CNN in the suggested CNN-RF version is trained to learn the characteristics over multiple days and hours using big and fluctuating wise metre data using convolution and down sampling processes. Finally, in order to determine whether the consumer is using electrical energy, The learned functions are used to train the RF category. Last but not least, the complexity matrix and receiver-operating characteristic (ROC) contours are used to evaluate the CNN-RF model's performance on the test dataset.

METHODOLOGY

People who employ terms like "am"—which may also be translated as "short articles" or "auxiliary verbs"—are reasonable and belong in this group. We can use the MRC thesaurus, which includes all terms in these categories, to make credible predictions about particular groups based on their tweets.

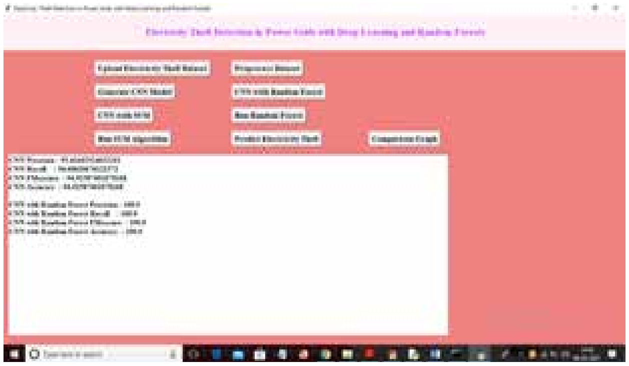
Individuals who use adjectives like "ugly," "unpleasant," "sad," etc., and those who exhibit neuroticism are likely to fall into this category. We can predict the score of this category by looking for these terms in tweets.

A person who has a large number of friends, fans, or followers on Twitter is likely to fall under the category of extroversion because of how well they get along with others.

Members of this category are conscientious and open about their hardworking beliefs in their posts.

So, we can predict a person's personality by assessing their Twitter profile and blog posts for the aforementioned five functions: openness (O), conscientiousness (C), extroversion (E), reasonableness (A), and neuroticism (N).

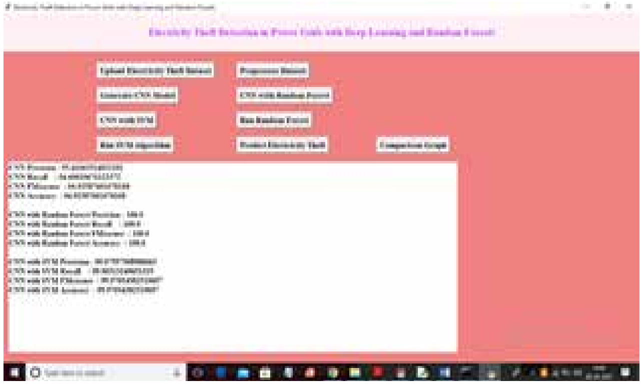
Using the Pearson correlation technique, we will determine the overall quality of all five functions by averaging the data from tweets. A person is considered to be part of a certain category if their rating for that function is more than 0.1. A person's uniqueness originates from many categories if their 0.1 worth of more than 1 trait is significant. For example, one can expect the same person to be sincere and diligent, among other qualities. We will utilize SVM, Random Forest, Naïve Bayes, and Logistic Regression for all features to assess the accuracy of the dataset and methods.


Electricity Theft Detection in Power Grids with Deep Learning and Random Forests
Upload Electricity Theft Dataset
Preprocess Dataset
Generate CNN Model
CNN with Random Forest
CNN with SVM
Run Random Forest
Run SVM Algorithm
Predict Electricity Theft
Comparison Graph

CNN Precision : [illegible]
CNN Recall : [illegible]
CNN FMeasure : [illegible]
CNN Accuracy : [illegible]

CNN with Random Forest Precision : 100.0
CNN with Random Forest Recall : 100.0
CNN with Random Forest FMeasure : 100.0
CNN with Random Forest Accuracy : 100.0

To train the dataset using CNN and SVM, use the “CNN with SVM” button; we achieved 100% accuracy with CNN-RF.

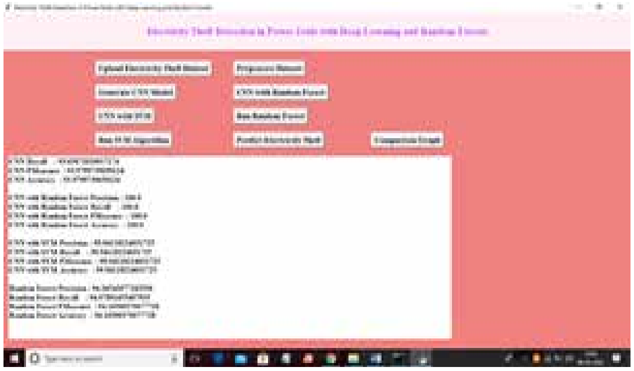

Electricity Theft Detection in Power Grids with Deep Learning and Random Forest
Upload Electricity Theft Dataset
Preprocess Dataset
Generate CNN Model
CNN with Random Forest
CNN with SVM
Run Random Forest
Run SVM Algorithm
Predict Electricity Theft
Comparison Graph

CNN Precision : [illegible]
CNN Recall : [illegible]
CNN FMeasure : [illegible]
CNN Accuracy : [illegible]

CNN with Random Forest Precision : 100.0
CNN with Random Forest Recall : 100.0
CNN with Random Forest FMeasure : 100.0
CNN with Random Forest Accuracy : 100.0

CNN with SVM Precision : [illegible]
CNN with SVM Recall : [illegible]
CNN with SVM FMeasure : [illegible]
CNN with SVM Accuracy : [illegible]

After achieving 99% accuracy using CNN-SVM, we may proceed to train RF on the dataset independently by clicking the “Run Random Forest” button.


Electricity Theft Detection in Power Grids with Deep Learning and Random Forest
Upload Electricity Theft Dataset
Preprocess Dataset
Generate CNN Model
CNN with Random Forest
CNN with SVM
Run Random Forest
Run SVM Algorithm
Predict Electricity Theft
Comparison Graph

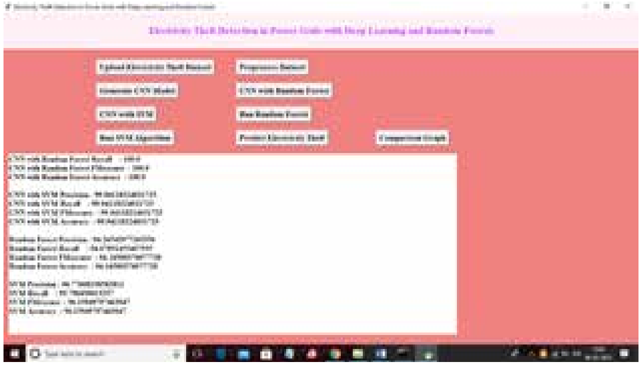
CNN Recall : [illegible]
CNN FMeasure : [illegible]
CNN Accuracy : [illegible]

CNN with Random Forest Precision : 100.0
CNN with Random Forest Recall : 100.0
CNN with Random Forest FMeasure : 100.0
CNN with Random Forest Accuracy : 100.0

CNN with SVM Precision : [illegible]
CNN with SVM Recall : [illegible]
CNN with SVM FMeasure : [illegible]
CNN with SVM Accuracy : [illegible]

Random Forest Precision : [illegible]
Random Forest Recall : [illegible]
Random Forest FMeasure : [illegible]
Random Forest Accuracy : [illegible]

To train SVM on the previously described dataset, click the “Run SVM Algorithm” button. Using Random Woodland alone, we were able to get 94% accuracy.


Electricity Theft Detection in Power Grids with Deep Learning and Random Forest
Upload Electricity Theft Dataset
Preprocess Dataset
Generate CNN Model
CNN with Random Forest
CNN with SVM
Run Random Forest
Run SVM Algorithm
Predict Electricity Theft
Comparison Graph

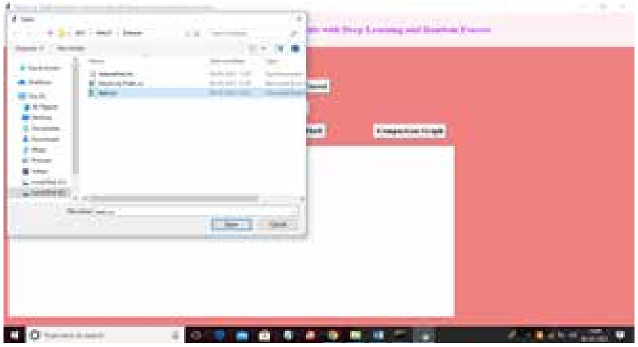
CNN with Random Forest Recall : 100.0
CNN with Random Forest FMeasure : 100.0
CNN with Random Forest Accuracy : 100.0

CNN with SVM Precision : [illegible]
CNN with SVM Recall : [illegible]
CNN with SVM FMeasure : [illegible]
CNN with SVM Accuracy : [illegible]

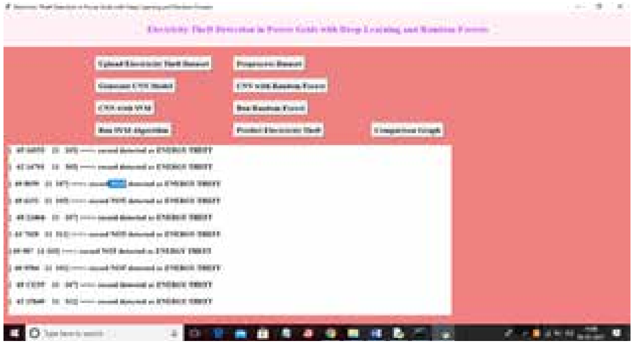
Random Forest Precision : [illegible]
Random Forest Recall : [illegible]
Random Forest FMeasure : [illegible]
Random Forest Accuracy : [illegible]

SVM Precision : [illegible]
SVM Recall : [illegible]
SVM FMeasure : [illegible]
SVM Accuracy : [illegible]

After achieving 96% accuracy using just SVM, you may proceed to submit test data by clicking the “Predict Electricity Theft” button on the top panel.

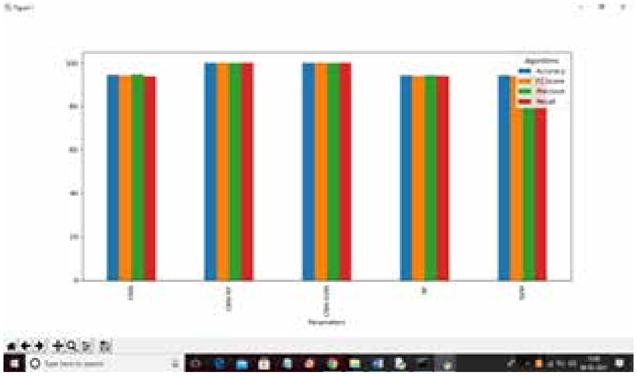


You may access the test data and see the prediction result below by going to the previous page, choose the “test.csv” file to upload, and then clicking the “Open” option.


Electricity Theft Detection in Power Grids with Deep Learning and Random Forest
Upload Electricity Theft Dataset
Preprocess Dataset
Generate CNN Model
CNN with Random Forest
CNN with SVM
Run Random Forest
Run SVM Algorithm
Predict Electricity Theft
Comparison Graph

[illegible] ----> record detected as ENERGY THEFT
[illegible] ----> record detected as ENERGY THEFT
[illegible] ----> record detected as ENERGY THEFT
[illegible] ----> record NOT detected as ENERGY THEFT
[illegible] ----> record detected as ENERGY THEFT
[illegible] ----> record NOT detected as ENERGY THEFT
[illegible] ----> record NOT detected as ENERGY THEFT
[illegible] ----> record NOT detected as ENERGY THEFT
[illegible] ----> record detected as ENERGY THEFT
[illegible] ----> record detected as ENERGY THEFT

In the screen above, the test data is included in square brackets that indicate the prediction result: “record detected as ENERGY THEFT” or “record NOT detected as ENERGY THEFT.” Click the “Comparison Graph” button to view the graph below.



All algorithms in the following graph are accurately executed by CNN-RF, with 100% precision. The x-axis shows the names of the algorithms, while the y-axis shows the metrics for recall, accuracy, FSCORE, and accuracy.

CONCLUSION

This research presents a novel CNN-RF architecture for detecting power theft. Here, the RF serves as the output classifier and the CNN checks wise metre data as an automated attribute extractor. In order to avoid optimising too many parameters, which increases the likelihood of over fitting, a fully connected layer with a failure rate of 0.4 is constructed during training. In addition, the SMOT technique is used to overcome the issue of data inequality. The exact same problem is used as a benchmark for applying several machine learning and deep learning algorithms, including SVM, RF, GBDT, and LR. All of these approaches have been tested on SEAI and LCL datasets. According to the findings, the suggested CNN-RF version is a very attractive classification method for detecting electrical power theft in homes: One advantage of hybrid design is that functions may be quickly extracted, unlike many other traditional classifiers that rely on the access to well-designed functions, which can be a time-consuming and tedious process. Since RF and CNN are two of the most widely used and successful classifiers for detecting electrical energy theft, the second part of the equation is that the hybrid model integrates their advantages. Future work will undoubtedly concentrate on examining how the granularity and duration of smart metre data may influence customers' personal privacy, since the identification of power theft impacts this privacy. It would be a worthwhile endeavour to look at expanding the proposed hybrid CNN-RF version to other uses, such as lots predicting.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "USING DEEP NEURAL NETWORKS TO DETECT ELECTRICITY THEFT IN SMART GRIDS", which gave us excellent resources and assistance to complete our assignment. I want to express my gratitude to our chairman, director, deans, head of the department, computer science and engineering department, guide, and teaching and non-teaching faculty members for their insightful advice and direction in all facets of our work.

REFERENCES

1. S. S. S. R. Depuru, L. Wang, and V. Devabhaktuni, "Electricity theft: overview, issues, prevention and a smart meter based approach to control theft," Energy Policy, vol. 39, no. 2, pp. 1007–1015, 2011.View at: Publisher Site | Google Scholar
2. J. P. Navani, N. K. Sharma, and S. Sapra, "Technical and non-technical losses in power system and their economic consequences on Indian economy," International Journal of Electronics and Computer Science Engineering, vol. 1, no. 2, pp. 757–761, 2012. View at: Google Scholar
3. S. McLaughlin, B. Holbert, A. Fawaz, R. Berthier, and S. Zonouz, "A multi-sensor energy theft detection framework for advanced metering infrastructures," IEEE Journal on Selected Areas in Communications, vol. 31, no. 7, pp. 1319–1330, 2013.View at: Publisher Site | Google Scholar
4. P. McDaniel and S. McLaughlin, "Security and privacy challenges in the smart grid," IEEE Security & Privacy Magazine, vol. 7, no. 3, pp. 75–77, 2009.View at: Publisher Site | Google Scholar
5. T. B. Smith, "Electricity theft: a comparative analysis," Energy Policy, vol. 32, no. 1, pp. 2067–2076, 2004. View at: Publisher Site | Google Scholar
6. J. I. Guerrero, C. León, I. Monedero, F. Biscarri, and J. Biscarri, "Improving knowledge-based systems with statistical techniques, text mining, and neural networks for non-technical loss detection," Knowledge-Based Systems, vol. 71, no. 4, pp. 376–388, 2014.View at: Publisher Site | Google Scholar
7. C. C. O. Ramos, A. N. Souza, G. Chiachia, A. X. Falcão, and J. P. Papa, "A novel algorithm for feature selection using harmony search and its application for non-technical losses detection," Computers & Electrical Engineering, vol. 37, no. 6, pp. 886–894, 2011.View at: Publisher Site | Google Scholar
8. P. Glauner, J. A. Meira, P. Valtchev, R. State, and F. Bettinger, "The challenge of non-technical loss detection using artificial intelligence: a surveyficial intelligence: a survey," International Journal of Computational Intelligence Systems, vol. 10, no. 1, pp. 760–775, 2017.View at: Publisher Site | Google Scholar
9. S.-C. Huang, Y.-L. Lo, and C.-N. Lu, "Non-technical loss detection using state estimation and analysis of variance," IEEE Transactions on Power Systems, vol. 28, no. 3, pp. 2959–2966, 2013.View at: Publisher Site | Google Scholar
10. O. Rahmati, H. R. Pourghasemi, and A. M. Melesse, "Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: a case study at Mehran region, Iran," CATENA, vol. 137, pp. 360–372, 2016.View at: Publisher Site | Google Scholar

Traffic Rules Violation Detection System by using Deep Learning

Sheggari Tejasri, Ankammagari Sumedha,
Ch Akshaya
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

B. P. Deepak Kumar
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ bhattudheepak@gmail.com

ABSTRACT

When it comes to socialism, security, and the compliance with safety regulations, as well as our own security concerns, real-time identification technologies are crucial. To prevent truck drivers from causing accidents or hitting people, traffic regulations are put in place. The rules of the road are crucial for everyone's safety. Additionally, they are to aid in managing the flow of online traffic for much improved efficiency. Management of Website Traffic According to that, offences are the primary cause of collisions, and India ranks first in the nation for road casualties. Because it is a manual procedure, the current system has some limitations when it comes to discovering guideline violations, on many occasions, we have found that the system becomes corrupted. An AI-developed system is another viable choice. Our technology can detect a wide variety of policy infractions, such as a truck driver disobeying a red light or failing to wear a safety helmet, among many others. One important aspect is the use of pre-installed video cameras to detect these infractions. By using an ML-based system, we can detect lawbreakers using image processing, get their licence plate numbers, categorise their offences as needed, and then punish them. Which will help make the enforcement of online traffic rules more effective.

KEYWORDS: *ML, CNN, Perdition, GUI, Sign, Auto fine, Speed detection, AI.*

INTRODUCTION

Worldwide, online traffic violations are becoming increasingly dangerous because of the increasing number of automobiles in densely populated places, which may cause massive amounts of traffic [1]. Extreme property damage and accidents may occur as a result, which may be frightening for people. We need web traffic violations finding technologies to fix the alarming problem and eliminate all those repercussions [2]. In order to do this, the system creates appropriate traffic regulations and applies penalties to those who do not adhere to them. As authorities track vehicles and roadways, a discovery system should be used to identify violations in real-time [3]. Because the technology can detect violations faster than humans, traffic regulators may utilise this to efficiently maintain safe streets. The online traffic violation system has an intuitive graphical user interface (GUI) that makes it easy for users to navigate, monitor, and respond to traffic infractions [4]. The capability to identify a prevalent kind of crime is present here. The primary focus of this system is to accurately detect and monitor the vehicles and their activities. In today's dynamic world, most developing countries are facing a significant issue with traffic regulation offences [5]. More and more people are breaking traffic laws, and the number of motorcycles on the road is also on the rise. It has always been difficult and risky to manage traffic in order to find violations. Despite this, it is still a formidable challenge since website traffic monitoring is now fully automated [6]. There were non-standard circumstances at the time of the picture's capture, including different plate size, turns, and lighting. Successfully managing website traffic rule breaches is a crucial component of this profession [7]. An automated method for taking pictures using a computer and a camera is part of the proposed upgrade [8]. The work provides Automatic Number Plate Recognition (ANPR) services, as well as additional image control techniques for plate localization and character recognition, to help speed up and ensure accurate number plate recognition. Once the registration number is known, the SMS-based component notifies the vehicle owners of their traffic violations [9]. All that is needed for number plate discovery in this project is the capacity to instantly extract and identify the personalities of a car number plate from an image [10, 11]. Using a built-in personality identification programme, this device is able to take pictures, identify specific people in them, and then extract their identities. Extensive controls are necessary to prevent accidents on motorcycles because of their widespread usage and cheap cost. Damage to safety helmets may result in substantial costs, since they are mandated by traffic laws [12].

LITERATURE SURVEY

Tong, Aniruddha, et al. In the year 2020 The proposed system first uses YOLO-based object identification to find motorcycles, and then it examines each bike for specific violations, such not wearing a helmet or not crossing the street at the designated crosswalk. In order to identify cases of headgear infringement, a classifier based on a Convolutional Neural Network (CNN) is used. Among those cited is Ruben J. Franklin together with colleagues. In the year 2020 An effective tool for monitoring and penalising website traffic breaches is an infraction finding system that uses computer vision. We advise this system to be constructed using YOLOV3 items discovery for website traffic violation discoveries such signal violation, bike speed, and motorbike count. [2]

A group of researchers led by Chetan Kumar B. In the year 2020 Convolutional semantic networks (CNNs) and other things discovery formulae are used by applications that monitor online traffic. One surprise layer is required for each input and output of a neural network. the third

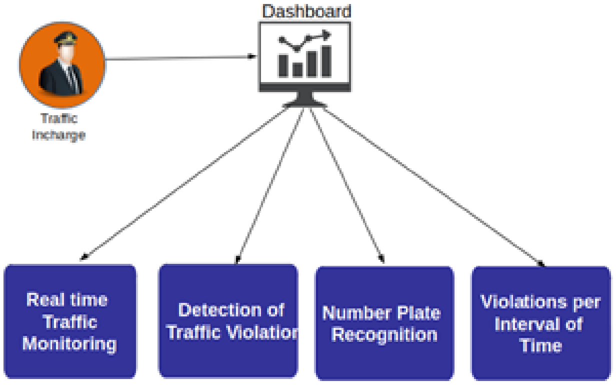
The other authors are Siddharth Tripathi. In the year 2019— They have really used a clever called CBITS in this little piece. Functions like pollution surveillance and accident detection are among those that will be reviewed. Regarding point [4], the authors are Helen Rose Mampilayil and colleagues. In the year 2019— This research presents a method that can detect, automatically and without human intervention, infractions of one-way online traffic regulations. Since three-wheeled vehicles were more likely to violate one-way traffic restrictions, they were considered. [5]

EXISTING SYSTEM

Using the capabilities of these high-monitoring systems and combining them with Deep Learning to identify the offences is the concept of our solution. Human error and technical restrictions will be eradicated with this system. For the sake of socialism, safety, and regulatory compliance, as well as our own safety and security, real-time identification technologies are very critical. Legislation enacted to protect pedestrians and drivers from harm emphasises the significance of traffic laws for the protection of all road users. They are also meant to help control the flow of website traffic for better efficiency. Penalties for violating online traffic standards may range from a fine to a prison sentence or even a ban for driving trucks altogether, depending on the severity of the offence. It detects vehicles that transgress the traffic rules outlined on the internet, including but not limited to: failing to use a turn signal, driving in the other way, not donning a safety helmet, and many more offences. These lawbreakers often get away with it because of clerical or technological mistakes, but accidents may also happen from time to time.

PROPOSED SYSTEM

Deep learning algorithms like Convolutional Neural Networks (CNNs) may take in an input picture, categorise its elements according to their relative significance (using learnable weights and predispositions), and then distinguish between them. When compared to other category formulae, the amount of pre-processing required by a ConvNet is much lower. With proper training, ConvNets may discover these filters/characteristics, in contrast to crude systems that need hand-engineered filters. The visual cortex served as an inspiration for the design of convolutional neural networks (CNNs), which mimic the connection pattern of neurons in the human brain. Every single neuron has a very small "Responsive Field" (the area of the visual field where it is most sensitive to stimuli). All of the visual space is covered by an overlap of such fields. By using the right filters, a ConvNet may effectively capture an image's spatial and temporal relationships. The approach is more suited to the picture dataset as a whole because of the reuse of weights and the lowering of criteria. The network may be trained to better recognise the picture's class, to put it simply.


Dashboard
Traffic Incharge
Real time Traffic Monitoring
Detection of Traffic Violatior
Number Plate Recognition
Violations per Interval of Time

METHODOLOGY

The system is fed video footage as input and the objects in motion are detected. The moving autos are sorted into certain classes using a detection version called YOLO version 3. The third iteration of the YOLO home detection algorithm is YOLOv3. Using a variety of methods, it ensures that data remains intact and can identify things with more precision. In order to create the classifier version, the Darknet-53 architecture is used. Car categorization serves the following purposes:

- Projection of Bounding Boxes
- Course Forecasting
- Predictions spanning various

Feature Extraction Module Forecasts for Bounding Boxes: Since it is a standalone network, each component of its convolutional area must be assessed independently in order to provide real-time object detection and categorization. By using logistic regression, this algorithm becomes ready for a respectable score. Here, 1 represents the overall overlap of the bounding boxes on the objects. Any erroneous assumption in this process results in loss of classification and identification, and it only predicts one bounding box ahead for one ground fact item. Additional bounding box priors will exist, the values of which may fall between the best and threshold limits. Except for category loss, these types of blunders will only result in recognition loss. For class prediction, this approach ditches the standard Softmax layer in favour of course-specific logistic classifiers. In order to achieve multi-label classification, this procedure is executed. Using the multi tag category, each box is able to identify the possible classes it contains. Projection on various scales: It detects boxes at three distinct scales for range recognition. Afterwards, qualities may be extracted from each scale using a method akin to functioning pyramid networks. YOLOv3 gains the ability to make predictions at various ranges by using the aforementioned method. A maximum of three bounding boxes priors per scale may be achieved by dividing the bounding boxes priors generated from measurement clusters into three ranges. So, in all, there will be nine bounded boxes.

Retrieve Attributes: Compared to YOLO version2, it uses a new network called Darknet-53, which includes more features and 53 convolutional layers. Compared to Darknet-19, it is much stronger. Additionally, it outperforms ResNet-101/ResNet-152 in terms of efficacy.

Vehicles are located by use of the YOLOv3 version. The inspection of infraction cases follows the detection of trucks. The customer's attention is drawn to a traffic queue crossing the road in the preview of the provided video clip material. Because of the placement, this line indicates that the traffic light is red. A bordering box of eco-friendly colour surrounds the objects as they are spotted. Moving trucks are in breach of traffic laws if they cross the cent reline at the red light. Following the detection of the infringement, the shade of the bounding box around the truck becomes red.

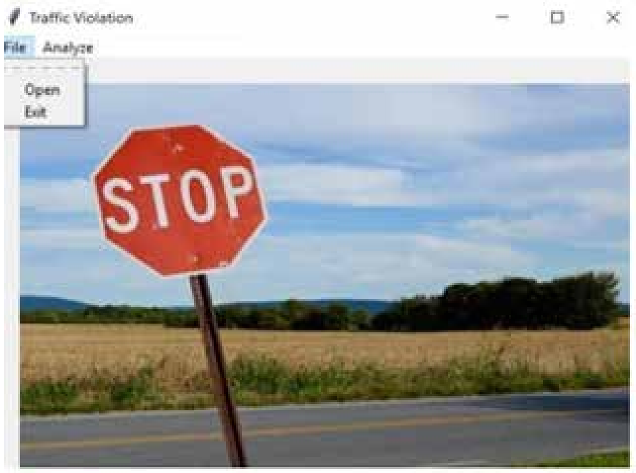


Fig.1. Home page

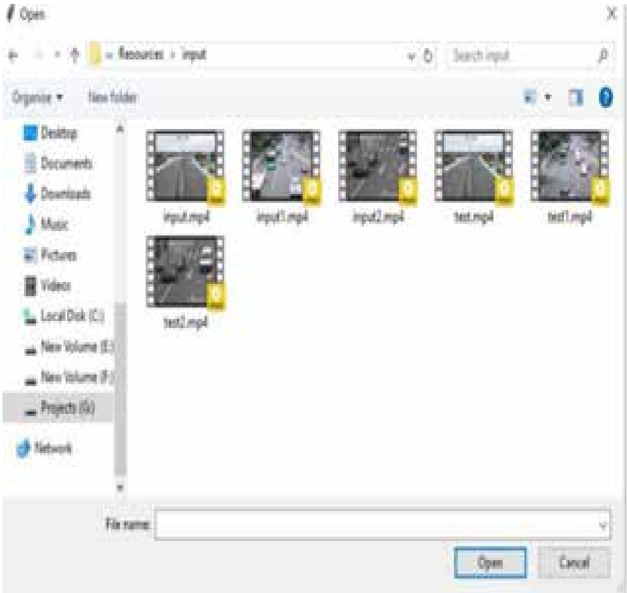


Fig.2. Input video

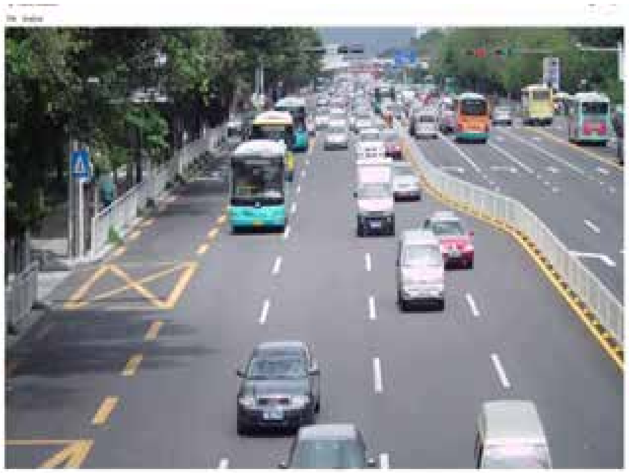


Fig.3. Vehicle detection system

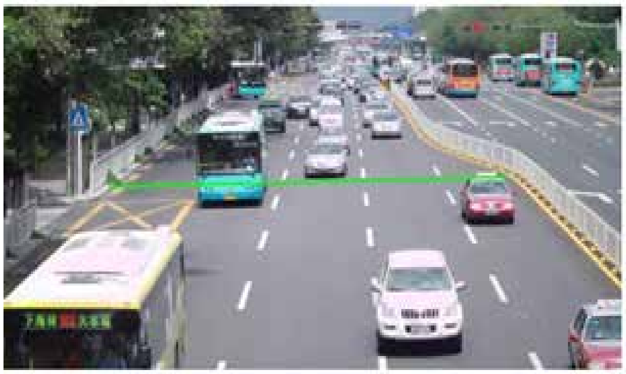


Fig.4. Driving signal line

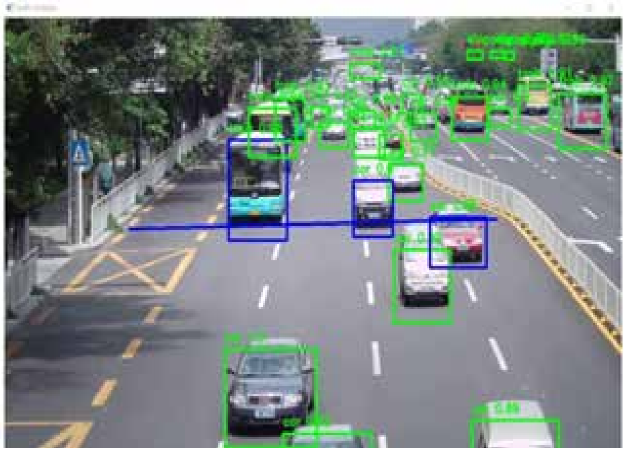


Fig.5. Vehicle classification.

CONCLUSION

Driving ahead of the designated traffic queue is considered a traffic offence on the road. Compared to humans, the proposed system is much more efficient and operates much faster. Most of us know that traffic policemen are the ones who record individuals breaking website traffic laws, however online traffic cops can't detect and record many violations at once. The following formula successfully identified the kind of violation specified above, namely, crossing the traffic signal. For violations of traffic signals, the current system provides detection. Also, the machine can definitely refine one piece of data at a time. The program's sluggish runtime is another area that may benefit from a machine with a high processing rate. Since the current formula increases the system's runtime by ignoring several other needless items, it necessitates more research before it can be applied to other high-level picture processing techniques. We can improve the current system's performance by replacing the current algorithm with one from OpenCV. Our next goal is to enhance the system's functionality by detecting the licence plate of a vehicle that disobeys the traffic signal and by adding other online traffic offence issues.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "TRAFFIC RULES VIOLATION DETECTION SYSTEM BY USING DEEP LEARNING", which provided good facilities and support to accomplish our work. I sincerely thank our

Chairman, Director, Deans, Head of the Department, Department of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. Aniruddha Tonge, S. Chandak, R. Khiste, U. Khan and L. A. Bewoor, "Traffic Rules Violation Detection using Deep Learning," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1250-1257,doi: 10.1109/ICECA49313.2020.9297495.

2. Ruben.J Franklin and Mohana, "Traffic Signal Violation Detection using Artificial Intelligence and Deep Learning, "2020 5th International Conference on Communication and Electronics Systems (ICCES), 2020, PP. 839-844, doi: 10.1109/ICCES48766.2020.9137873.

3. Chetan Kumar B, R. Punitha and Mohana, "Performance Analysis of Object Detection Algorithm for Intelligent Traffic Surveillance System," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 573,579,doi:10.1109/ICIRCA48905.2020.9182793.

4. Siddharth Tripathi, Uthsav Shetty, Asif Hasnain, Rohini Hallikar,"Cloud Based Intelligent Traffic System to Implement Traffic Rules Violation Detection and Accident Detection Units", Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019) IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978- 1- 5386-9439- 8.

5. Helen Rose Mampilayil and R. K., "Deep learning-based Detection of One-Way Traffic Rule Violation of ThreeWheeler Vehicles," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 1453- 1457, doi: 10.1109/ICCS45141.2019.9065638.

6. Ali Şentas, S. Kul and A. Sayar, "Real-Time Traffic Rules Infringing Determination Over the Video Stream: Wrong Way and Clearway Violation Detection," 2019International Artificial Intelligence and Data Processing Symposium (IDAP), 2019, pp. 1-4, doi:10.1109/IDAP.2019.8875889.

7. M. Purohit and A. R. Yadav, "Comparison of feature extraction techniques to recognize traffic rule violations using low processing embedded system," 2018 5thInternational Conference on Signal Processing and Integrated Networks (SPIN), 2018,pp. 154-158, doi: 10.1109/SPIN.2018.8474067

8. S. P. Mani Raj, B. Rupa, P. S. Sravanthi and G. K. Sushma, "Smart and Digitalized Traffic Rules Montioring System," 2018 3rd International Conference on Communication and Electronics Systems (ICCES), 2018, pp. 969-973, doi: 10.1109/CESYS.2018.8724086.

9. Shashank Singh Yadav, V. Vijayakumar and J. Athanesious, "Detection of Anomalies in Traffic Scene Surveillance," 2018 Tenth International Conference on Advanced Computing (ICoAC), 2018, pp. 286-291, doi: 10.1109/ICoAC44903.2018.8939111.

10. N. Krittayanawach, et al., " Robust Compression Technique for YOLOv3 on Real-Time Vehicle Detection", 11th International Conference on Information Technology and Electrical Engineering (ICITEE),Pattaya,Thailand,2019.

11. github.com/anmspro/Traffic Signal Violation-Detection-System.

12. Mohana, et.al., "Simulation of Object Detection Algorithms for Video Survillance Applications", 2 ndInternational Conference on I-SMAC (IoT in Social, Mobile ,Analytics and Cloud),2018.

13. Yojan Chitkara, et. al.," Background Modelling techniques for foreground detection and Tracking using Gaussian Mixture model" International Conference on Computing Methodologies and Communication,2019.

14. Mohana, et.al., "Performance Evaluation of Background Modeling Methods for Object Detection and Tracking," International Conferenceon Inventive Systems and Control,2020.

15. N. Jain, et.al.,"Performance Analysis of Object Detection and Tracking Algorithms for Traffic Surveillance Applications using Neural Networks," 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud),2019.

Early Pest Detection from Crop using Image Processing and Computational Intelligence

Bathula Pravalika, Mohammed Mubeen
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

Voruganti Naresh Kumar
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ nareshkumar99890@gmail.com

ABSTRACT

One of the biggest problems in the agricultural sector is the rapid detection of bugs. Using insecticides is the most convenient way to handle the bug infestation. But pesticides, when used in excess, are harmful not just to humans but also to plants and animals. Parasite infection prevention is the goal of integrated pest control, which uses both physical and biological methods. Digital image processing and equipment vision In the realm of agricultural scientific inquiry, processing finds extensive use; this is particularly true in the area of plant security, where it inevitably leads to the administration of crops. This study examines a novel method for the first identification of parasites. Using a digital video camera, one may capture images of the fallen leaves that have been affected by parasites. Using attribute extraction and picture classification algorithms, we can detect insects on fallen leaves by refining the photographs of insects on the leaves until they become a grey image. Using a digital video camera, the pictures are captured. Once transferred to a computer, the images are then rendered using the MATLAB programme. After that, the picture is converted to a gray scale version using the RGB photo, and then the feature extraction methods are performed on it. To determine which types of pests are present, the Support Vector Device classifier is used.

KEYWORDS: *MATLAB, RGB, Crop, Image processing, Pest, SVM, AI.*

INTRODUCTION

Farming is the main economic activity in India. Typically, farming is the main source of income for 70% of the population [1]. Therefore, increasing plant efficiency is a pressing issue right now. Investigations in this field are being conducted by the majority of researchers. This becomes a piece of cake when you use their cutting-edge methods and put them into practice. However, “bug infection” on plants is today one of the most important issues. Greenhouse plants are the main subject of this article [2]. A variety of plants are grown in greenhouses. For instance, fruits and vegetables such as cucumbers, tomatoes, potatoes, and so on, as well as flowering plants like roses and jasmine. The three most prevalent types of insects that may harm these verdant houseplants are trips, aphids, and white-flies [3]. The use of pesticides is one strategy for controlling the insect infestation. Certain types of insects may be controlled with the use of pesticides. Chemicals have a negative impact on ecosystems and the environment [4].

Air, water, and soil will undoubtedly be polluted by the excessive use of chemicals. Pesticide suspensions are carried by the wind and end up contaminating other places. Our primary emphasis in this study is on finding bugs early on. This necessitates keeping an eye on the plants from time to time [5]. The usage of cameras allows for the acquisition of photos. Next, photo processing procedures must be used to the obtained picture in order to analyse the picture materials [6]. Picture interpretation for insect identification is the main focus of this article.

Farming is the main industry in India. The agricultural sector typically provides income for 70% of the population. Therefore, increasing plant efficiency is a key focus. This area is receiving a lot of attention from researchers. This is a breeze when you use their cutting-edge methods and practical tools [7]. The problem of "pest infection" on plants is one of the most important ones right now. The focus of this article is mostly on crops grown in greenhouses. In greenhouses, many different kinds of crops are grown. vegetables (cucumber, potato, tomato, etc.) and flowers (rose, jasmine, etc.) are examples. Whiteflies, aphids, and thrips are among the most common pests that may harm these verdant houseplants. Using pesticides is one way to control the bug infestation [8]. Some pests can be controlled with the use of pesticides. Toxic pesticides degrade ecosystems and harm the air we breathe. When chemicals are used excessively, they contaminate the air, water, and soil. Dispersions of pesticides contaminate different areas when carried by the wind. The focus of this study is the first identification of insects [9]. This indicates the need for regular plant monitoring. By use of cameras, images are captured [10]. The next step is to apply image processing techniques to the acquired picture in order to analyse the photo materials. Picture analysis for pest finding is the main focus of this article [11].

LITERATURE SURVEY

Here, we'll take a look at the pros and cons of some of the current techniques used to spot parasites in greenhouse plants early on. Below, we have outlined the strategies along with their benefits and drawbacks.

Finding Insects with the Use of Video Analysis

Photo processing techniques and a knowledge-based approach are both included in this role. [1] Whiteflies are the only insects it can detect. When compared to the results obtained by manual methods, this system's accuracy and reliability are far higher. In reality, it is an interdisciplinary cognitive vision system that uses a wide variety of methods, including computer vision, AI, image processing, and more. For this task, they used white flies as the screening parasite and rose plants as the screening crop. It was not easy to get detected. As a result, they collected adult flies. Nevertheless, there were also some problems with the detection of adulthood. At any point throughout the picture-taking process, the adult may take flight. Therefore, they decided to scan the rose leaves while the flies weren't active. Finding white flies at the start of the project is the future goal.

Method that employs Delicate Snares

Using video clip analysis, the goal of insect detection via camera network [2] is to identify parasite infections on plants. Finding and counting the insects using the conventional methods would undoubtedly take much more time. This is why they came up with an automated technique that uses video assessment. Five wireless electronic cameras were used in the greenhouse. As a crop to test, they chose climbing. This task requires the usage of sticky traps. All it takes to set up a sticky trap is a sticky substance with coloured dots to entice bugs. Their method for pest identification included using video division algorithms in conjunction with finding and adaption strategies. No matter the weather, the adaptive system will work. The long-term goal of this technology is the early detection of novel insect species.

PROBLEM STATEMENT

It is common practice to use computer vision, artificial intelligence, or deep understanding breakthroughs to identify plant conditions; however, it is uncommon to compare many methods for the same task; instead, just one way is often used. The majority of automated bug detection and identification services only consider a single technology solution, rather than exploring all of the possible options. Recent years have seen tremendous progress in computer vision and object identification. Item classification and object identification are only two examples of the many visualization-related computer vision issues that use ILSVRC the ImageNet public dataset as their standard. Canine, Salient Regions, SURF, SIFT, MSER, and other feature detection algorithms used to be the gold standard for picture categorization problems. With these features, several finding methods are used for function extraction.

METHODOLOGY

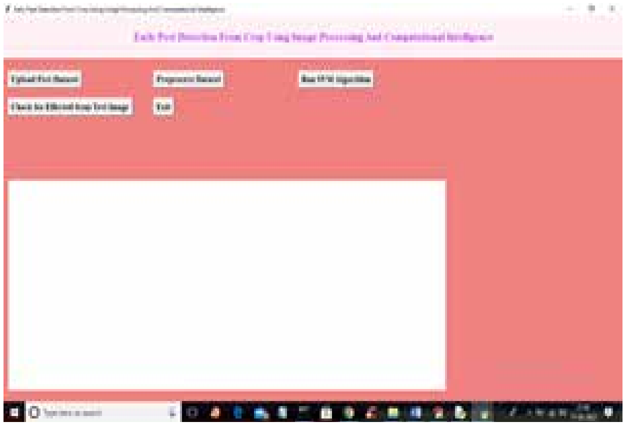
Upload Pest Dataset

This module is used to upload datasets to the application.

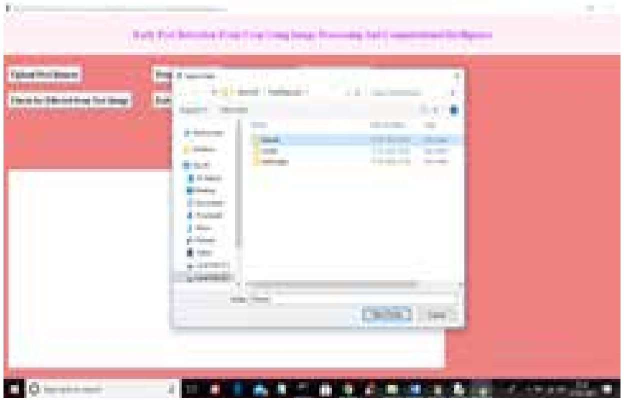
Then, we'll pre-process the datasets by acquiring images from the dataset, filtering them to greyscale, normalising them, and finally, splitting the dataset into a train and test part. About 80% of the images will be used for training, and 20% will be used for testing.

To run the SVM algorithm, we will feed the processed photos into it as training data, and then we will determine how accurate its predictions are.

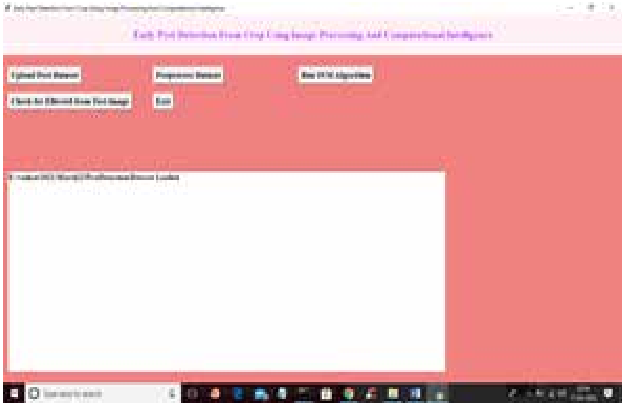
This module allows us to input a test picture and use support vector machines to forecast if the image contains aphids, white fly larvae, or is unaffected.


Early Pest Detection From Crop Using Image Processing And Computational Intelligence
Upload Pest Dataset
Preprocess Dataset
Run SVM Algorithm
Check for Effected from Test Image
Exit

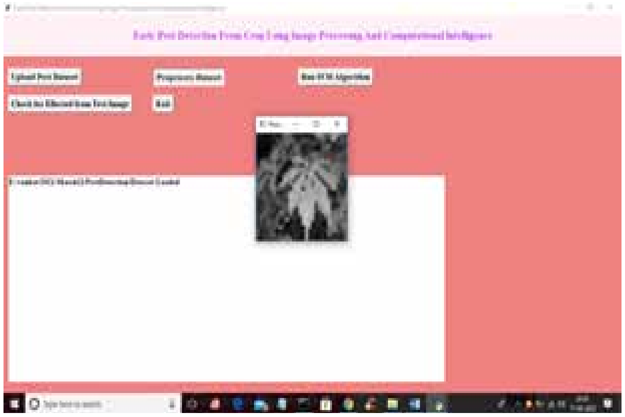
To upload a dataset, go to the first screen and look for the "Upload Pest Dataset" button.


Early Pest Detection From Crop Using Image Processing And Computational Intelligence
Upload Pest Dataset
Check for Effected from Test Image
Dataset
Select Folder
Cancel

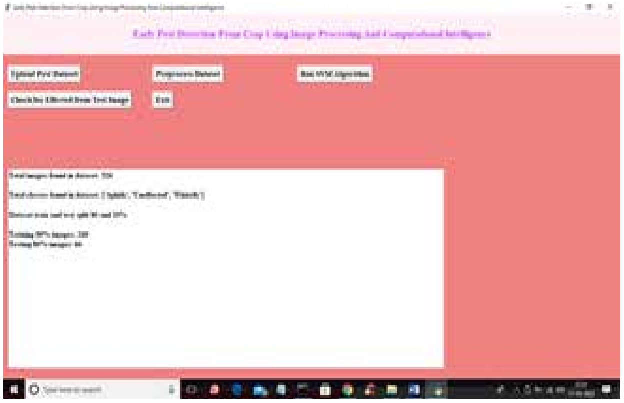
To load the dataset, go to the first screen and choose the "Dataset" folder. Then, click the "Select Folder" button. This will bring you to the second screen.


Early Pest Detection From Crop Using Image Processing And Computational Intelligence
Upload Pest Dataset
Preprocess Dataset
Run SVM Algorithm
Check for Effected from Test Image
Exit

To read and normalise the pictures in the loaded dataset, split it into a train and test set, and then click the "Preprocess Dataset" button on the top screen.


Early Pest Detection From Crop Using Image Processing And Computational Intelligence
Upload Pest Dataset
Preprocess Dataset
Run SVM Algorithm
Check for Effected from Test Image
Exit

The upper screen is showing a processed grey picture; to see the screen below, dismiss the image.


Early Pest Detection From Crop Using Image Processing And Computational Intelligence
Upload Pest Dataset
Preprocess Dataset
Run SVM Algorithm
Check for Effected from Test Image
Exit

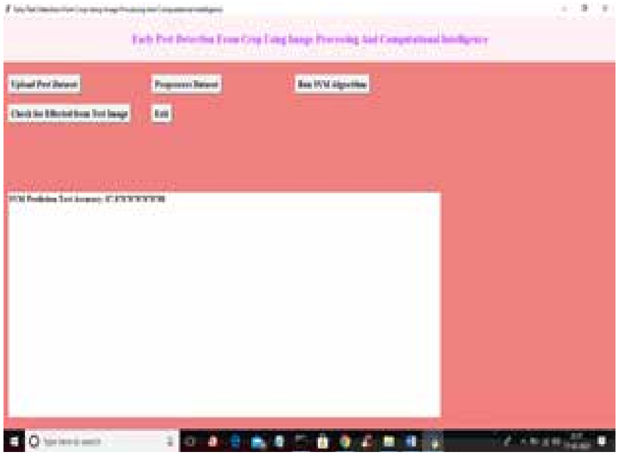
Total images found in dataset: [illegible]

Total classes found in dataset: ['Aphids', 'Unaffected', 'Whitefly']

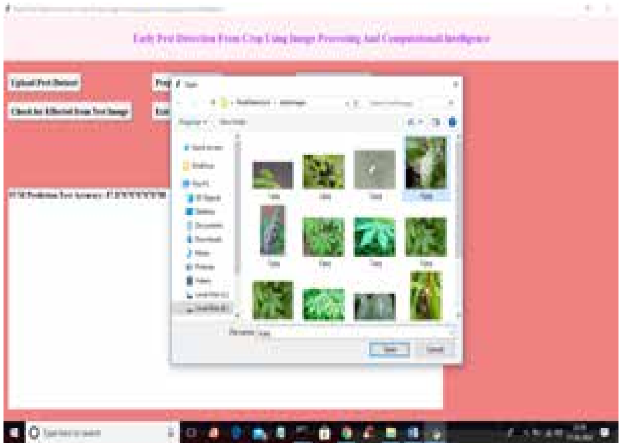
Dataset train and test split 80 and 20%

Training 80% images: [illegible]
Testing 20% images: [illegible]

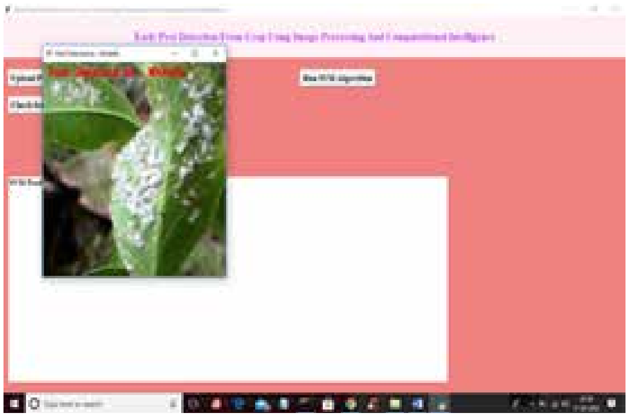
Click the "Run SVM Formula" button to train SVM with improved photographs and then calculate its prediction accuracy. On the above screen, you can observe a range of images and courses found in the dataset.


Early Pest Detection From Crop Using Image Processing And Computational Intelligence
Upload Pest Dataset
Preprocess Dataset
Run SVM Algorithm
Check for Effected from Test Image
Exit

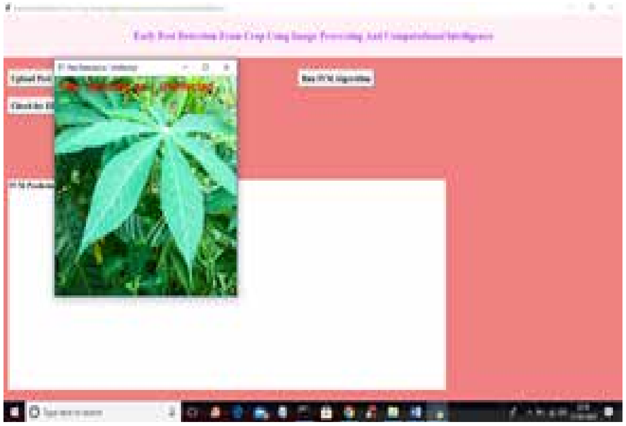
After you've achieved 87% prediction accuracy using SVM, as demonstrated on the previous page, click the "Check for Effected from Test Image" button to submit a test image that looks like the one below.


Early Pest Detection From Crop Using Image Processing And Computational Intelligence
Upload Pest Dataset
Check for Effected from Test Image
Open
Open
Cancel

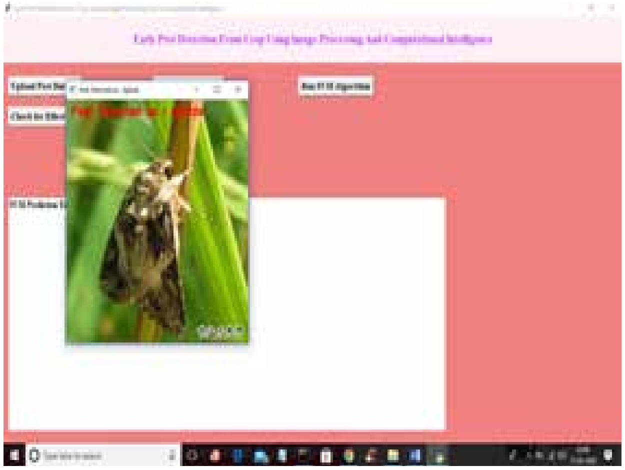
Click the "Open" button after choosing and uploading the 4.jpg file on the previous screen to get the following result.


Early Pest Detection From Crop Using Image Processing And Computational Intelligence
Run SVM Algorithm

You may submit and test more photographs in the same way; in the following panel, the red text indicates that SVM predicted or classed the uploaded image as "whitefly."


Early Pest Detection From Crop Using Image Processing And Computational Intelligence
Run SVM Algorithm

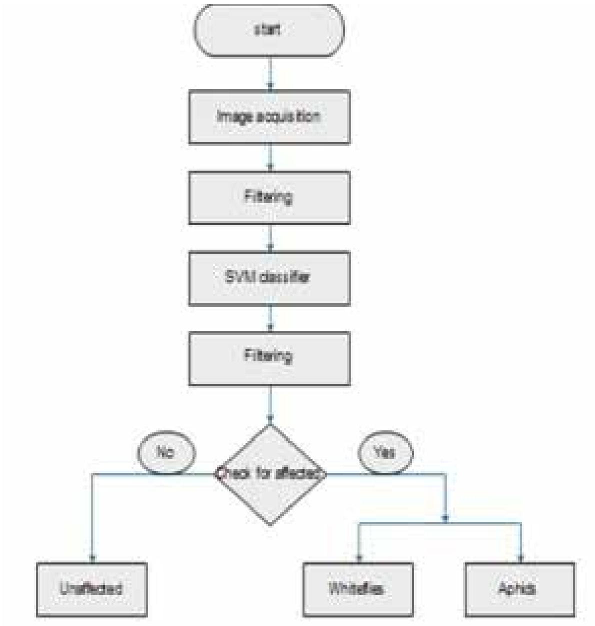
The uploaded picture is expected to be labelled as "Uneffected" on the above screen since it does not contain any pests.


Early Pest Detection From Crop Using Image Processing And Computational Intelligence
Run SVM Algorithm

In the above screen, uploaded image is classifier as 'Aphids'.

FLOWCHART

Figure 3 shows a flow diagram of the proposed system. A camera is used to capture the pictures, which are then filtered using bicubic filters to remove any unwanted noise. Actually, here is where the images are pre-processed. In order to identify the pest infection, svm classification is the next stage. Once again, it is applied to the svm in order to determine the kind of pest if the picture is impacted.



CONCLUSION

The field of pest detection has invested in image processing systems. Our main goal is to detect pests like trips, aphids, and white flies on greenhouse crops. We provide a fresh method for identifying pests at an early stage. Using a pan-tilt-zoom camera allows us to see finer details. So long as we do this, we can get the shot without disturbing the bugs. It exemplifies how teams of different backgrounds worked together to create a system that is both automated and very adaptable. Rapid pest identification was made possible by the prototype technology, just as promised. On top of being easy to use, it performs about the same as a regular manual method. We want to decrease pesticide use by detecting pests early on.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "EARLY PEST DETECTION FROM CROP USING IMAGE PROCESSING AND COMPUTATIONAL INTELLIGENCE", which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department Of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. Martin,V. Thonnat,. "A Learning Approach For Adaptive Image Segmentation." In Proceedings of IEEE Trans. Computers and Electronics in Agriculture.2008
2. Vincent Martin and Sabine Moisan, "Early Pest Detection in Greenhouses". INRIA Sophia Antipolis M´editerrann´ee, PULSAR team 2004 route des Lucioles, BP93
3. Jayamala K. Patil1 , Raj Kumar, "Advances In Image Processing For Detection Of Plant Diseases" Journal Of Advanced Bioinformatics Applications and Research ISSN 0976-2604 Vol 2, Issue 2, June-2011, pp 135-141
4. Ikhlef Bechar and Sabine Moisan, "On-line counting of pests in a greenhouse using computer vision". published in "VAIB 2010 - Visual Observation and Analysis of Animal and Insect Behavior (2010)"
5. Paul Boissarda, Vincent Martin, "A cognitive vision approach to early pest detection in greenhouse crops". computers and electronics in agriculture 62 (2008) 81–93
6. B.Cunha."Application of Image Processing in Characterisation of Plants." IEEE Conference on Industrial Electronics.2003.
7. Sreelakshmi m1, Padmanayana2" Early detection and classification of pests using image processing" international journal of innovative research in electrical, electronics, instrumentation and control engineering
8. Muhammad Danish Gondal,Yasir niaz khan "early pest detection from crop using image processing and computational intelligence".
9. Paul boissarda,Vincent Martin,"a cognitive vision approach to early pestdetection in green house crops".
10. Dheeb al Bashish, Malikbraik "detection and classification of leaf diseases using k means based segmentation and neural networks based classification."
11. Long zhou , Xiao-jun Tong IEEE explore digital library "application of two dimension wavelet transform in image process of pests in stored grain".
12. Preethi Rajan, Radhakrishnan B " a survey on different image processing techniques for pest identification and plant disease detection".ijscn international journal of computer science and network.
13. S. Arivazhagan, R. Newlin Shebiaj, S. Ananthi and S. Vishnu Varthini," detection of unhealthy region of plant leaves and classification of plant leaf disease using texture features".

Comparison of Machine Learning Algorithms for Predicting Crime Hotspots

Padigela Sahithi Reddy, Kirthi Santhoshilaxmi
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

K Srujan Raju
Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ ksrujanraju@gmail.com

ABSTRACT

Criminological forecasting is essential for formulating and implementing police policies. Predictions are now most often made using machine learning. However, there has been a dearth of thorough comparisons including the use of machine learning to the task of crime prediction. This study uses public property crime history data from 2015–2018 from a section of a large coastal city in southeastern China to assess the prediction capabilities of several machine learning methods. Based only on findings obtained from historical crime data, the LSTM model outperformed other models such as convolutional neural networks, naive Bayes, support vector machines, random forests, and KNN. Urban road network density and points of interest (POIs) are two examples of the built environment data used to train the LSTM model. The model that takes built environment characteristics into account outperforms the one that used historical crime data alone in terms of forecast accuracy. As a result, elements connected to criminological theory and past crime statistics should be considered when making predictions about the future. When it comes to criminal conduct prediction, not all ML models are created equal.

KEYWORDS: *KNN, POI, LSTM, Crime, Hotspots, Public data.*

INTRODUCTION

In recent years, there has been an exponential growth in the amount of spatial and temporal data linked to public safety in general. But not all data has been put to good use in solving genuine problems. Several researchers have developed versions to predict criminal behaviour, which may aid in crime prevention. [1] When calibrating their prediction algorithms, many relied only on data from past criminal activities. The current state of crime prediction research is primarily concerned with two main areas: forecasts of criminal activity danger areas [2, 3] and hotspots [4, 5]. Based on the "routine activity theory" and other relevant factors, criminal activity threat location prediction takes into account the relationship between criminal tasks and the physical environment. [6] Conventional approaches to estimating the likelihood of criminal offenses often use crime statistics to identify problem areas, with the expectation that this trend will persist into the future. [7] Case in point: the surface risk version is effective for long-term, stable hotspot prediction of criminal activity by considering crime-related environmental factors and criminal activity background data, with an eye towards crime-related gatherings and locations. [2] in A number of studies have integrated data from mobile phones, criminal records, land usage, financial statistics, and market and financial statistics to conduct empirical study on criminal activity predicted across varying time periods. A criminal offence hotspot prediction aims to predict where future crime occurrences are most likely to cluster [8]. Estimating bit densities is a common method [9]-- [12] Designing with auto-correlations of past occurrences in mind, whether temporal or geographical, yields better results than designs that don't. in [13] Algorithms based on artificial intelligence have recently become rather popular. K-Nearest Neighbours (KNN), the random forest technique, neural networks, support vector machines (SVMs), Bayesian models, and many more are among the most widely used approaches. [6] There were comparisons between linear techniques for criminal offence pattern forecasting, Bayesian models and BP semantic networks, and approaches to crime forecasting based on spatiotemporal bit thickness and random forests. [12] One of the most successful supervised learning algorithms among them is KNN. A popular maker-discovery variant, the SVM can do more than simply do categorization and regression tasks; it can also spot outliers. Multiple domains have shown the Random Forest formula's great prediction accuracy and robustness to non-linear relational data. An ageless category algorithm, Naive Bayes (NB) does not take into account missing data and uses just two criteria. To handle even more facility category concerns, convolutional semantic networks (CNNs) may increase their expression power with a truly deep layer, and they have excellent extensibility. One of the most influential neural networks for processing data with significant time collection trends is the Long Short-Term Memory (LSTM) network, which extracts time-series properties from functions. This study will compare and contrast the six machine learning formulae mentioned above, and then recommend the optimal one for demonstrating predictive power with and without covariates.

SURVEY OF RESEARCH

A. The current study delves into the geographical patterns of criminal activity and its variables in Portland, Oregon, analysing an architectural design of violent crime. The study provides results from an international ordinary least squares version that are taken from a random structure, using conventional structural activities that are deemed to be applicable for all places within the research study region. Then, as a substitute for these conventional approaches to crime modelling, we provide geographically weighted regression (GWR). A collection of mappable parameter quotations and room-varying t-values of value are produced by the GWR technique, which approximates a neighbourhood version.

B. We generate an enormous amount of data via social networks. More than 230 million people use the microblogging service Twitter every day, and they post more than 500 million tweets. Recommend using Twitter's publicly available data to predict future crime rates. The crime rate has been on the rise recently. While there are a number of strategies being used by crime stoppers to reduce crime rates, none of them have focused on using the language used (offensive vs. non-offensive) in Tweets as a source of information to predict crime rates. This research proposes the hypothesis that analysing tweets for linguistic patterns might be a useful method for predicting citywide crime rates.

C. Information mining, which is short for "information evaluation methods," has recently been popular for analysing data on criminal behaviour that has been collected from various sources in order to identify trends and patterns. It may also be linked to immediately inform offenders and used to increase efficiency in repairing crimes quicker. But data mining techniques abound.

D. There are two very important goals to the research study that is detailed below. The first is to foretell illicit gun behaviour using risk surface modelling (RTM). In order to assess the dispersion of future shooting threats, RTM's danger landscape maps incorporate a wide range of contextual information pertinent to the probability structure of captures.

EXISTING SYSTEM

Cohen and Felson jointly proposed routine task theory in 1979, and it has since been refined via the incorporation of additional ideas. Motivated criminals, suitable targets, and an inability to defend in terms of both time and distance are the three components that, according to this theory, must come together for the commission of the vast majority of crimes, especially predacious crimes.

Cornish and Clarke advocated for the logical selection hypothesis. The idea is that the criminal's decisions about location, objectives, and tactics may be better understood in light of the reasonable trade-offs between effort, danger, and profit. By combining the reasonable selection idea with the regular duties theory, criminal activity pattern theory provides a clearer explanation of the geographical distribution of criminal occurrences. People build their "cognitive map" and "activity space" by the things they do on a daily basis. Also, potential criminals need to use their mental maps to zero in on specific spots in a somewhat familiar area to conduct crimes. The perpetrator of a crime will choose places where the "criminal possibility overlaps with cognitive room" according to their reasonable competence, rather than avoiding areas they are unfamiliar with. Because of their blatant characteristics of "creating" or "attracting" criminal offence, these locations inevitably become crime hotspots. For this reason, in addition to crime statistics, environmental factors in the locations should be considered when trying to identify potential trouble zones.

PROPOSED SYSTEM

For crime forecasting, the proposed system employs the arbitrary woodland algorithm, KNN, SVM, and LSTM algorithms. The first step in calibrating the versions is to use only past criminal offence data. By comparing, the most dependable design would be shown. The second step is to increase the forecast accuracy by adding established setting data like road network density and poi as variables to the predictive design. K-closest neighbour (KNN) algorithms take instance function vectors as input, calculate distance between training set and fresh data include value, and then choose the nearest K category. If k is less than 1, the data to be studied is the next-door neighbour course that is closest. Category choice regulation in KNN is based on distance-based heavy voting or bulk balloting. The input situation's categorization is determined by the majority of the k nearby training examples. Bayesian theory anticipates the incident potential of an occasion based on knowledge about the evidence of an occurrence in the area of probability and statistics. The naïve Bayes (NB) classifier is an AI category technique that assumes that each function is independent and is based on Bayesian theory. To solve the probability that an input object belongs to a certain class, NB classifiers rely on conditional probabilities.

WORKING METHODOLOGY

Distributor of Solutions

In order to access this component, the Admin must provide their valid login credentials. A variety of features will be available to him after his visit, including Criminal Activity Data Sets with Sight Information andFind information on convicted criminals, find Examine Areas Likely to See Increased Criminal Activity, In addition to all remote customers, you may see the results of crime ratio searches, counts of criminal activity, ratios of criminal activity discovered, information by place and day, and more. You may also see the results of crime percentages computed with the help of vector makers.

Authorization and Access for Customers

Here, the business may see all consumer data and provide them access to legal protections. Personal Information such as Name, Address, Email, and Phone Number of the User.

This component has a number of users. Before you start working on any projects, make sure you sign up. Upon registration, the customer's information will undoubtedly be added to the database. After enrollment is complete, he must log in using the permitted username and password. After coming to the site, users may see their profiles, search for crime statistics, and access collections of blog posts on criminal activities.

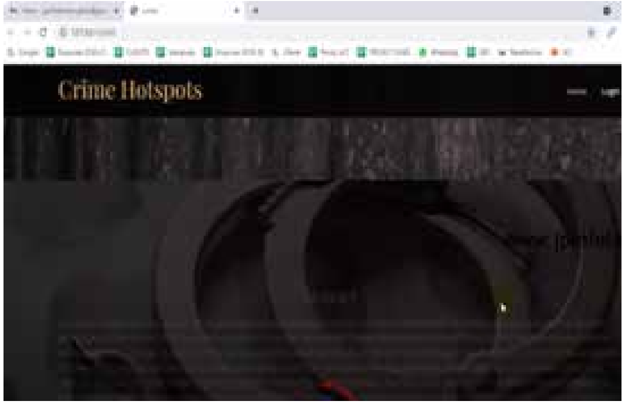


Fig.1. Home page

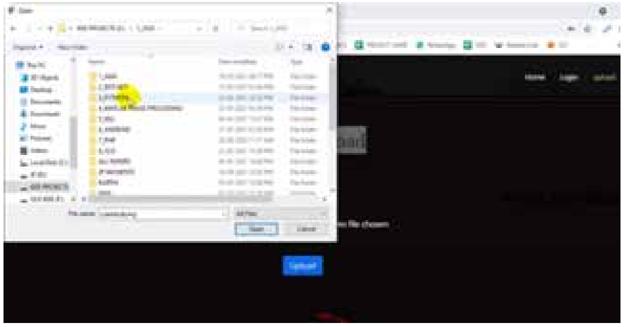
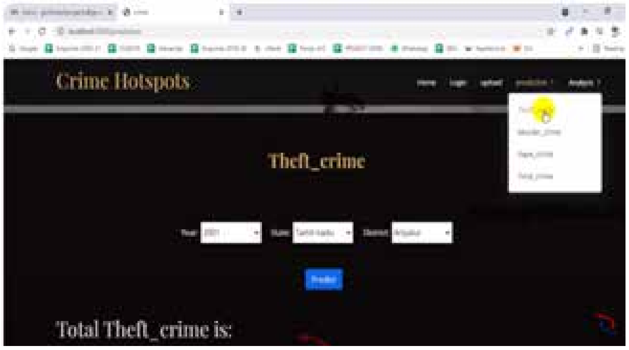


Fig.2. Data upload data

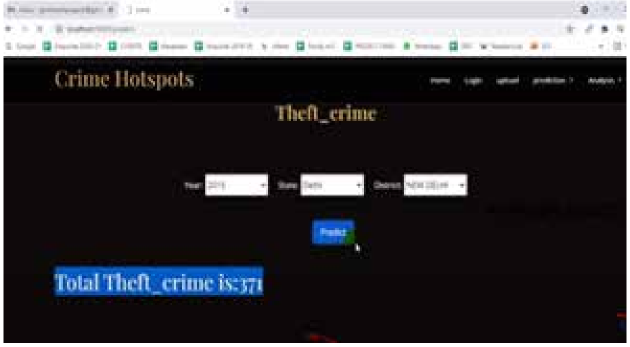


Crime Hotspots

Theft_crime

Total Theft_crime is:

Fig. 3. Theft crime detection

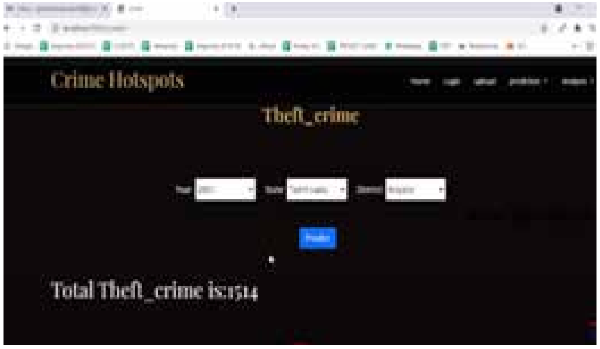


Crime Hotspots

Theft_crime

Total Theft_crime is:371

Fig.4. Total theft crime detection

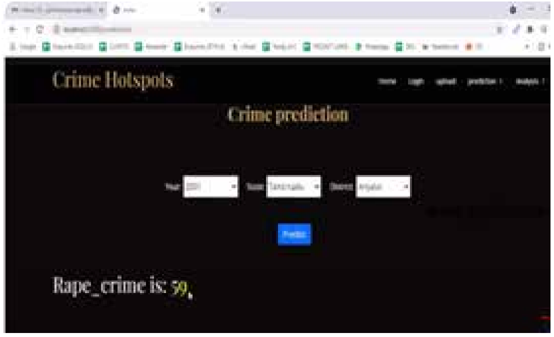


Crime Hotspots

Theft_crime

Total Theft_crime is:1514

Fig.5. Total theft crime detection

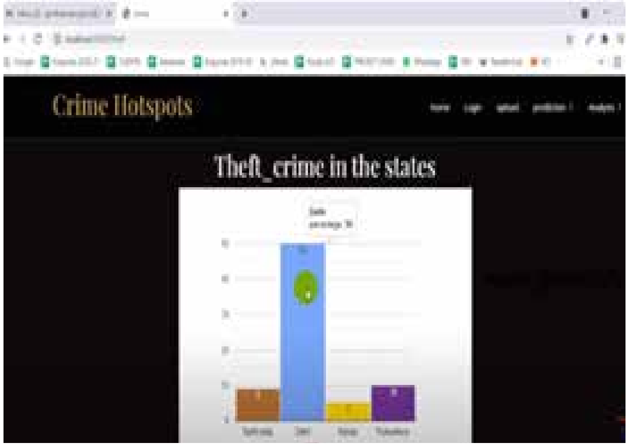


Crime Hotspots

Crime prediction

Rape_crime is: 59

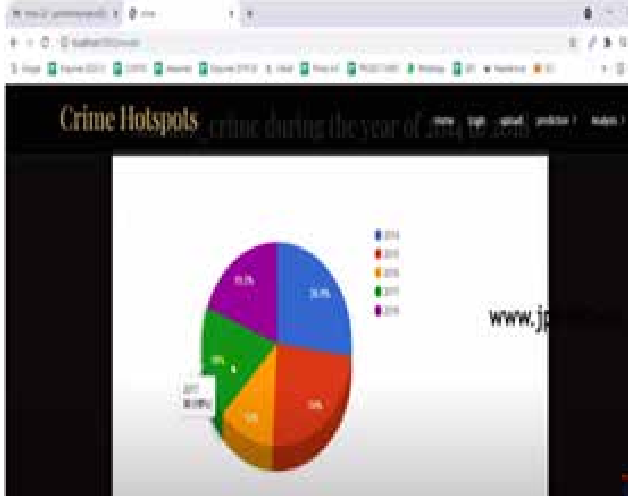
Fig.6. Type of crime detection



Crime Hotspots

Theft_crime in the states

Fig.7. Theft crime in states



Crime Hotspots

crime during the year of 2014 to 2018

Fig.8. Graph in different states.

CONCLUSION

In order to forecast the occurrence of crime hotspots in a community in a city in China’s southeastern coastal region, this article employs six artificial intelligence systems. The following findings are drawn to:1) Compared to other models, the LSTM version has much higher forecast precision’s. It is much more effective in extracting consistency and pattern from past crime statistics. 2) The LSTM version’s prediction accuracy are further improved by enhancing the built environment variables in urban areas. When compared to the original version that relied only on historical crime data, the improved prediction results are clear winners. We have enhanced the forecast accuracy of our designs compared to other designs. Using historical criminal offence data at a grid system size of 200 m200 m, Rummens et al. conducted an empirical research on the prediction of criminal offence hotspots. They used three designs: logistic regression, semantic network, and a blend of logistic regression and neural network. Two types of burglaries had the highest case hit rates in the biweekly projection: 31.97% and 32.95%, respectively (Liu et al.). In two weeks of testing at a 150m × 150m size, we used a random forest design to predict where the heat would be most intense. A case hit rate of 52.3% and an average grid hit price of 46.6% were normal for the version. When compared to the findings of earlier studies, the LSTM model used in this work achieved an improved instance struck price of 59.9% and an average grid hit rate of 57.6%. There are still areas that might be enhanced for the next research study. The prediction's temporal resolution is the first. The level of criminal activity changes with time, as Felson et al. revealed. A number of studies have shown that it helps to monitor the variety of dangers throughout the day.The forecast window was set at two weeks. Even with a one-day correction, it misses the impact of changes in crime that occur within a week. If the forecast home window is restricted to a day of the week or an hour within a day, the sparsity of information makes the prediction of criminal offence occasions problematic. The current state of this complex problem is devoid of any practical solution. The grid's spatial resolution is the second. A 150 150 metre grid is used in this research. Changing the grid sizes and their effect on prediction accuracy is something that needs further investigation. Third, several more research study sites should be investigated to determine the robustness and generalizability of this paper's searches. Nonetheless, given the study's sample size, the results have proven useful in a current hotspot criminal activity prevention experiment conducted by the regional police department.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "COMPARISON OF MACHINE LEARNING ALGORITHMS FOR PREDICTING CRIME HOTSPOTS", which provided good facilities and support to accomplish our work. Sincerely thank our Chairman, Director, Deans, Head Of the Department, Department of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. U. Thongsatapornwatana, ``A survey of data mining techniques for analyzing crime patterns,'' in Proc. 2nd Asian Conf. Defence Technol. (ACDT), Jan. 2016, pp. 123128.
2. J. M. Caplan, L. W. Kennedy, and J. Miller, ``Risk terrain modeling: Brokering criminological theory and GIS methods for crime forecasting,'' Justice Quart., vol. 28, no. 2, pp. 360381, Apr. 2011.
3. M. Cahill and G. Mulligan, ``Using geographically weighted regression to explore local crime patterns,'' Social Sci. Comput. Rev., vol. 25, no. 2, pp. 174193, May 2007.
4. A. Almehmadi, Z. Joudaki, and R. Jalali, ``Language usage on Twitter predicts crime rates,'' in Proc. 10th Int. Conf. Secur. Inf. Netw. (SIN), 2017, pp. 307310.
5. H. Berestycki and J.-P. Nadal, ``Self-organised critical hot spots of criminal activity,'' Eur. J. Appl. Math., vol. 21, nos. 45, pp. 371399, Oct. 2010.
6. K. C. Baumgartner, S. Ferrari, and C. G. Salfati, ``Bayesian network modeling of offender behavior for criminal proling,'' in Proc. 44th IEEE Conf. Decis. Control, Eur. Control Conf. (CDC-ECC), Dec. 2005, pp. 27022709.
7. W. Gorr and R. Harries, ``Introduction to crime forecasting,'' Int. J. Fore- casting, vol. 19, no. 4, pp. 551555, Oct. 2003.
8. W. H. Li, L.Wen, and Y. B. Chen, ``Application of improved GA-BP neural network model in property crime prediction,'' Geomatics Inf. Sci. Wuhan Univ., vol. 42, no. 8, pp. 11101116, 2017.
9. R. Haining, ``Mapping and analysing crime data: Lessons from research and practice,'' Int. J. Geogr. Inf. Sci., vol. 16, no. 5, pp. 203507, 2002.
10. S. Chainey, L. Tompson, and S. Uhlig, ``The utility of hotspot mapping for predicting spatial patterns of crime,'' Secur. J., vol. 21, nos. 12, pp. 428, Feb. 2008.

11. S. Chainey and J. Ratcliffe, ``GIS and crime mapping,'' Soc. Sci. Comput. Rev., vol. 25, no. 2, pp. 279282, 2005.

12. L. Lin,W. J. Liu, andW.W. Liao, ``Comparison of random forest algorithm and space-time kernel density mapping for crime hotspot prediction,'' Prog. Geogr., vol. 37, no. 6, pp. 761771, 2018.

13. C. L. X. Liu, S. H. Zhou, and C. Jiang, ``Spatial heterogeneity of microspatial factors' effects on street robberies: A case study of DP Peninsula,'' Geograph. Res., vol. 36, no. 12, pp. 24922504, 2017.

14. M. I. Jordan and T. M. Mitchell, ``Machine learning: Trends, perspectives, and prospects,'' Science, vol. 349, no. 6245, pp. 255260, Jul. 2015.

15. X. Zhao and J. Tang, ``Modeling temporal-spatial correlations for crime prediction,'' in Proc. Int. Conf. Inf. Knowl. Manag. Proc., vol. F1318, 2017, pp. 497506.

Predicting Employees Under Stress for Pre-Emptive Remediation using Machine Learning Algorithm

Sharma Astha, Kondapalli Chandrakala, Mohammad Irfan
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

DTV Dharmajee Rao
Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ dtvdrao@gmail.com

ABSTRACT

Working from home and coping with limited workers at company sites are only two examples of the innovative and varied ways that services and organisations have evolved to cope with the continuing COVID-19 epidemic. Employees have also adapted to new working environments and customisation's, which has led to mental tension and sleepiness for many as they adapt to the new normal and change their professional and personal lives. The new normal is here to stay for the near future. In this role, stress levels among employees have been predicted using data visualisation methods and machine learning algorithms. An information-based design may be developed to assist in predicting whether an employee is going to be under stress or not. In this case, we use the XGB classifier for our forecasts, and the results demonstrate that this method helps us get the most reliable version performance possible. Working hours, workload, age, and duty ambiguity all negatively affect employee performance, according to XGB classifier interpretation. In light of the above, the additional variables do not provide very useful information. Consequently, it is concluded that all perspectives would lead to a decrease in employee portrayal due to longer work hours, more function ambiguity, and the task itself.

KEYWORDS: *XGB, Covid 19, ML, WHO, Stress, High efficiency.*

INTRODUCTION

A global epidemic threatening the whole universe was announced by the World Health Organisation (WHO) as corona infection (COVID-19) on March 11, 2020 [3]. The coronavirus is responsible for the spread of the infectious illness COVID-19. The term "corona infections" refers to a large group of diseases that may cause anything from the common cold to serious consequences. That claims that 202 nations were infected with the virus as of March 31, 2020. This has caused a precipitous fall in development across a number of industries, including the securities industry. Employees are also affected since they start to worry when they can't cope with long periods of tension and uncertainty. The service industry is seeing encouraging expansion in the use of machine learning and AI. In [11], the employees' behaviour pattern is examined. In comparison, they aren't happy because of the long hours and high workload. The major objective of this research is to assess the effect of stress on the look of staff members. Furthermore, this impacts bodily ailments and a lack of commitment to one's job. But in the present, COVID-19 has put the world's people in an unprecedented situation. We want to gauge the extent to which workers experience worry and stress as a result of a current epidemic via this position. This is where AI formulae come into play, trying to figure out whether workers are experiencing stress or not.

LITERATURE SURVEY

Mental Wellness Forecasting Formulas Based on Data Mining Vijaya Pinjarkar, Aadesh Aachaliya, Hardik Jatta, and Vidit Laijawala wrote the book. Released in 2020.

In conclusion, a soldier's emotional, social, and mental health reflect their overall health. A person's thoughts, emotions, and reactions to situations are impacted by it. Stress, social anxiety, clinical depression, obsessive-compulsive disorder (OCD), substance abuse, workplace problems, and personality disorders are all factors that might affect mental health and mental illness.

Drawbacks: - The use of small datasets leads to less precise results.

-Methods for mining data to forecast psychological health.

- A mountain of data is required.

Title: AI-Based Methods for Forecasting Employee Stress in Active Roles The 2020 edition of A Dharun Magazine was written by U Srinivasulu Reddy and Aditya Vivek Thota.

In conclusion, anxiety disorders are common among working IT professionals in today's industry. Tensions among employees are on the rise as a result of shifting lifestyles and company cultures. Using machine learning techniques, this article will examine stress and anxiety patterns in working adults and identify the factors that significantly affect anxiety levels.

Drawbacks: - Fewer factors are used for tension forecasting.

Using the boosting formula in real time is not a good idea.

We utilise Ready Equipment to predict stress.

Using the Naïve Bayesian Category Formula, Analysing the Degree of Stress and Anxiety in Students Magazine year 2020 by Monisha S, Meera R, VijaySwaminath.R, and Dr. Arun Raj L

The combination of public opinion and overall academic efficiency has placed emotional strain on students. Reducing the stress elements that are often mentioned is crucial for helping adolescents succeed academically and engage in social duties. As a result, fewer individuals will suffer from certain health issues, such as headaches caused by migraines or difficulty wearing glasses, among others.

Drawbacks: - This theory only applies to college students and cannot be used to predict the stress levels of regular people.

- Additional processing time is required for these specific calculations.

Results are much less efficient.

EXISTING SYSTEM

Staff members, such as personnel managers, are understandably concerned about the prevalence of stress and anxiety in the workplace. Although there has been a lot of focus on anxiety management from both academics and practitioners for a long time, new perspectives and research are needed. This research offers evaluative results, drawing on the emerging topic of organisational efficiency, which has implications for reducing workplace stress. For a better understanding of adaptability in stress and anxiety symptoms, it may be essential to look at positive sources of efficiency, optimism, and strength, according to data from a large sample of working people across different organisations and sectors. Many studies have been published in nations that actively seek to foster and enhance innovation, both monetarily and socially, and there have been a great deal of investigations and experiments conducted in the last few years. Over the last several years, anxiety has become one of the most common "work disorders." Nearly three billion people worldwide report feeling stressed out on the job, which has a negative impact on their ability to do daily tasks.

Negative Aspects

There has been no Principal Component Analysis (PCA) implementation in the system. Unused XGB Classifier (XS Increase) in the system.

PROPOSED SYSTEM

Data cleansing is one of the most important and productive things you can do when working with data. We won't see much improved performance in design execution until we do that. Hence, it should be able to deal with null, empty, and non-existent values.

There are 3895 null values in this data set. Mean, typical, and flooring techniques are used to deal with numerical data, while mode is used to correct the categorical null values. In this case, it can also move down null values, although it can cause some data to disappear. This version execution is so preferred indefinitely.

Here is the variable that matters: the target. This shows that the data is imbalanced and that there are many ways to make it more balanced. It is also possible to use'smote' active specification techniques to resample data, oversimple data, or under sample data. Additionally, it may regress by making use of improved algorithms that provide the most trustworthy efficiency model, or by making use of bagging and improving strategies, formulae, and tools such as logistic regression, best examination measures, changing performance metrics, the ROC curve, for example.

Advantages

A vast quantity of datasets should be accurately examined and educated by the system.

Machine learning algorithms were created using the suggested method to assess and train the datasets.

Building Blocks

Supplier of Services

The Service Provider must enter the correct client login credentials in order to access this module. Upon successful login, he will be able to access many processes, including the following: Browse, Train & Test Data Sets, and Login. Sight-Trained and -Checked Accuracy Results, Accuracy in Bar Charts, Find the Staff Member Stress Prediction Kind Ratio, View the Staff Member Tension Forecast Type Get the Forecasted Data Sets, Results for Sight All Remote Users, Sight Worker Tension Forecast Kind Ratio

People may be seen and licensed

The admin may see a complete list of registered users in this section. The administrator is able to see user details such as name, email, and address, and they also have the power to authorise people.

Solo Traveller

All all, there are n clients in this section. Before conducting any operations, the customer should join up. An individual's data will be securely stored in the database after they sign up. He must provide the authorised customer name and password in order to log in once the registration is effective. Individuals may perform things like "REGISTER AND LOGIN," "PREDICTION OF WORKER STRESS TYPE," and "VIEW YOUR PROFILE" if the login process is successful.

OPERATION

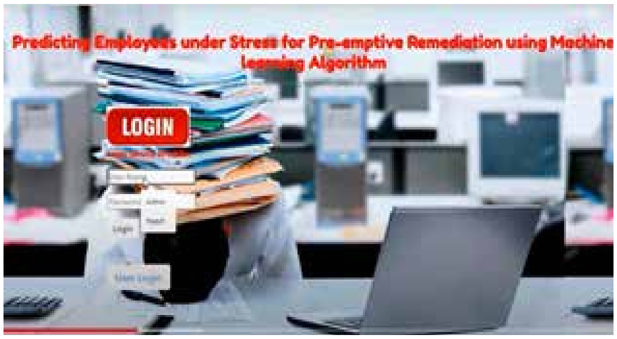
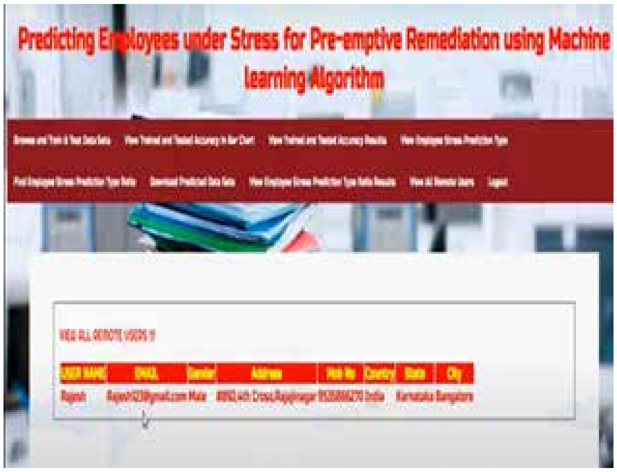
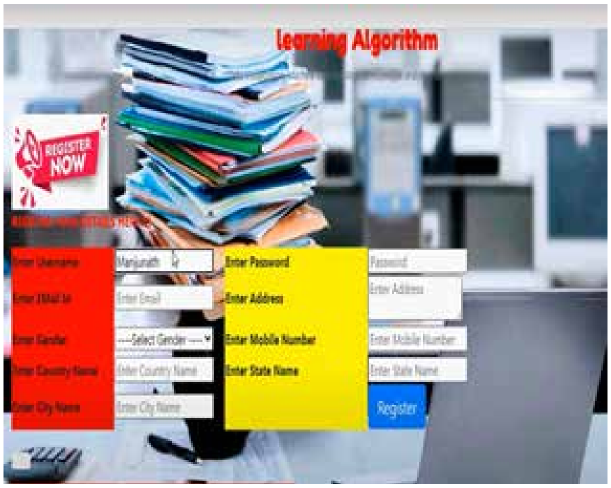

Predicting Employees under Stress for Pre-emptive Remediation using Machine learning Algorithm
LOGIN
Login
Reset

Fig. 1. Login page


Predicting Employees under Stress for Pre-emptive Remediation using Machine learning Algorithm
Browse and Train & Test Data Sets
View Trained and Tested Accuracy in Bar Chart
View Trained and Tested Accuracy Results
View Employee Stress Prediction Type
Find Employee Stress Prediction Type Ratio
Download Predicted Data Sets
View Employee Stress Prediction Type Ratio Results
View All Remote Users
Logout
VIEW ALL REMOTE USERS !!!

USER NAME	EMAIL	Gender	Address	Mob No	Country	State	City
Rajesh	Rajesh123@gmail.com	Male	#892,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore

Fig. 2. login details with mail id



Enter Username: Manjunath

Fig. 3. Login details verification

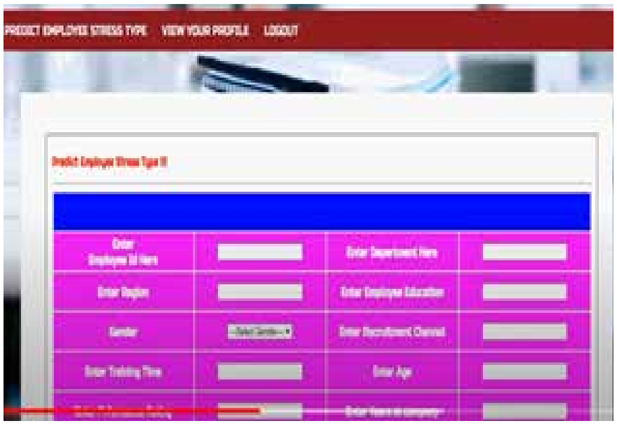


Fig. 4. Profile details

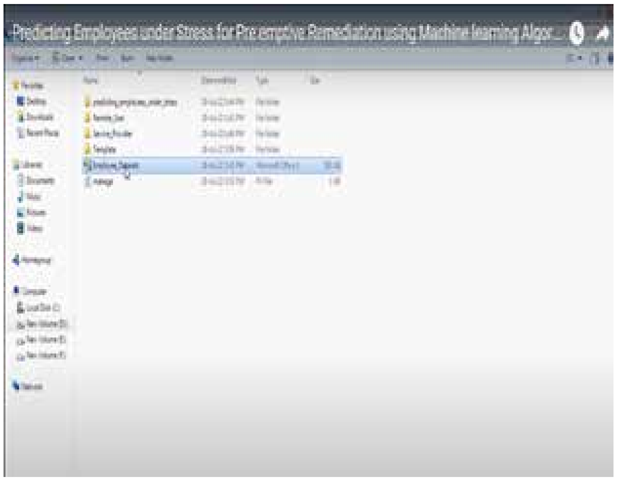

Predicting Employees under Stress for Pre emptive Remediation using Machine learning Algor...

Fig. 5. Data set name

CONCLUSION

In order to test our version and achieve much greater efficiency, we use the XGB classifier. Using a combination of a decision tree-based formula, the gradient boosting structure work method for evaluation, and a complexity matrix that tells us how many right values our version is expecting, this is one of the greatest optimisation techniques. XG Boost outperforms other slope enhancement algorithms by a factor of ten due to its superior anticipating power. It offers a range of regularisation techniques that improve overall performance while reducing the risk of overfitting. Thus, it is further known as the "regularised improving" approach. Examples of its values are true positive, true negative, false favourable, and wrong negative. Applied for determining how well the category model works.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "PREDICTING EMPLOYEES UNDER STRESS FOR PRE-EMPTIVE REMEDIATION USING MACHINE LEARNING ALGORITHM", which provided good facilities and support to accomplish our work. Sincerely thank our Chairman, Director, Deans, Head of the Department, Department of Computer Science and Engineering, Guide and Teaching and Non-Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. Bhattacharyya, R., & Basu, S. (2018). India Inc looks to deal with rising stress in employees. Retrieved from 'The Economic Times'

2. OSMI Mental Health in Tech Survey Dataset, 2017 from Kaggle

3. Van den Broeck, J., Cunningham, S. A., Eeckels, R., & Herbst, K. (2005). Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS medicine, 2(10), e267.

4. Shwetha, S, Sahil, A, Anant Kumar J, (2017) Predictive analysis using classification techniques in healthcare domain, International Journal of Linguistics & Computing Research, ISSN: 2456-8848, Vol. I, Issue. I, June-2017

5. Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for healthcare. International Journal of Bioscience and Biotechnology, 5(5), 241-266.

6. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., &Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.

7. Gender and Stress. (n.d.). Retrieved from APA press release 2010

8. Vidit Laijawala, Aadesh Aachaliya, Hardik Jatta, Vijaya Pinjarkar. Classification Algorithms based Mental Health Prediction using Data Mining (2020).

9. U Srinivasulu Reddy, Aditya Vivek Thota, A Dharun. Machine Learning Techniques for Stress Prediction in Working Employees (2020).

10. Monisha S, Meera R, Vijay Swaminath.R, Dr. Arun Raj L. Predictive Analysis of Student Stress Level Using Naive Bayesian Classification Algorithm (2020).

11. Fang Li. Research on the College Students Psychological Health Management based on Data Mining and Cloud Platform (2016).

12. Jigang Zheng, Jingmei Zhang. The Application of Association Rules Mining in the Analysis of Students Test Scores (2016).

13. Jenny K. Hyun, Brian C. Quinn, Temina Madon, Steve Lustig. Graduate Student Mental Health: Needs Assessment and Utilization of Counseling Services (2014).

14. Shivendra Jena, Harish Chandra Tiwari. Stress and mental health problems in 1st year medical students: a survey of two medical colleges in Kanpur, India (2014).

15. Honglei Zhu, Zhigang Xu. An Effective Algorithm for Mining Positive and Negative Association Rules (2008).

Detection of Cyberbullying on Social Media using Machine Learning

Mekala Jaya Laxmi, Mankala Vishnu Vardhan
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

Nuthanakanti Bhaskar
Associate Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ bhaskar4n@gmail.com

ABSTRACT

Young people today, referred to as "digital natives," have grown up in a period when new technologies have dominated society and communications have become almost instantaneous. As a result, building relationships with others and communities has never been easier. Teenagers are increasingly using social networking sites, which has exposed them to bullying. Adversarial remarks have a negative psychological impact on teenagers and demoralize them. In this work, we have developed supervised learning approaches for cyberbullying detection. Cyberbullying is when someone is harassed online using technology. Despite the fact that it has always been a problem, awareness of its effects on youth has grown recently. We can create rules to automatically identify cyberbullying content and identify language patterns that bullies and their victims employ through machine learning. There is a significant amount of bullying-related content on the website kagle.com, where we gathered the data for our study.

KEYWORDS: *Cyber bullying, Data encryption, Image.*

INTRODUCTION

In order to facilitate the production and dissemination of user-generated content, social networks assemble a suite of web-based applications that build upon the conceptual and technical underpinnings of Web 2.0. Many individuals take advantage of the abundance of information and the ease of contact offered by social media networks. However, there are potential downsides to social media, such as cyberbullying, which may impact people's lives, especially those of young people. Cyberbullying is defined as the hostile and purposeful use of electronic communication tools, such as the Internet, by one or more individuals or groups to harass, threaten, or otherwise harm another person or group. Unlike the more common kind of bullying that occurs in the classroom or in private conversations, cyberbullying on social media may happen whenever and anywhere it wants. They are free to hurt their classmates' feelings whenever they want to torment them online because they don't have to confront anybody. Because everyone, especially young people, is always online, victims of harassment are easy prey. According to [2], the victimisation rate resulting from cyberbullying may range from 10% to 40%. Nearly half(43%) of American young people have experienced cyberbullying at some point in their lives. Third In order to prevent cyberbullying from occurring, it is important to be able to spot bullying communications quickly and report them without delay. In order to amass a social network like Twitter, a strong system where one may freely communicate or assert anything, positive or negative. The act of ending one's own life is known as suicide. In a 2014 global research on guarding against suicide, the World Wellness Company found that among those aged 15 to 29, suicide ranks as the second greatest cause of death globally [3]. An estimated 800,000 people annually take their own lives. Every year, more individuals attempt suicide than actually do it. In the overall population, the most significant risk factor for self-destruction is a history of suicidal edition or behaviour. In the Philippines, the rate of self-destruction for males is 5.8 per 100,000, while for women it is 1.9 per 100,000, and for all sexes it is 3.8 per 100,000. The number of instances affected per 100,000 individuals is the basis for the rate. [2] The idea that the impoverished are more likely to suffer from severe depression and acts of self-destruction is false. The rising rates of major depression and suicide among college students from middle-class and affluent families are well-documented [4].

RELATED STUDY

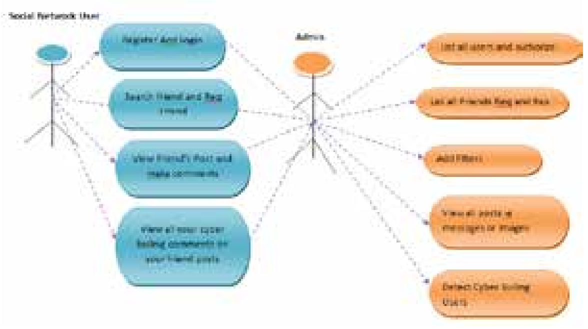
Websites belonging to social media networks are the main causes of cyberbullying. The ever-changing character of these platforms contributes to the expansion of violent conduct in cyberspace. Recognising the aggressor becomes more challenging due to the anonymous characteristic of user profiles. A social network's prominence stems from the fact that it is structured around interconnected networks. The problem arises, however, when unchecked complaints or bullying messages make it into the network. As an example of a popular social media site, consider Facebook and Twitter. Facebook users have around 150 billion connections, which gives an idea of how bullying content may propagate throughout the network in a short amount of time ([3]). Determining these abusive communications across such a large network by hand is difficult. It would be ideal if there was an automated system that could detect these types of problems and respond appropriately. Teens and women are the most common victims. [4] Work that has such a negative impact on people's mental and physical health increases the risk of depression and suicide attempts [2]. Cyberbullying can only be effectively controlled with tools that can recognise and monitor incidents automatically.

METHODOLOGY

Scientific understanding of massive data is aided by equipment or deep knowledge algorithms [1]. In the pre-big-data age, it was impossible to get your hands on the wealth of information on people and their cultures that is available now [2]. Social networks (SM) are a primary source for data pertaining to humans. Applying AI algorithms to SM data allows us to change past data in order to predict how different applications will work in the future. One potential use of machine learning algorithms is the detection and prevention of harmful human behaviours, such as cyberbullying [3]. Deep insight from raw data may be uncovered by large-scale information analysis, leading to surprising new knowledge [1]. By combining massive amounts of data with machine learning algorithms, big data analytics has improved a number of applications and made future prediction a real possibility. [4] In order to identify and control aggressive behaviour, it is necessary to combine methods and perspectives from a variety of disciplines and fields of study in order to conduct a thorough analysis of data pertaining to human behaviour and communication. With the availability of large-scale data comes a plethora of new possibilities for quantitative discovery, as well as novel computational tools, multidisciplinary approaches, and research study topics. However, using conventional wisdom (analytical procedures) here is a challenge for scalability and accuracy. Commonly, these methods rely on pre-arranged behavioural data and sparse human networks (typical social media networks). Several problems arise when these methods are used to large-scale online social networks (OSNs) in terms of level and range. One side of the coin is that the proliferation of OSNs makes it easier for bad actors to launch and spread their malicious campaigns. However, OSNs provide crucial information for learning about human behaviour and communication at large scales, and academics may use this information to develop efficient methods of identifying and reducing harmful behaviours. Offenders may find networks to do wrongdoing and instruments to carry them out on OSNs. Consequently, in order to identify and restrict malicious activities in facility systems, it is optimal to use methods that address both the content of the website and the network.

RESULTS EXPLANATION

Previous research on automated services for cyberbullying detection was insufficient. One of the main reasons there aren't enough training datasets is because of this. Supervised methods make use of publicly accessible datasets that are more focused on generic belief analysis. The frequency of bullying messages is low, even if they are published every day in comparison to the hundreds of hundreds of messages submitted per second. Due to the fact that random sampling will provide very few bullying messages, gathering enough training data is a significant challenge. We chose two separate datasets that were recently released and have anything to do with the social media platforms YouTube and FormSpring.me.


Social Network User
Register And Login
Search Friend and Req Friend
View Friend's Post and make comments
View all your cyber bulling comments on your friend posts
Admin
List all users and authorize
List all Friends Req and Res
Add Filters
View all posts ie messages or images
Detect Cyber bulling Users

Two components, the System Module and the User Component, make up this style layout. We can find the system and data source in the system module. The system has already been pre-trained from kagle.com, and we will be pulling the ready dataset from there. The data source is directly connected to the system. Clients may add friends and send each other friend requests in the user module of the app, but only once they log in. Here we're utilising the SVM formula to cover the comments with a star sign before sending them back to the user and making them public. If the quotes added by the customer do not contain any bullying words, they will be considered typical comments and sent to the public with no problems.

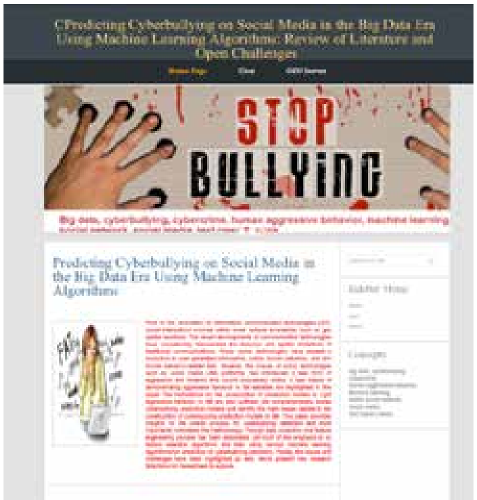

CPredicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges
STOP BULLYING
Big data, cyberbullying, cybercrime, human aggressive behavior, machine learning, social network, social media, text classif
Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms

Fig. 1. Home page

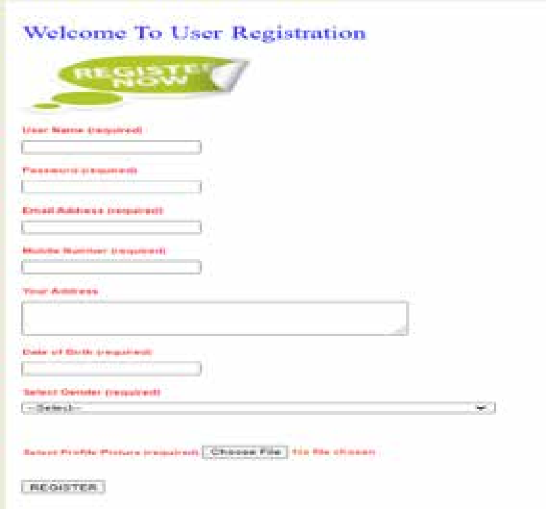


Fig. 2. Registation page

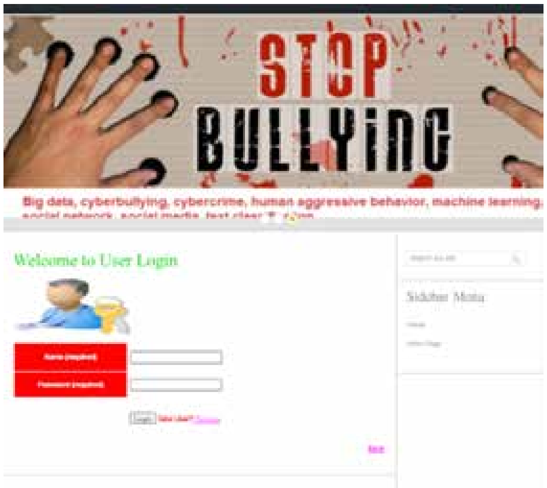

STOP BULLYING
Big data, cyberbullying, cybercrime, human aggressive behavior, machine learning, social network, social media, text classif
Welcome to User Login
Name (required)
Password (required)
Login

Fig. 3. User login page

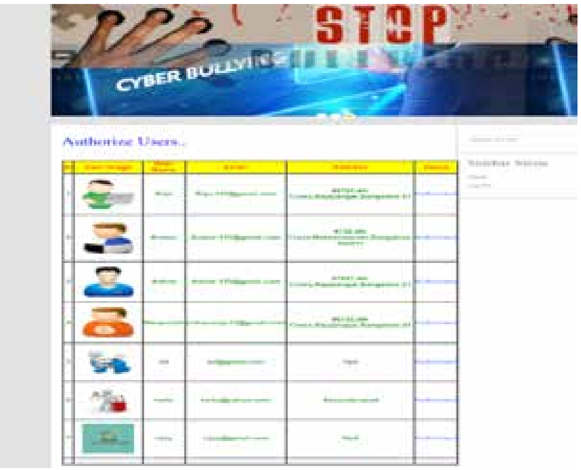


Fig. 4. Authorize details

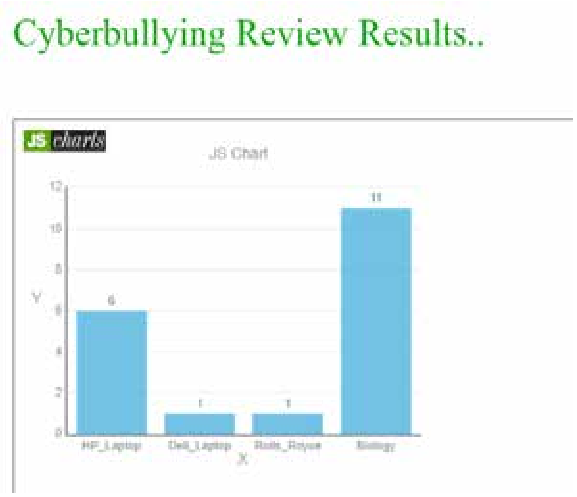


Fig.4.5. OUTPUT results

CONCLUSION

This study analysed previous literature to identify hostile behaviour on social media websites using AI techniques. We focused on four areas of cyberbullying message detection using AI methods: data collecting, function engineering, version building, and assessment of built cyberbullying discovery designs. There was also a synopsis of many discriminative qualities used to identify cyberbullying on online social networking sites. In addition, the best classifiers for monitoring equipment that can detect cyberbullying communications on online social networking sites were detected. The significance of evaluation metrics in quickly identifying the important criteria for comparing the various device learning formulae is one of the main contributions of the current work. We highlighted the most important aspects of cyberbullying detection using machine learning techniques, especially monitored learning, and we compiled and identified them. We have used f-measure, which provides the area under the curve characteristic for modelling behavioural behaviours, together with precision and accuracy recall, to achieve this purpose. In the end, we were able to identify and talk about the primary issues and unanswered research questions.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "DETECTION OF CYBERBULLYING ON SOCIAL MEDIA USING MACHINE LEARNING", which gave us excellent resources and assistance to complete our assignment. We sincerely appreciate the invaluable advice and direction that our chairman, director, deans, head of the department, computer science and engineering department, guide, and teaching and non-teaching faculty members have provided for every facet of our work.

REFERENCES

1. V. Subrahmanian and S. Kumar, ``Predicting human behavior: The next frontiers,'' Science, vol. 355, no. 6324, p. 489, 2017.
2. H. Lauw, J. C. Shafer, R. Agrawal, and A. Ntoulas, ``Homophily in the digital world: A LiveJournal case study,'' IEEE Internet Comput., vol. 14, no. 2, pp. 15_23, Mar./Apr. 2010.
3. M. A. Al-Garadi, K. D. Varathan, and S. D. Ravana, ``Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network,'' Comput. Hum. Behav., vol. 63, pp. 433_443, Oct. 2016.
4. L. Phillips, C. Dowling, K. Shaffer, N. Hodas, and S. Volkova, ``Using social media to predict the future: A systematic literature review,'' 2017, arXiv:1706.06134. [Online]. Available: https://arxiv.org/abs/1706.06134
5. H. Quan, J. Wu, and Y. Shi, ``Online social networks & social network services: A technical survey,'' in Pervasive Communication Handbook. Boca Raton, FL, USA: CRC Press, 2011, p. 4.
6. J. K. Peterson and J. Densley, ``Is social media a gang? Toward a selection, facilitation, or enhancement explanation of cyber violence,'' Aggression Violent Behav., 2016.
7. BBC. (2012). Huge Rise in Social Media. [Online]. Available: http://www.bbc.com/news/uk-20851797
8. P. A.Watters and N. Phair, ``Detecting illicit drugs on social media using automated social media intelligence analysis (ASMIA),'' in Cyberspace Safety and Security. Berlin, Germany: Springer, 2012, pp. 66_76.
9. M. Fire, R. Goldschmidt, and Y. Elovici, ``Online social networks: Threats and solutions,'' IEEE Commun. Surveys Tuts., vol. 16, no. 4, pp. 2019_2036, 4th Quart., 2014.
10. N. M. Shekokar and K. B. Kansara, ``Security against sybil attack in social network,'' in Proc. Int. Conf. Inf. Commun. Embedded Syst. (ICICES), 2016, pp. 1_5.

Train Time Delay Prediction for High-Speed Traindispatching Based on Spatio-Temporal Graphconvolutional Network

Reddi Malla Ramya Bhavani,
Mohammed Aftab
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

G Vinesh Shanker
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ gogikarvinesh@gmail.com

ABSTRACT

A better train hold-up prediction may lead to better train departures, which in turn helps the dispatcher have a better idea of the train's operating status and make more informed dispatch decisions. Many variables, including traveller circulation, error, harsh weather, and dispatching technique, might impact on the delay of a single train. Typically, dispatchers are only able to determine the separation time of a single train based on their own knowledge and experience. Predicting when trains will be delayed using current methods does not take into account the location and temporal dependency of the many trains and tracks. The goal of this study is to estimate the total variety of delays experienced by a particular station, which represents the cumulative impact of train delays over a certain time period, rather than the precise amount of time that a single train will be late. We offer a deep learning structure, the train spatio-temporal graph convolutional network (TSTGCN), to predict the overall effect of a train delay at a single station on train departures and emergency plans. The majority of the elements in the suggested design are either ongoing, occurring once a week, or everyday. Each component is made up of a spatiotemporal convolution and a spatiotemporal interest system for efficient spatiotemporal information acquisition. The weighted fusion of the three components yields the final prediction. In contrast to the most recent sophisticated criteria for predicting train delays, TSTGCN clearly performs better in studies conducted using data from the China Railway Traveller Ticket System.

KEYWORDS: *TSTGCN, CNN, ML, Train time delay, Weather environment.*

INTRODUCTION

Complex systems of interconnected infrastructure, including tracks and stations, and moving objects, most notably trains, make up railway networks [1]. Train status is influenced by a variety of factors, including internal interactions between facilities and products in the systems, external factors like weather, and other environmental factors [2]. When discussing the state of a railway network, the most typical moving thing to be discussed is trains. —delays are often employed as a measure of the variance between the actual and planned process strategies. As a result, operational train delays are crucial metrics for gauging the health of a rail network. As a result, railway controllers and drivers must prioritise network-oriented train hold-up development if they want to improve traffic management and rescheduling approaches[3]. Many cutting-edge machine-learning methods have found applications in railway systems as a result of the abundance of available data. One of the most popular areas is delay forecasting and breeding. Nevertheless, the majority of earlier studies on hold-up forecasting and propagation have been train-centric [4]. In other words, these studies mostly dealt with trains and attempted to predict the delays that each train would experience at stations downstream. However, Actually, dispatchers should be considerably more aware of the organised network states, such as the delays in the railway network. This study examines the various train types and stations within the systems, with a focus on network-based train hold-up design [5]. To find out how train network delays are growing, the suggested network-oriented approach takes into account all trains on the network at once and predicts when other trains will experience delays after a certain amount of time has passed. With the network-oriented approach, dispatchers will be able to see the big picture of the railway network's health and formulate a global strategy for improvement, rather than a piecemeal one.

There are energising excellent characteristics to the train hold-up. Not only can delayed trains impact on their own tasks, but they also affect the activities of other trains in the same area. Consequently, one of the fundamental aims of train dispatching is train postpone expectation [6]. When dealing with dispatching, the train defer assumption is very relevant. One of the main purposes of train delay expectation is to be ready for the impact of train activity blockage and delay spread. This will help with continual wager assessment, early criticism of sending off and constant adjustment of multi-mode transportation plans for emergencies [7]. It may help dispatchers figure out how trains are doing, the risk of gauge delays, and make sensible decisions about website traffic sending out. Consequently, the rapid rail route traffic order robotizing framework may benefit by concentrating on the assumption model of train delay. Numerous studies have been conducted to analyze and forecast train delays. For example, Milinkovi et al. proposed a fluffy Petri net model to simulate train activity and the traffic cycle in the railway structure; Tikhonov first examined the connection between the presence of tourist trains on postponement and several elements of the rail line structure before applying SVM to the postponement analysis; A train delay projection model based on Bayesian organization was created by Corman, Kecman, and Lessan; Yaghin The majority of these checks revolve around the possibility of a single train being delayed [8]. A number of factors may cause a train to be late, including but not limited to: course deficiencies, problems with the train and communication networks, terrible weather, heavy passenger loads, and on-site departures. If these factors are disregarded, the accuracy of the prediction will be diminished [9]. Furthermore, the spatial residential or commercial features of trains and programme are not considered. It is easy to see how delays affect the train job as a whole, and it's also easy to see how certain programs at certain cross stations may have a multiplicity of effects [10].

SURVEY OF RESEARCH

In the past, predictions regarding train delays have shown to be rather accurate. Most of it is usually composed of the following groups: Operates based on situation estimation and simulation data; (5) operates based on real performance data, taking into account the spatiotemporal features of the train procedure while ignoring external variables; (6) operates based on actual data, disregarding the spatiotemporal attributes of the train procedure; (7) operates based on actual data, taking into account external variables, but not the spatiotemporal attributes of the train operation. There are studies that do not use data from real train operations. Using the interpretative structure version, Wang et al. [5] analysed the train hold-up by looking at the four dimensions of people, tools, environment, and monitoring, and by adding 14 additional important influencing components. Following scenario analysis, Ma [6] determined the factors that affect train delay levels, calculated the matching weight using a specialist racking up method and an analytical power structure procedure, used genetics and data degeneration to solve the models of different scenarios, and used instance simulation to solve the train procedure change model, all with the goal of improving and adjusting the train delay design. While some research does exclude the train's spatiotemporal characteristics, it is based on real efficiency data. For example, in their study, Huang et al. [7] utilised random forest regression to predict train delays based on several independent variables, comprising the overall interval buffer time for each quit, the train's hold-up time at the first late terminal, the total hold-up time of the train across all terminals, and a 0-1 variable that indicates if the train is delayed during the Zhuzhou West Changsha South interval. Oneto et al. [8] proposed a fast knowing formula for shallow and deep severe knowing machines to predict train delays. The formula makes use of fully utilized memory scale data processing technology and actionable information from a vast amount of historical train operation data from the Italian train network. Even though the spatiotemporal aspects are frequently disregarded, certain research do account for external factors. For example, in their study, Oneto et al. [9] didn't use train procedure historical data but instead relied on fixed rules set by train facilities specialists based on classical univariate data. To improve the model even further, they used weather data from the national meteorological solution. However, train procedure data changes over time and space, so a model that follows expert-defined guidelines lacks flexibility and mobility and makes it hard to understand the legislation governing train procedure in the data.

EXISTING SYSTEM

To evaluate the efficacy of various train designs, Zhaoxia and Zhongying built a simulation system for train hold-ups that includes a graphical tracer and "controlled randomness"; Xin used the theory of discrete-occasion dynamic systems to develop a formula for state characteristics and a model for hold-up breeding; Kecman and Goverde approximated the train operation time using a time-event network diagram with dynamic weights; Carey developed a stochastic estimate method–based inspection system for train hold-up propagation simulations. Conventional mathematical model-driven methods rely on assumptions that make them ill-equipped to handle the complicated data produced by actual train operations and to aid in train dispatching during train hold-ups. Guo created a model using linear regression to predict hold-ups. When it comes to intelligent solutions, computational methods such as Bayesian networks and Unclear networks may greatly enhance the resolution of uncertainty modeling in train operation. The train procedure and price quote hold-up can be replicated with a fuzzy Petri web design by Milinkovi et al.; the complexity and dependency nature of train procedures can be solved with a Bayesian network-based design by Lessan; a stochastic model for anticipating the propagation of delay based on Bayesian networks, offered by Corman and Kecman, can be defined as the effect that the forecast horizon and inbound information about running trains carry the possibility of the hold-up. The experimental results suggest that a well-designed hybrid Bayesian network architecture may achieve low error and high accuracy based on domain expertise, knowledge judgments, and input from regional authorities.

PROPOSED SYSTEM

For the first time, to the best of our knowledge, we provide a collective cumulative impact estimate for train dispatching under the delay scenario. In order to predict the arrival delays at a single terminal within a certain time frame, The model known as TSTGCN is created. In the proposed version, the temporal and geographical dependencies are considered. A comprehensive map of China's high-speed rail network, including station locations and route length data, is currently under development. Additionally, we have constructed a 16-week actual operating data set for China's high-speed railway, comprising 727 endpoints, all routes linking these terminals, and 1,954,176 delay records for the period from October 8, 2019 to January 27, 2020. By comparing the recommend outright mistake (MAE), origin mean settled error (RMSE), and mean absolute percentage error (MAPE) to ANN, SVR, RF, and LSTM standards, the efficacy of the train hold-up prediction is assessed.

WORKING METHODOLOGY

Data about train operations is characteristic of spatiotemporal networks. Train operations on a genuine high-speed railway network are highly dependent on one another in space and time, and there is a close relationship between the two. Because of spatial dependancy, which is the direct interaction between neighbouring stations, delays at one terminal will have an effect on delays at another. What we mean by "temporal relevance" is that the pattern of a certain length of time spent waiting at a given station is identical to that of the previous few days or weeks. The fact that distinct terminals have different common influences in the spatial dimension is described by spatiotemporal correlation.

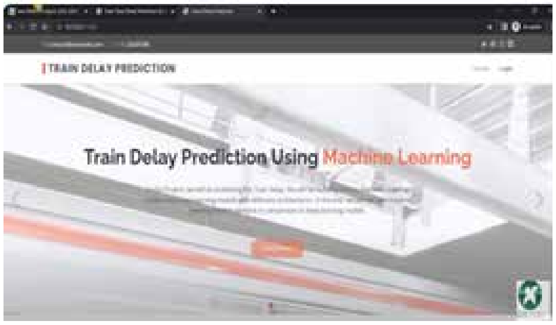


Fig.1. Home page

There is a strong dynamic correlation in the spatiotemporal dimension of the train operation data of high-speed railways. This is due to the fact that even a single terminal can have different effects on nearby terminals over time, and that different stations' historical observation data can affect the hold-up status of the terminal and its nearby stations at different future times. The study uses three methods to illustrate data: the most recent hourly collection, the instant collection of today and one week, and the historical collection. There are two dimensions to time and space that are shared by weather and major holidays. From a temporal measuring standpoint, for a single station, the weather will vary more in a week than in a day, and more in a day than in an hour. From a geographical measuring standpoint, different stations experience different climates in the same length of time. As an example, it's safe to assume that the weather in nearby stations will be similar, while the climate at terminals farther out will be different. Therefore, we in the team think that meteorological components have spatiotemporal properties. We at the organisation think that important holiday components have temporary features, especially when it comes to major vacations.

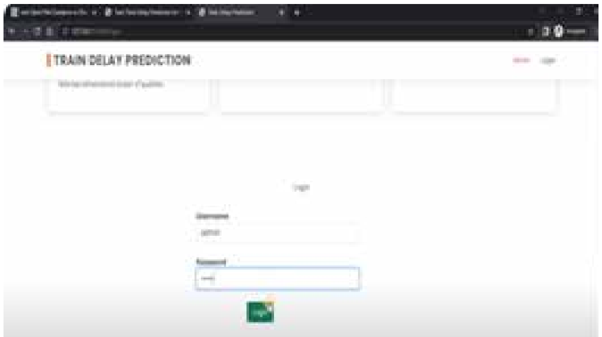


Fig.2. User details

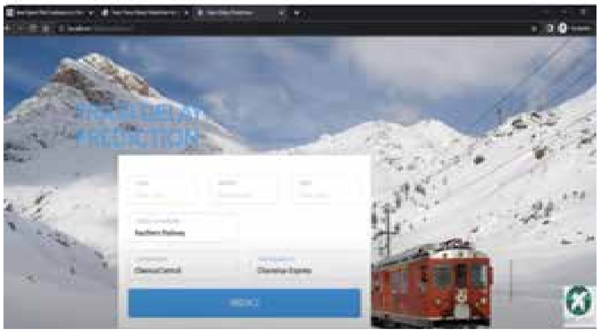


Fig.3. Delay detection

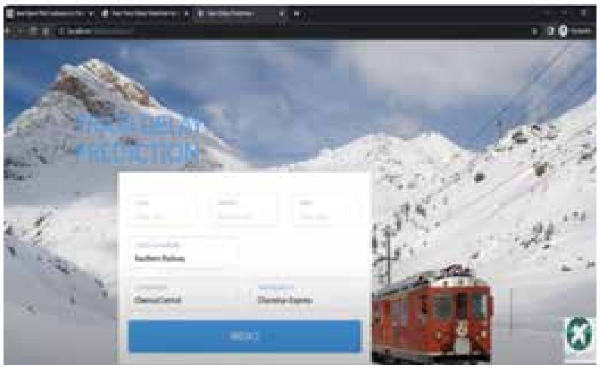


Fig.4. Output results.

CONCLUSION

This work provides a TSTGCN design based on focus system to consider vivid spatio-temporal correlation of high-speed train operating information and to forecast the cumulative impact of train arrival delays for railway dispatching. By collecting the spatio-temporal features of data from training methods, this methodology combines spatio-temporal attention with convolution to better predict outcomes. We evaluate the prediction effect of several models in the speculative phase using MAE, RMSE, and MAPE, and we compare our TSTGCN with ANN, SVR, RF, and LSTM versions. The experimental results demonstrate that TSTGCN is more effective than other approaches in predicting the consequences of delays in train dispatching.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "TRAIN TIME DELAY PREDICTION FOR HIGH-SPEED TRAINDISPATCHING BASED ON SPATIO-TEMPORAL GRAPHCONVOLUTIONAL NETWORK", which gave us excellent resources and assistance to complete our assignment. We sincerely appreciate the invaluable advice and direction that our chairman, director, deans, head of the department, computer science and engineering department, guide, and teaching and non-teaching faculty members have provided for every facet of our work.

REFERENCES

1. S. Jing, "Research on delay prediction of high speed railway train based on data analysis," Ph.D. dissertation, Southwest Jiaotong Univ., Chengdu, China, 2019.

2. R. G. Xu, “Multiple traffic jams in full velocity difference model with reaction-time delay,” Int. J. Simul. Model., vol. 14, no. 2, pp. 325–334, Jun. 2015.

3. S. Milinkovi´c, M. Markovi´c, S. Veskovi´c, M. Ivi´c, and N. Pavlovi´c, “A fuzzy Petri net model to estimate train delays,” Simul. Model. Pract. Theory, vol. 33, pp. 144–157, Apr. 2013.

4. N. Markovi´c, S. Milinkovi´c, K. S. Tikhonov, and P. Schonfeld, “Analyzing passenger train arrival delays with support vector regression,” Transp. Res. C, Emerg. Technol., vol. 56, pp. 251–262, Jul. 2015.

5. F. Corman and P. Kecman, “Stochastic prediction of train delays in realtime using Bayesian networks,” Transp. Res. C, Emerg. Technol., vol. 95, pp. 599–615, Oct. 2018.

6. J. Lessan, L. Fu, and C. Wen, “A hybrid Bayesian network model for predicting delays in train operations,” Comput. Ind. Eng., vol. 127, pp. 1214–1222, Jan. 2019.

7. M. Yaghini, M. M. Khoshraftar, and M. Seyedabadi, “Railway passenger train delay prediction via neural network model,” J. Adv. Transp., vol. 47, no. 3, pp. 355–368, Apr. 2013.

8. H. Ping, W. Chao, L. Zhongcan, Y. Yuxiang, and P. Qiyuan, “A neural network model for real-time prediction of high-speed railway delays,” China Saf. Sci. J., vol. 29, no. S1, pp. 24–30, 2019.

9. W. Chao, L. Zhongcan, H. Ping, T. Rui, M. Weiwei, and L. Li, “Progress and perspective of data-driven train delay propagation,” China Saf. Sci. J., vol. 29, no. S2, p. 1, 2019.

10. Y. Zhaoxia and D. Zhongying, “Simulation system of train delay propagation,” J. China Railway Soc., vol. 17, no. 2, pp. 17–24, 1995.

11. W. Xin, N. Lei, and L. Wen-Jun, “Study on robustness of high-speed train working diagram based on EMU utilization,” Railway Transp. Economy, vol. 36, no. 11, pp. 50–55, 2014.

12. P. Kecman and R. M. Goverde, “Online data-driven adaptive prediction of train event times,” IEEE Trans. Intell. Transp. Syst., vol. 16, no. 1, pp. 465–474, Feb. 2015.

13. M. Carey and S. Carville, “Testing schedule performance and reliability for train stations,” J. Oper. Res. Soc., vol. 51, no. 6, p. 666, Jun. 2000.

14. W. Chao, Y. Xiong, H. Ping, L. Zhongcan, and T. Youhua, “Review on conflict detection and resolution on railway train operation,” China Saf. Sci. J., vol. 28, no. S2, pp. 70–77, 2018.

15. J. Yuan and I. A. Hansen, “Optimizing capacity utilization of stations by estimating knock-on train delays,” Transp. Res. B, Methodol., vol. 41, no. 2, pp. 202–217, Feb. 2007.

Phishing URL Detection A Real-case Scenario through Login URLS

Bkataru Anusha, Akula Pravalika,
Marla Rishikesh
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

Bagam Laxmaiah
Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ blaxmanphd@gmail.com

ABSTRACT

Attacks known as “phishing” have emerged as a major threat to online security. In these attacks, malicious actors create bogus websites to deceive users into divulging sensitive information. The creation of phoney login web pages that mimic real ones in order to steal user credentials is a common tactic. This study examines the examination of login links as a means of detecting phishing links in real-world circumstances. With the evolution of phishing attacks, it is becoming more difficult for users to differentiate between legitimate and fraudulent websites. In order to trick their targets into divulging their login credentials, phishers often exploit login pages of popular services, including banking, email, and social networking site systems. To protect consumers from identity theft, financial loss, and unauthorised data access, it is vital to discover these phishing URLs. Finding phishing links in real-world scenarios is the focus of this research, which suggests a method that targets login URLs in particular. Web content analysis, evaluation of URL properties, and machine learning approaches are all part of the process. It is possible to determine if a URL is a phishing attempt based on characteristics including domain similarity, SSL certificate validity, examination of web page content, and URL structure.

KEYWORDS: *URL, SSL, Phishing attacks, SVM, Dataset.*

INTRODUCTION

Cybercriminals pose a significant and ongoing threat to users’ personal information, credentials, and financial assets via phishing attacks. These attacks trick victims into giving up sensitive information by using a variety of social engineering approaches. Making misleading login links that seem like real websites is a common and efficient tactic used by phishers to trick people into giving over their login credentials. The prevalence of phishing attacks has grown in recent years due to the proliferation of online solutions and the growing dependence on electronic systems. At its core, phishing is an attempt to trick users into unknowingly divulging critical information or financial data by making them provide their login credentials. Phishing attempts that exploit login URLs are particularly worrisome since they prey on people’s faith and reliance on legitimate online solutions. These attacks take advantage of people’s habits of inputting their credentials on known login pages, which makes them less likely to notice differences that may point to a fraudulent website. It is crucial to develop new algorithms that can effectively identify phishing URLs in real-world scenarios, as traditional phishing detection systems often fail to keep up with the evolving tactics utilised by attackers.

In this research, we zero in on the challenges posed by phishing attacks that use misleading login links. The goal of this research project is to develop and test a comprehensive strategy for detecting phishing URLs by analysing real-life scenarios. This method will enhance the ability to reliably identify these malicious URLs. This technique aims to provide a long-term defence against the increasing danger of phishing attacks by using machine learning algorithms, online content analysis, and URL characteristics. Subsequent sections will carefully investigate the methodologies used, training and evaluation of datasets, outcomes, and consequences of the results. This study adds to the larger effort to improve internet security and protect people from falling victim to phishing attacks with the use of misleading login URLs by advancing the status of phishing detection technology.

Statement of the problem

By capitalising on both psychological and technological weaknesses, phishing attacks continue to pose a constant and ever-changing hazard to cybersecurity. The development of deceptive login URLs, which mimic legitimate websites in an effort to fool users into divulging their login credentials, is an especially potent variation of these attacks. One of the biggest challenges in protecting consumers' private data and electronic assets is detecting and preventing these attacks.

When applied to real-world situations, the issue centres on the difficulty of accurately detecting phishing links, particularly those that use login web pages. Typical heuristics and patterns used by traditional anti-phishing technologies are easy to circumvent using new strategies that take advantage of consumers' comfort with tried-and-true login interfaces. The clever obfuscation tactics used by thieves and the quick growth of internet businesses both make this problem worse.

The goal of this research is to develop a solid method for detecting phishing links in real-world settings, with an emphasis on those that use fraudulent login web pages. The goal of this approach is to overcome the limitations of current detection techniques by using sophisticated methodologies including AI, online content evaluation, and URL characteristics assessment. Here are the main concerns that need fixing:

The methods used by attackers are always evolving in order to evade traditional methods of detection. New and sophisticated phishing URL variations that use deceptive login pages should be able to be detected using the suggested approach.

User Depend On Exploitation: Phishers take advantage of people's unspoken trust in well-known login forms. The challenge is in creating methods that can detect little discrepancies in actual login links that people may overlook.

Case in Point: Phishing assaults take place in dynamic and diverse settings. To be practically useful, the method should be able to analyse URLs in real-time across a variety of sectors and service types.

Ensuring Accurate Results: Finding the sweet spot between accurate detection and reducing false positives is crucial. In order to prevent legitimate Links from being mistakenly identified as phishing efforts, the approach must reach a high level of accuracy.

Flexibility: The discovery method must be able to adjust to new threats without requiring frequent manual updates, since opponents are always changing their tactics.

A multi-pronged approach using recent advances in web research, machine learning, and cybersecurity is required to address these challenges. This research aims to help enhance online safety for people, organisations, and electronic ecological communities by shedding light on the dangers of phishing attacks that use false login links.

LITERATURE SURVEY

The kernel-based method was proposed by Rashmi Karnik et al. as a model for categorization. We categorise phishing in this. With this technique, we can find phishing and malware sites with an estimated accuracy of 95%.

In order to prevent phishing attacks, Andrei Butnaru et al. compared Google Safe browsers with a monitored Machine Learning algorithm that was based on a unique combination of phishing attacks.

One of the most effective ways to identify these harmful works is artificial intelligence, according to Vahid Shahrivari et al. The Reason being, machine learning techniques can identify the vast majority of phishing attacks based on their shared characteristics. In this context, phishing website prediction is accomplished using a large number of machine learning-based classifiers. The capacity to generate variants for specific tasks, such as phishing detection, is the primary advantage of artificial intelligence. Machine learning models may be a great tool for dealing with phishing, which is a classification challenge.

In order to identify phishing websites, Ammara Zamir et al. presented a methodology that makes use of a piling model. You may examine Phishing aspects using attribute choice formulae such as details gain, gain ratio, Relief-F, and recursive function removal (RFE). We combine the finest and worst qualities to create two new features. Primary component assessment using a variety of device discovery techniques, such as neural network [NN] and random woodland [RF], makes use of nagging. Two heaping models, heaping1 (RF + NN + Bagging) and heaping2 (kNN + RF + Bagging), are used to improve category accuracy by merging the best possible score classifiers.

Deep neural networks (DNNs), convolutional neural networks (CNNs), long short-term memories (LSTMs), and gated persistent devices (GRUs) were proposed by Nguyet Quang Do, Ali Selamat, and colleagues in their study on phishing detection. Extensive experiments were conducted to examine the impact of parameter adjustment on the performance precision of the deep - knowledge models, in order to assess the behaviour of these designs. In which several designs' varying degrees of accuracy are shown by each variation.

A Machine Learning-based connection finding treatment was proposed by Ashit Kumar Dutta. To identify the phishing URL, an RNN is used. In particular, it is tested with 7,900 malicious sites and 5,800 legitimate ones. When compared to existing strategies, the final product of this technique demonstrates remarkable performance. In order to learn how phishing domain names appear and how to distinguish them from legitimate ones, Atharva Deshpande et al. proposed using a mix of AI algorithms and natural language processing approaches. In order to identify phishing websites using a cross-breed machine, Ms. Sophiaya Shikalgar et al. suggested a set of equipment learning classifiers. Determining method is an amalgam of several classifiers interacting to get an excellent prediction outcome. Classifiers differ in how they operate and the categories into which they fall. Takes use of a database of URLs that contains 29,05 Links in an unstructured manner.

A group of people led by Nureni Ayofe Azeez tried to solve this problem by focusing on two main areas. The primary concern is how to identify suspicious URLs on social media and how to protect users from clicking on fake or untrustworthy links. In order to train using characteristics gathered from the social media network and for further handling, it modifies six machine learning methods: AdaBoost, Slope Boost, random forest, Linear SVM, decision tree, and Naïve Bayes classifier. We looked at 532,403 items in total. Finally, 87,083 messages were deemed suitable for version education. With a 95% accuracy and a 97% precision, AdaBoost performs well compared to the others. Dear Ademola, Machine learning was suggested by Philip Abidoye and Boniface Kabaso to correctly classify the information in order to identify the phishing Link functionalities that attackers may use.

An innovative approach to phishing site detection using AI techniques was outlined by R. Kiruthiga and D. Akila, who also suggested a category version for identifying phishing assaults. Moreover, it offers a way to detect phishing email attacks with great accuracy by combining natural language processing with artificial intelligence.

EXISTING SYSTEM

Prevalent solutions identify phishing websites by textual characteristics or visual similarities; however, these can be easily circumvented. Instead, a method can be proposed to ascertain the real domain name of a checkout page by analysing common components of websites. Website trademarks comprise unique messages and images. According to the authors, the method achieves very high accuracy and very low error rates. In order to create a dynamic and extensible system for discovering new and existing phishing domains, Aaron Blum et. al. investigated the possibility of combining confidence-heavy classification with content-based phishing link discovery. The authors also claim that, unlike reactive blacklisting strategies, their system can detect emerging dangers and provide improved security against zero-hour threats. Engage in phishing attacks and identify zero-hour phishing attempts. On the other hand, the characteristics aren't always present in these attacks, and the false positive rate in detection is rather significant. You may use this element to include a new page into your main website. By using the "iframe" element, phishers may make their content invisible, meaning it will not be surrounded by any visible structure. Customers may provide sensitive information thinking the inserted website is an integral component of the main website since the border of the added page is invisible.

PROPOSED SYSTEM

We are more vulnerable to cyber crime since we have moved much of our banking, employment-related, and other daily responsibilities online. One of the most prevalent threats to internet users is phishing attacks that use URLs. This form of attack takes use of people's frailties rather than flaws in software. This malicious software targets both people and organisations, tricking them into clicking on seemingly secure links that steal sensitive information or install malware on our computer. Phishing URL detection, or the process of labelling a URL as legitimate or malicious, makes use of a variety of maker learning algorithms. Improving the accuracy and efficiency of current designs is a constant goal of research. As part of this project, we will evaluate several machine learning techniques, datasets, and URL characteristics that were used to train AI models to achieve this goal. There is discussion and analysis of the efficacy of different maker learning algorithms as well as methods to improve the accuracy of their activities. In order to help researchers keep up with the latest developments in the area and create better versions of phishing detection tools, we're working on a survey resource.

WORKING METHODOLOGY

Several techniques exist for detecting phishing URLs in actual-world contexts by means of analysing login URLs. These strategies seek for fake web sites that mimic actual login pages. Effectively identifying phishing URLs targeting login credentials involves the ranges indicated below, which give a excessive-degree summary of the technique:

Gathering and Preparing URLs: Collect a wide style of URLs from actual situations. This series of URLs should consist of both actual and malicious ones, and it must span many distinct styles of agencies and services. Get the domain, subdomain, path, arguments, and protocol information out of the URLs and get them ready to be used.

Extraction of Features: Get more statistics out of the URLs, such how lengthy they may be, if they consist of hyphens or digits, and information about the area's repute. In order to educate and evaluate gadgets getting to know models, these traits will be used.

The next step is to educate gadgets getting to know fashions using the dataset that has been produced. Common algorithms consist of neural networks, selection bushes, random forests, and help vector machines. You need to mark each URL in the dataset as both real or a phishing try.

Content Analysis: Retrieve the website's content material for any URLs that may be related to phishing. Examine the navigation, look for login forms, and notice any outside hyperlinks used on the internet site. The legitimacy of the internet site may be ascertained with the usage of this analysis.

Verifying an SSL Certificate: Verify that the URL's SSL certificate remains legitimate. Make positive that the SSL certificates is from a relied on certificate authority and that it suits the area. One sign of a phishing effort is a certificate this is either invalid or does not in shape.

Verifying Your Domain's Credibility: To verify whether or not the URL corresponds to any domains which can be regarded to be malicious, seek advice from a database that has this records. The presence of the URL in the database suggests that it might be an attempt at phishing.

Track how users engage with URLs, inclusive of how they click on, type, and circulate the mouse. Find times of uncommon interest on respectable login sites; this can be a sign of phishing.

Determining Thresholds: Establish appropriate thresholds for feature evaluation and model predictions. When those conditions are met, a URL can be marked as doubtlessly dangerous. To reduce the wide variety of false positives and negatives, the standards ought to be balanced.

Reporting and Scanning in Real Time: Put the detection techniques into movement in a real-time place. Perform function evaluation and practice the taught system, gaining knowledge of the model each time customers try to go to URLs. Make a file and provide warnings if a URL is determined to be malicious.

Feedback and User Education: Inform users approximately the dangers of phishing, how to spot malicious URLs, and what to do if they find one. In order to make the device greater accurate over time, we encourage users to provide comments on flagged URLs.

Updating databases and gadget getting to know fashions for acknowledged dangerous web sites on a ordinary basis is an crucial part of continuous monitoring and development. In order to combat new phishing techniques as they emerge, it is critical to live aware and modify the technique .

The methodology's efficacy is dependent on a trifecta of era improvements, user training, and steady investigation into new styles of phishing. Keeping a phishing URL detection device that works in actual-international situations with login URLs calls for constant refinement, input from customers, and cooperation amongst cybersecurity professionals.

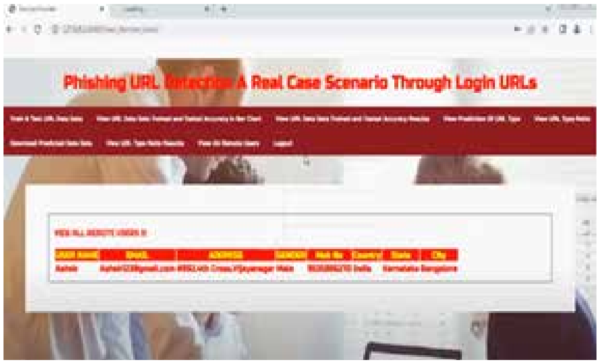
OPERATION

Fig.1. Web URL page

The continuous fight to guard digital environments from cyber assaults is based heavily on phishing URL detection. The upward thrust of phishing assaults is a major problem for clients, businesses, and information integrity due to the net's growing importance in our lives. Sensitive generation solutions, person schooling, and steady monitoring are necessary for the critical endeavour of detecting phishing URLs and shielding customers from fraudulent schemes.

Phishing URL Detection A Real Case Scenario Through Login URLs

Fig. 2. Admin login page


Phishing URL Detection A Real Case Scenario Through Login URLs

VIEW ALL REMOTE USERS !!!

USER NAME	EMAIL	ADDRESS	GENDER	Mob No	Country	State	City
Ashok	[illegible]	[illegible]	Male	[illegible]	India	Karnataka	Bangalore

Fig. 3. User details

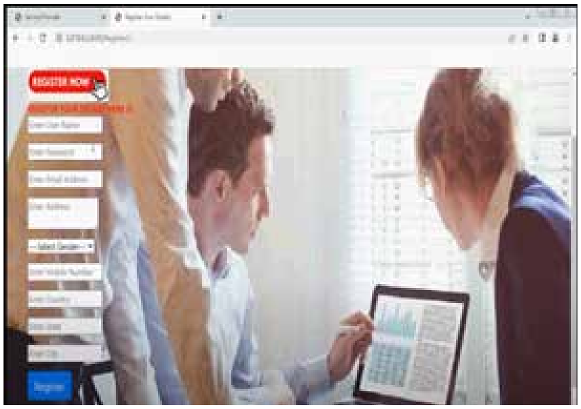
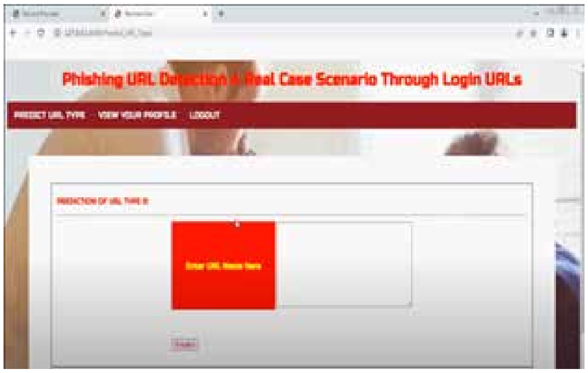


Fig. 4. Register details


Phishing URL Detection A Real Case Scenario Through Login URLs
PREDICT URL TYPE
VIEW YOUR PROFILE
LOGOUT
Enter URL Name here

PREDICTION OF URL TYPE !!!

Fig .5. URL upload page

Many special techniques have been devised to perceive phishing URLs, drawing on latest trends in device gaining knowledge of, online content analysis, person behaviour monitoring, and area reputation evaluation. The aim of these methods is to hit upon malicious URLs that masquerade as authentic web sites, taking advantage of users' familiarity with famous interfaces like login pages to trick them into giving up touchy statistics.

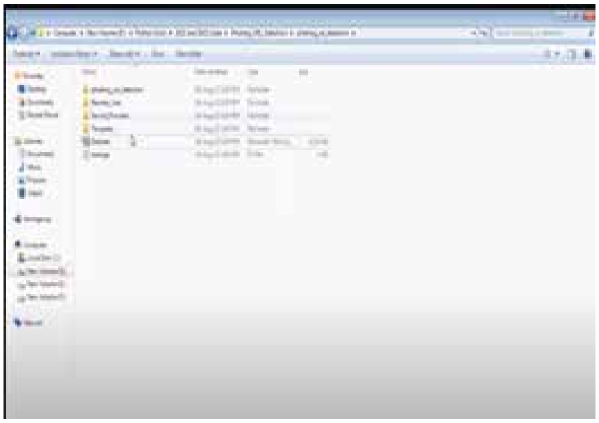
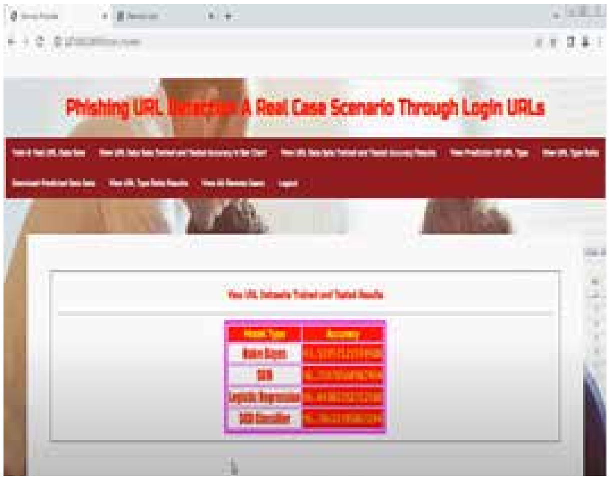


Fig. 6. Upload dataset details


Phishing URL Detection A Real Case Scenario Through Login URLs

View URL Datasets Trained and Tested Results

Model Type	Accuracy
Naive Bayes	[illegible]
SVM	[illegible]
Logistic Regression	[illegible]
SGD Classifier	[illegible]

Fig. 7. URL dataset with accuracy.

Problems nonetheless exist with phishing URL detection, even if it has made a few developments. In order to stay one step in advance of cyber criminals, defences want to be flexible and brief to respond. There must be continuous a look at and innovation in this location, since the net is continually changing, new attack routes are usually acting, and phishing tries are usually turned into smarter.

Combating phishing attempts relies heavily on teaching users. A safer online revel can be carried out by instructing users approximately phishing, the importance of cautiously examining URLs, and the risks of revealing personal information.

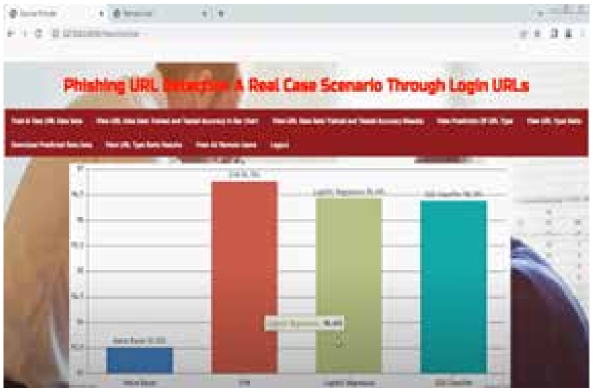


Fig. 8. Output Graphs

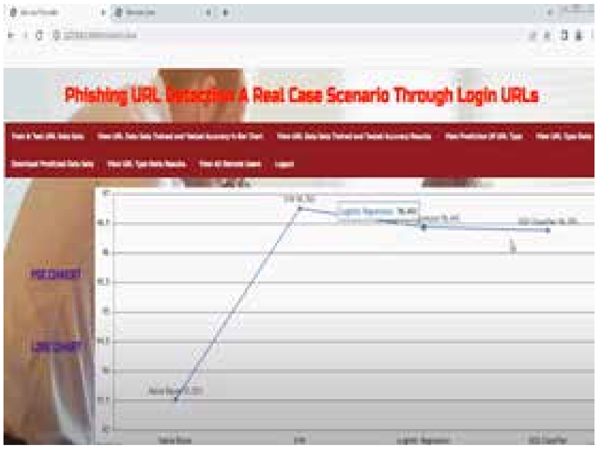
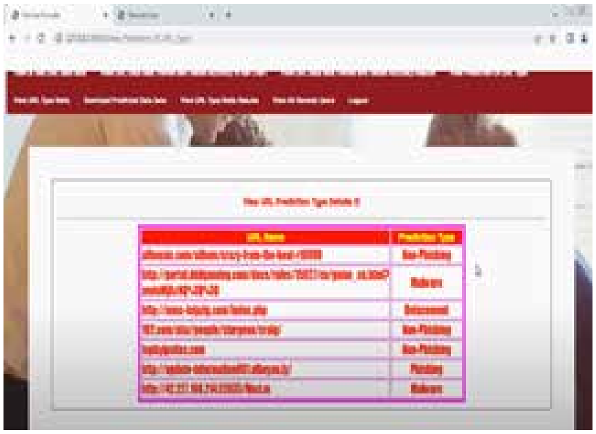


Fig. 9. Accuracy levels



View URL Prediction Type Details !!!

URL Name	Prediction Type
[illegible]	Non-Phishing
[illegible]	Malware
[illegible]	Defacement
[illegible]	Non-Phishing
[illegible]	Non-Phishing
[illegible]	Phishing
[illegible]	Malware

Fig.10. Phishing detected

CONCLUSION

Ultimately, academics, cybersecurity experts, innovative service providers, and consumers all play a part in the never-ending fight against phishing link detection. Keeping one step ahead of cyber criminals requires a multi-pronged approach that combines cutting-edge technical solutions with user education and understanding, a dedication to building a safe and secure digital ecological community, and the perseverance to face new challenges as they arise. Together, we can take action to reduce the dangers posed by phishing attempts and build a safer internet for everyone by being proactive and working together.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled “PHISHING URL DETECTION A REAL-CASE SCENARIO THROUGH LOGIN URLS”, which provided good facilities and support to accomplish our work. Sincerely thank our Chairman, Director, Deans, Head of the Department, Department of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. Anti-phishing Working Group (APWG) Phishing Activity Trends Report 4th quarter 2020, https://docs.apwg.org/reports/apwg trends report q4 2020.pdf
2. FBI Internet Crime Report 2020, https://www.ic3.gov/Media/PDF/AnnualReport/2020 IC3Report.pdf
3. Verizon 2020 Data Breach Investigation Report, https://enterprise.verizon.com/resources/reports/2020-databreachinvestigations- report.pdf
4. World Health Organization, Communicating for Health, Cyber Security, https://www.who.int/about/communications/cyber-security
5. Ye Cao, Weili Han, and Yueran Le, “Anti-phishing based on automated individual white-list,” Proceedings of the 4th ACM workshop on Digital identity management-DIM 08, pp. 51-60, 2008
6. M. Sharifi, and S. H. Siadati, “A phishing sites blacklist generator,” 2008 IEEE/ACS International Conference on Computer Systems and Applications, pp. 840-843, 2008
7. N. Abdelhamid, A. Ayesh, and F. Thabtah, “Phishing detection based associative classification data mining,” Expert Systems with Applications, vol. 41, no.13, pp. 5948-5959, 2014
8. L. Wenyin, G. Huang, L. Xiaoyue, Z. Min, and X. Deng, “Detection of phishing webpages based on visual similarity,” Special interest tracks and posters of the 14th international conference on World Wide WebWWW 05, pp. 1060-1061, 2005
9. C. L. Tan, K. L. Chiew et al., “Phishing website detection using url assisted brand name weighting system,” 2014 International Symposium on Intelligent Signal Processing and Communication Systems(ISPACS), IEEE, pp. 054-059, 2014
10. K. L. Chiew, E. H. Chang, W. K. Tiong et al., “Utilisation of website logo for phishing detection,” Computers & Security, vol. 54, pp. 16-26, 2015
11. K. M. kumar, K. Alekhya, “Detecting phishing websites using fuzzy logic,” International Journal of Advanced Research in Computer Engineering Technology(IJARCET), vol. 5, no. 10, 2016.

Prediction of Air Pollution Using Machine Learning

Kodarmur Sindhu, Kadira Poojitha,
Gundavaram Suryavamshi Vardhan Rao
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

K Praveen Kumar
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ praveenkumar.cse@cmrtc.ac.in

ABSTRACT

People are causing air pollution via their activities, automation, and urbanisation. Air pollutants include carbon monoxide, notorious oxide, chrysene, and others. Meteorological factors including air velocity, wind direction, relative humidity, and temperature control the concentration of air pollutants in the surrounding air. Machine learning (ML) is a considerable improvement over older methods like chance and statistics for predicting air quality, which are notoriously difficult to pull off. In order to forecast the relative humidity of the air, this method employs Linear Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), and the Random Forest Method (RF), and it uses Origin Mean Square Error to determine how accurate it is. The method takes into account a number of criteria, including CO, TIN oxide, nonmetallic hydrocarbons, Benzene, Titanium, NO, Tungsten, Indium oxide, temperature, and so on.

KEYWORDS: *SVM, DT, RF, LR, ML, CO, NO.*

INTRODUCTION

All of the vital things occurring in the environment are impacted by the pollution that humans cause through their daily activities, such as air pollution and noise pollution. The atmosphere becomes much hotter if the moisture level is increasing rapidly. Transportation and industry account for 75% of all gases in the environment, along with CO, SO2, and other fragments. This is a major contributor to the ever-increasing pollution levels [1]. There has been a terrifying increase in the pace at which the expanding scene, autos, and inventions are damaging the air. Hence, we have really gathered some quality data, such as the number of automobiles and pollutant connections, in order to make a prediction about the pollution level in a certain location of Delhi. Whatever is immediately around us is everything that is considered part of the environment. Human activity and natural disasters are contaminating the environment, with air pollution being one of the most severe forms. Weather conditions, including temperature, humidity, wind speed and direction, and loved ones' moisture levels, determine the concentration of air contaminants in the surrounding air. Since perspiration won't evaporate into the air if there's a lot of moisture, it makes us feel hotter. The expansion of urban areas is a leading cause of air pollution because the proliferation of vehicles on the road releases even more harmful gases into the atmosphere. Another important contributor to this problem is the rise of automated machinery. Nitrogen oxide (NO), carbon monoxide (CO), particulate matter (PM), sulphur dioxide (SO2), and other similar substances are major carcinogens. Inadequate oxidation of propellants (gas, petroleum, etc.) results in the production of carbon monoxide gas. The combustion of thermal fuel produces nitrogen oxides; carbon monoxide causes headaches and vomiting; benzene is a byproduct of smoking and may aggravate respiratory problems; and nitrogen oxides can make you feel woozy and sick to your stomach. Even more harmful to human health are particles with a diameter of 2.5 micrometres or less.It is imperative that steps be taken to reduce air pollution in the area. One way to gauge air quality is by using the Air Top Quality Index (AQI). In the past, experts relied on traditional techniques like probability and statistics to predict air quality, but such approaches were cumbersome.

The development of contemporary technology has made it quite easy to collect data about air pollution using sensors. Energy assessment is required for the examination of raw data in order to find the pollutants. Algorithms for learning equipment, recursive neural networks, deep learning, and convolutional neural networks all ensure accurate future AQI predictions, allowing for timely and appropriate action.Machine learning, a subfield of AI, makes use of three distinct learning algorithms: monitored learning, unsupervised learning, and reinforcement learning. We have used the supervised discovering approach in the recommended task. Linear Regression, Nearest Next-door Neighbour, Support Vector Machine, bit SVM, Naive Bayes, and Random Forest are just a few of the many supervised learning formulae. Our method uses Random Forest to accurately predict air pollution since it outperforms all other models.

LITERATURE SURVEY

For the purpose of predicting the intensity of air pollution, Ishan et al. [1] detailed the advantages of the Bidirectional Lengthy-Brief Memory [BiLSTM] technique. The suggested method improved forecasting by identifying the most significant, immediate, and long-term effects of PM2.5 intensity levels. The suggested method makes predictions at6,12, and 24 hours intervals. After 12 hours, the results are consistent, but after 6 and 24 hours, they are inconsistent. For the purpose of air quality prediction, Chao Zhang et al. [2] suggested using an online service. The consumer was able to transmit photographs of air pollution thanks to the solution they supplied for their cell phone. There are two parts to the method that is suggested. a) data about the location of general practitioners so that they may get air quality assessments from local air quality terminals. b) In order to forecast the air quality, they have used a convolutional semantic network and a lexicon to analyse the user-submitted photos. When compared to other algorithms like PAPLE, DL, and PCALL, the suggested technique has a much lower mistake price. However, there is a downside to this method when it comes to finding security, which makes the findings less precise. To determine the air quality, Ruijun Yang et al. [3] built a DAG using data collected from the town known as Shanghai and used the Prejudice network. Each version's training and testing datasets are partitioned. The lack of consideration for geographical and social environment qualities is a drawback of this technique, since these variables might affect the outcomes. In order to gather air quality data, Temesegan Walelign Ayele et.al. [4] suggested an Internet of Things (IoT) based method. They have calculated the expected air quality using the Long Short-Term Memory [LSTM] approach. By drastically cutting down on training time, the suggested approach achieved much higher accuracy. On the other hand, other approaches, like the Random woodland methodology, may improve the accuracy even more. In order to predict the presence of two important air pollutants, Nadjet Djebbri et.al.[5] suggested a nonlinear, fabrication-based regression strategy.

Organisation for Parts

The Service Provider must provide a valid username and password in order to access this module. Following successful login, he will be able to access many processes, including: Train Data Sets and Sight Childbirth Prediction, Sight Train and Exam Results, Sight Air Quality/Pollution Facts, Determine the Ratio of Air Quality to Pollution Forecasts Using Data Sets, Find Out What Happens When You Predict Air Pollution and Quality, Access All Remote Users, Install Trained Data Establishments, and Accredit Individuals.

The administrator may see a list of all registered clients in this section. Here the admin may see the user's information (name, email, and address) and approve or disapprove their access.

Solo Traveller

Here you may find n different types of customers. Before undergoing any operations, individuals are need to sign up. A person's details will be stored in the database as soon as they register. He will be prompted to provide his licenced customer name and password upon successful registration. After the login process is complete, the user will be able to do things like see their profile, predict the kind of air pollution, and register and log in.

Submitted form

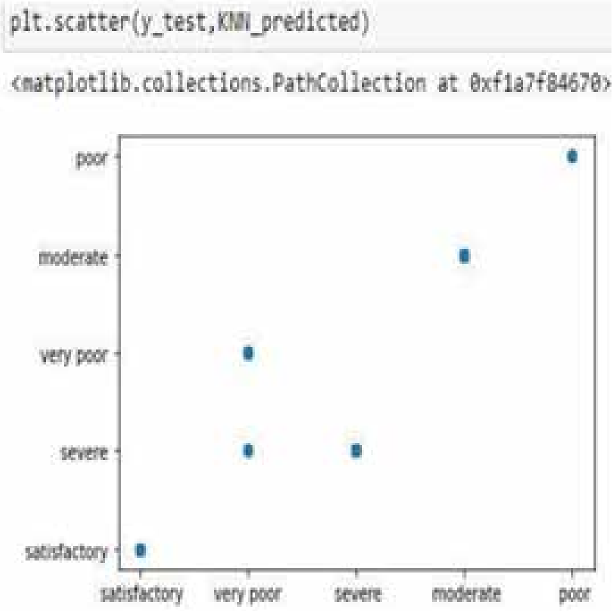
Data on air pollutants is collected from the sensors, processed, and then stored in a database. Quality selection and normalisation are two of the many characteristics that have been applied to this dataset throughout its pre-processing. When the dataset is ready, it is split into two parts: one for training and one for testing. The next step is to apply an AI algorithm using the training dataset. In order to assess the findings, the obtained data is compared to the testing dataset.

Machine Learning Edition In order to foretell the air pollution, a Machine Learning method is used. Multi-Level Perception (ML) is a branch of AI that enables software programme to accurately predict outcomes without being specifically programmed to do so. A.I. Algorithms use existing historical data as input and utilise it to predict future results. With Machine Learning, a person may feed a computer a mountain of data, and the machine will do all the heavy lifting in terms of analysis and drawing conclusions. The KNN algorithm is an AI tool for predicting air pollution levels. You may think of K-Nearest Neighbours (KNN) as an example of a monitored AI system. KNN does a lot of hard category work with a very basic look. One name for KNN is "lazy learning algorithm" since it skips the training process. Instead, it trains on the whole dataset and finds a new data factor at the same time. The name "non-parametric knowing approach" comes from the fact that it does not assume anything. Involvement in Support Vector Machines:

Identify the range of values that fall between the test data and each sample of the training data.

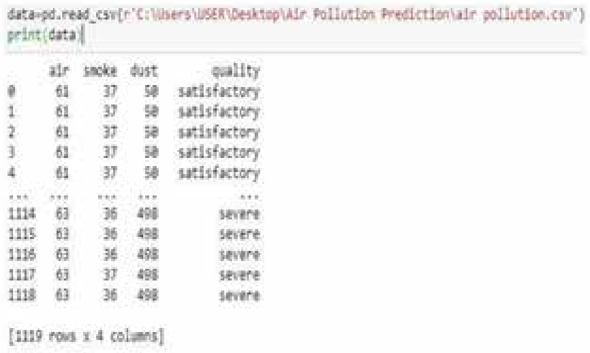
- The range may be determined using the formulas for the distances in either the Euclidean, Minkowski, or Manhattan dimensions.
- Arrange the predicted ranges from highest to lowest.
- Choose the classes.
- Determine the accuracy of the model and rebuild it if necessary. - The outcome will be determined by the class with the highest votes.

The capacity to gain useful example statistics, like approaches, variances, and connections with different other specifications, is an additional function to attempt to stationers a time series. If a series is stationary, then these statistics can only be used to predict future practices. For instance, if the series is consistently increasing with time, the sample mean and difference will undervalue the mean and variance in good periods, but they will climb with example dimension. More importantly, if the collection's mean, variation, and correlations with other variables are not explicitly stated, then the collection's variance and mean are also not verbalised. When projecting regression designs that were fitted to non stationary data, it is important to use care.



```
plt.scatter(y_test,KNN_predicted)
```

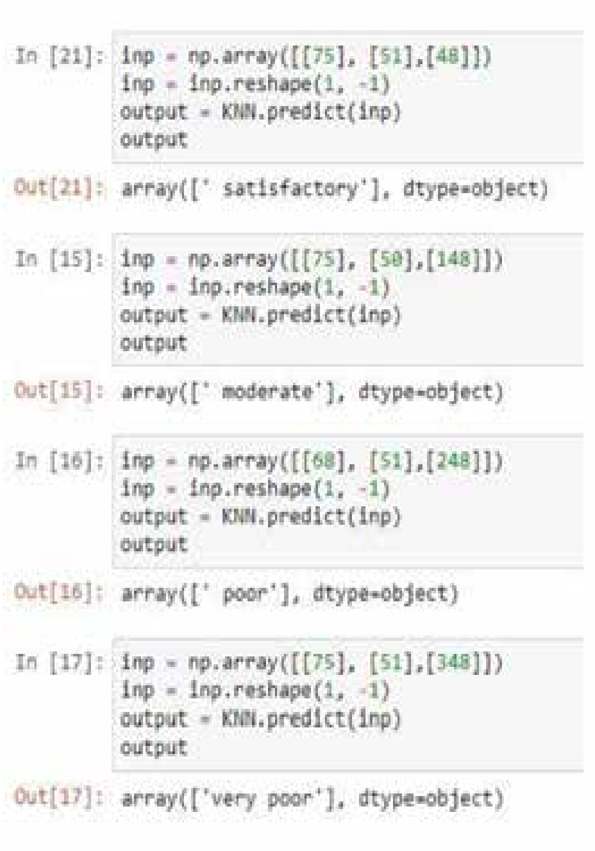
```
<matplotlib.collections.PathCollection at 0xf1a7f84670>
```



```
data=pd.read_csv(r'C:\Users\USER\Desktop\Air Pollution Prediction\air pollution.csv')
print(data)
```

```
      air  smoke  dust       quality
0      61     37    50  satisfactory
1      61     37    50  satisfactory
2      61     37    50  satisfactory
3      61     37    50  satisfactory
4      61     37    50  satisfactory
...   ...    ...   ...           ...
1114   63     36   498        severe
1115   63     36   498        severe
1116   63     36   498        severe
1117   63     37   498        severe
1118   63     36   498        severe

[1119 rows x 4 columns]
```

Fig.1. Dataset details



```
In [21]: inp = np.array([[75], [51],[48]])
         inp = inp.reshape(1, -1)
         output = KNN.predict(inp)
         output

Out[21]: array([' satisfactory'], dtype=object)

In [15]: inp = np.array([[75], [50],[148]])
         inp = inp.reshape(1, -1)
         output = KNN.predict(inp)
         output

Out[15]: array([' moderate'], dtype=object)

In [16]: inp = np.array([[68], [51],[248]])
         inp = inp.reshape(1, -1)
         output = KNN.predict(inp)
         output

Out[16]: array([' poor'], dtype=object)

In [17]: inp = np.array([[75], [51],[348]])
         inp = inp.reshape(1, -1)
         output = KNN.predict(inp)
         output

Out[17]: array(['very poor'], dtype=object)
```

Fig.2. Output results.

CONCLUSION

Components such as gas and particle problem determine the air quality. The air quality is negatively impacted by these contaminants, which, when inhaled repeatedly, may lead to significant health complications. The presence of these toxics may be determined and air quality monitored with the use of air quality monitoring devices, allowing for the practical improvement of air quality. As a result, production spikes and air pollution-related health issues are reduced. It has been shown that the AI-built forecast versions are much more consistent and dependable. Innovating current technologies and sensors have made data collection both fundamental and particular. In order to generate accurate and trustworthy predictions from such massive amounts of environmental data, only artificial intelligence (ML) algorithms are capable of handling the thorough review required. The KNN method, which is more suited to forecasting jobs, is used to predict air pollution. An impressive 99.1071% accuracy rate in predicting airborne pollution has been provided by the KNN machine learning system.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled “PREDICTION OF AIR POLLUTION USING MACHINE LEARNING”, which provided good facilities and support to accomplish our work. Sincerely thank our Chairman, Director, Deans, Head Of the Department, Department of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. Ni, X.Y.; Huang, H.; Du, W.P. “Relevance analysis and short-term prediction of PM 2.5 concentrations in Beijing based on multi-source data.” Atmos. Environ. 2017, 150, 146-161.

2. G. Corani and M. Scanagatta, “Air pollution prediction via multi-label classification,” Environ. Model. Softw., vol. 80, pp. 259-264,2016.

3. Mrs. A. GnanaSoundariMtech, (Phd) ,Mrs. J. GnanaJeslin M.E, (Phd), Akshaya A.C. “Indian Air Quality Prediction And Analysis Using Machine Learning”. International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 11, 2019 (Special Issue).

4. Suhasini V. Kottur , Dr. S. S. Mantha. “An Integrated Model Using Artificial Neural Network

5. RuchiRaturi, Dr. J.R. Prasad .“Recognition Of Future Air Quality Index Using Artificial Neural Network”. International Research Journal ofEngineering and Technology (IRJET) .e-ISSN: 2395-0056 p-ISSN: 2395-0072 Volume: 05 Issue: 03 Mar-2018

6. Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu .” Detection and Prediction of Air Pollution using Machine Learning Models”. International Journal o f Engineering Trends and Technology (IJETT) - volume 59 Issue 4 - May 2018

7. Gaganjot Kaur Kang, Jerry ZeyuGao, Sen Chiao, Shengqiang Lu, and Gang Xie.” Air Quality Prediction: Big Data and Machine Learning Approaches”. International Journal o f Environmental Science and Development, Vol. 9, No. 1, January 2018

8. PING-WEI SOH, JIA-WEI CHANG, AND JEN-WEI HUANG," Adaptive Deep Learning-Based Air Quality Prediction Model Using the Most Relevant Spatial-Temporal Relations," IEEE ACCESSJuly 30, 2018.Digital Object Identifier10.1109/ACCESS.2018.2849820.

9. Gaganjot Kaur Kang, Jerry Zeyu Gao, Sen Chiao, Shengqiang Lu, and Gang Xie,"Air Quality Prediction: Big Data and Machine Learning Approaches," International Journal of Environmental Science and Development, Vol. 9, No. 1, January2018.

10. Haripriya Ayyalasomayajula, Edgar Gabriel, Peggy Lindner and Daniel Price, "Air Quality Simulations using Big Data Programming Models," IEEE Second International Conference on Big Data Computing Service and Applications,2016.

Identification of Fake Profile in Social Network using NLP and Machine Learning

Polam Srija, Gopu Udaypal Reddy,
Madishetti Manusree
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

Srinu Vandanapu
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ svandhanapu@gmail.com

ABSTRACT

Millions of individuals all around the world use some kind of social networking site. Everyday life is significantly affected by user behaviour on social media sites like Facebook and Twitter, and this behaviour is often unfavourable. Spammers have circulated a great deal of harmful and irrelevant content via popular social networking platforms. For example, because to its meteoric rise to prominence, Twitter is now one of the most spammed platforms in history. False accounts market businesses or websites by sending out unwanted tweets to users, which wastes resources and hurts real consumers. Another factor contributing to the proliferation of dangerous things is the increased capability of disseminating incorrect information to people via the use of fraudulent identifications. Discovering Twitter spammers and fraudulent people has lately be a accepted study field (OSNs) in today's online social media, deceitful online material, centred on Spam URLs, hot subjects replete with junk, and phoney clients. Users, site content, charts, structures, and temporal factors are among the many parameters used to compare and evaluate the offered techniques. The offered paper is a great resource for researchers seeking the latest and most significant growth in Twitter spam detection on a single platform.

KEYWORDS: *URL, Social media, Twitter, Fake account, OSN, Spam.*

INTRODUCTION

A lot of studies have used Twitter, which is one of the mainly well-liked social media platforms. Nearly everyone uses Twitter these days. False Twitter user IDs were discovered in this survey since we too have fake consumers on that platform [1]. In this study, we will definitely find fake consumers by using deceptive online contented, URL-based spam detection, spam in favourite themes, and fake user identification. subsequently identify the deceitful client [2]. By posting often and on irrelevant subjects, the fake user would waste other people's time. In recent years, the popularity of social media websites has skyrocketed, including Facebook, Twitter, Instagram, MySpace, and Connected In. When it comes to social media platforms, Twitter is one of the most well-known and well-known [3]. Users of social networking sites may now post and communicate with ease thanks to Twitter. The Twitter network uses the word "tweet" to describe communications that are less than 280 characters long. The vast majority of the time, individuals post their opinions about various things, feelings, and other people's views on social networking sites [4]. A customer's largest mechanism for submitting comments and evaluations on things they have actually purchased might be these social networking platforms. Twitter ads now have a greater rate of spam information access than email ads since only 0.13 percent of users click on links [5]. Cybercriminals and social robots alike often target Twitter and other online social media networks because of the large number of people who use them to share critical information. Social bots are another name for spam crawlers that operate on social media networks.

Research has been done a lot in the domain of Twitter spam identification. A few surveys on fraudulent consumer identification from Twitter have also been carried out in order to incorporate the state-of-the-art at the moment. In their publication, Tingmin et al. [4] provide a comprehensive overview of current approaches and methodologies for Twitter spam detection. The methods used now are contrasted with the results of the survey that was mentioned before. Conversely, researchers examined the several actions conducted by spammers on the Twitter social media platform in [5]. Furthermore, the paper recognises that spammers are visible on Twitter and offers a literature review that supports this claim. No matter how many studies are conducted, there will always be a gap in the literature. Therefore, in an effort to close the area, we assess one of the most current increases in the detection of spammers and fake user identification on Twitter. Along with a comprehensive review of recent advances in the subject, this paper also offers a taxonomy of strategies for detecting spam on Twitter.

A choice in social media according to Wikipedia "concentrates on the advancement and confirmation of online social media networks for communities of individuals who share interests and conditioning or who have an interest in discovering the interests and conditions of others, and which asks for making use of software programme [6]." OCLC reports provide an explanation for the following social media platforms. The majority of the users of social networking sites like MySpace, Face publication, and Mixi are addicts who use these sites to trade products and services. Social media platforms and networks provide a number of advantages to an organization's participants. to aid in locating Learning communities and those involved in proficiency support may benefit from the increased social connections made possible by social media [7]. Even in a more informal setting, they have the potential to increase reading comprehension. assistance to members of a group Any employee, not just those who deal directly with students, may use a company's social media accounts. Advancements in technology may benefit from social media networks. Engaging in conversation with colleagues Important business information and thoughts on institutional solutions may be shared via social media (however this might lead to moral ventures) [8]. restriction of access to jobs and details Addicts may benefit from the ease of use of many social media platforms as it speeds up their access to additional tools and processes. The Face Publishing System exemplifies the way a social networking solution may serve as a platform for other devices. straight forward user interface One possible perk of social networks is the shared interface that transcends professional and interpersonal boundaries [9]. Thus, the treatments may be used with less training and assistance in a professional context since they are routinely utilised in the same capacity, with the same user interface, and with the same techniques that service duties may be familiar with. However, for those who thrive with clear boundaries between their responsibilities and their social conditioning, this may still be an issue [10].

RELATED WORK

Additionally, OSNs have taken a number of actions to safeguard sensitive data from various privacy concerns, as suggested by Shivangi Ghee Wala et al. Although these recommendations are important, designers believe that data security solutions are still lacking a well-defined conceptual framework. At its heart, this plan must be an idea of risk. In light of this, we propose an approach to threat monitoring for OSNs that will be implemented throughout the project. By linking risk levels to social media users, they entice people to consider how risky it would have been to contact them while divulging personal information. To calculate risk boundaries, they take into account the threat perceptions of their customers and employ measures of profitability and resemblance. We employ a dynamic risk assessment mentor technique, wherein individual threat behavior is demonstrated through a limited number of crucial individual exchanges. Another process for risk analysis that has been developed and evaluated using actual data is mentioned in this brief article.

In their publication from [2], Kaliyar et al. developed a technique they dubbed "Counterfeit Information Discovery Making Use Of a Deep Semantic Network." In contrast to the impacts of online discussion boards, which include face-to-face conversational formats, the incorporation of digital communication tools into co-located courses has received relatively less criticism. This study examined middle school students' expectations and impressions of two different communication styles in a co-located classroom setting: face-to-face (F2F) and computer-mediated interaction (CMC). Is there any research available in French? As a result, they distinguish between students who participate fully in in-person (F2F) classroom conversations and those who choose to keep quiet. These studies highlight the advantages of computer-mediated communication (CMC) over in-person interactions in co-locations and show that different students have varied opinions on F2F and CMC ("active" and "quiet," respectively). Network breaches and malicious cyberattacks present serious threats to public safety.

The third Aditi Gupta et al. proposed a method for identifying and removing false Facebook customer records. There is a serious problem with cyber criminals doing different evil deeds, putting people in grave danger to OSNs. There is already a thriving industry of record-based bootleg market administrators peddling these counterfeit goods. Because of Facebook's immense popularity and the difficulty in locating information about it, our research primarily focuses on detecting false information on this platform. Listed below are the trickiest aspects of our profession. It has really been a huge effort to gather data linking real and phoney Facebook profiles. The programming user interface and Facebook's stringent security mechanisms are constantly improving and adding new limits, making it harder to get user account information. The next stage is to use Facebook client channel data to identify customer profile behaviour and identify a set of 17 characteristics that are critical for distinguishing real Facebook users from fake ones. In the end, these highlights will be utilised to decide which AI-based classifiers shine out at recognising tasks out of a total of twelve classifiers.

Recognizing phony Twitter profiles The writers are Aktaş, B. Erçahin, D. Kilinç, and C. Akyol. The way individuals interact on social media sites like Facebook and Twitter has a significant impact on a lot of people's lives. The increased use of social media has resulted in a number of issues, one of which is the possibility of damaging content spreading through deception—people being led to believe they are someone else. The real world society could be seriously undermined by this predicament. In this study, we offer a classification technique for identifying fraudulent Twitter accounts. We pre-processed our dataset using the Worsening Reduction Discretization (EMD) method of monitored discretization on numerical features before analyzing the result of the Ignorant Bayes formula.

EXISTING SYSTEM

A lot of people's lives are improved as a result of their involvement with social media sites like Facebook and Twitter. The widespread use of social media has brought up several issues, one of which is the possibility that harmful online information might spread via the platform's ability to trick users into thinking they are someone they are not. Society on Earth might suffer serious harm as a result of this disease. Our research presents a category strategy for detecting phoney accounts on Twitter. To prepare our dataset for analysis, we used the EMD pre-processing approach on simplified numerical functions.

PROPOSED SYSTEM

A grouping of metadata-, online content-, communication-, and community-based criteria are used in the suggested strategy to recognize social spam bots on Twitter in order to identify phoney individuals. The majority of network-based qualities are not specified utilizing customer followers and underlying area structures when analyzing the defining functions of present techniques. This ignores the reality that a person's reputation inside a network is inherited from neighbors and followers rather than from those they follow. As a result, the system gives priority to using followers and neighborhood structures to define an individual's network-based attributes. I fake material, (ii) spam based on link, (iii) spam in famous topics, and (iv) fake persons are the four main categories that the system uses to divide a team of characteristics. Interaction- and community-based elements further subdivide the network group. Metadata characteristics are derived from additional information provided about a user's tweets, while content-based attributes aim to analyse a user's message publishing activity and the quality of the message they employ in posts. In order to deactivate network-based features, the user interaction network is used.

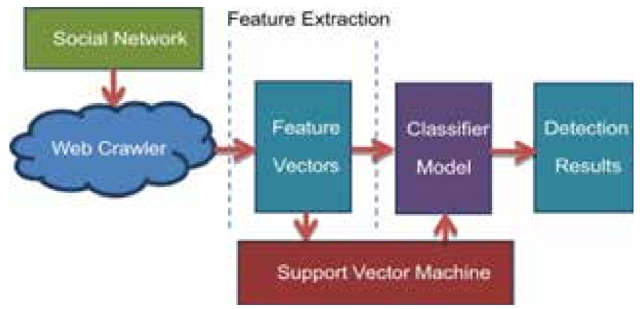


Fig.1. Spammer detection model

METHODOLOGY

The paper's author discusses a method for identify Twitter spam and fake accounts on the blog platform. In order to accomplish the task of detection, The author uses four different ways of detection: phony person identification, spam URL discovery, spam trending topic, and fraudulent online content. We will use the previously mentioned dataset to train the Random Woodland information mining algorithm, which will then recognize the percentage of spam to non-spam tweets in addition to erroneous and legitimate accounts. This process is similar to the previously described four-step process for formative whether a tweet is spam or not. Although we are utilizing the Random Forest classifier in this instance, authors frequently use various information mining techniques to determine whether or not a tweet is spam.

a summary of four techniques to ascertain whether a tweet is spam. You may compare the given ways using various criteria, such as customer features (such as retweets, tweets, and follows), material functions, and other features.

Online deception: When an account's number of followers is low compared to its variety of followers, it indicates that its online reputation is poor and that it is likely spam. Interactions between HTTP links, states and replies, warm topics, and tweet reputation online are all comparable services. A customer account is deemed spam if it quickly tweets out a large number of messages, according to the time function.

Identifying Spam links: The user-based features are determined by multiple parameters, such as the age of the account and the number of highly-loved listings and tweets by the user. The parsed JSON framework contains the discovered user-based properties. A few of the tweet-based characteristics are URLs, customer mentions, hashtags, and the quantity of retweets. We will undoubtedly identify whether a tweet has a spam link by using a type of artificial intelligence called Naive Bayes.

If one uses the Naive Bayes algorithm to identify tweet content, it is probable to ascertain if a trending topic contain phrase that are measured spam or not. This algorithm will search for terms with mature content, spam links, and duplicate tweets. If the Naive Bayes algorithm detects SPAM in a tweet, it will undoubtedly return 1, and if not, it will return 0.

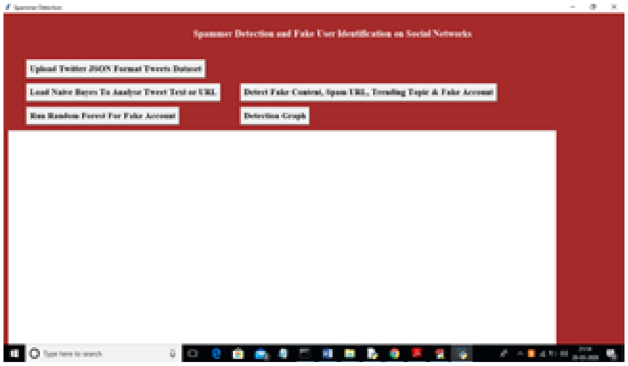
Improper Person Acknowledgement: Examples of these attributes include the age of the description, the extent of followers and fans, and the amount of followers. Content qualities are closely connected to user-generated content, in contrast to spammers who simply send out a few duplicate tweets. This is a result of the same content that spam crawlers frequently generate. This method takes relevant data from tweets and classifies them as spam or non-spam according to whether or not they include spam material. It does this by applying the Naive Bayes algorithm. In the future, the random forest method would be worn to train these features in order to identify whether or not an account is fraudulent. The features.txt files will contain every feature that was extracted. A naive Bayes classifier may be found in the "model" folder.

Using the methods described above, we can tell whether a tweet has real content or spam. If social media platforms can identify and delete these spam messages, their public image would greatly improve. If spam communications are not removed from social media networks, their attraction may decline. Keeping social media accounts spam-free is a great way to gain credibility as modern consumers rely on them heavily for news, company, and home information.

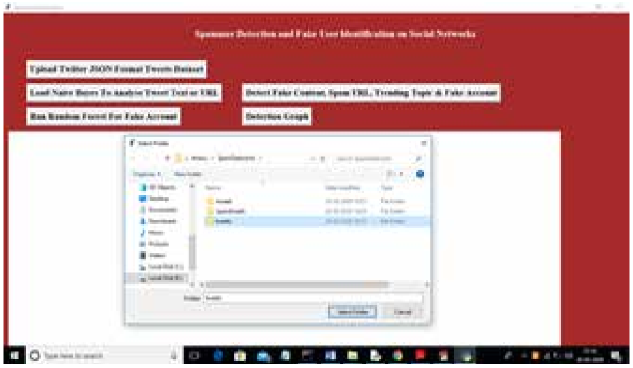
The JSON-formatted Twitter dataset that we are using for this project includes details about people, their tweets, the quantity of fans and followers they have, the amount of favorite tweets, and more. We analyze all of the data determining the authenticity of user accounts and whether they comprise spam or regular messages by utilizing the Python JSON API. All of those dataset files are located in the "tweets" folder.

IMPLEMENTATION

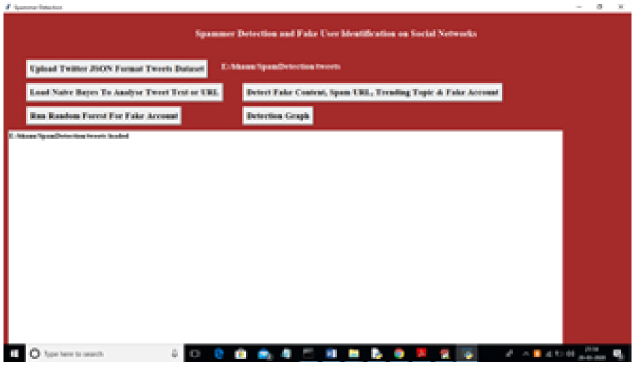
Double-click the "run.bat" file to start the project. This will display the screen that is seen below.


Spammer Detection
Spammer Detection and Fake User Identification on Social Networks
Upload Twitter JSON Format Tweets Dataset
Load Naive Bayes To Analyse Tweet Text or URL
Detect Fake Content, Spam URL, Trending Topic & Fake Account
Run Random Forest For Fake Account
Detection Graph
Type here to search

In the window described above, there is a button labeled "Upload Twitter JSON Format Tweets Dataset". Next, submit the folder with the tweets inside of it.

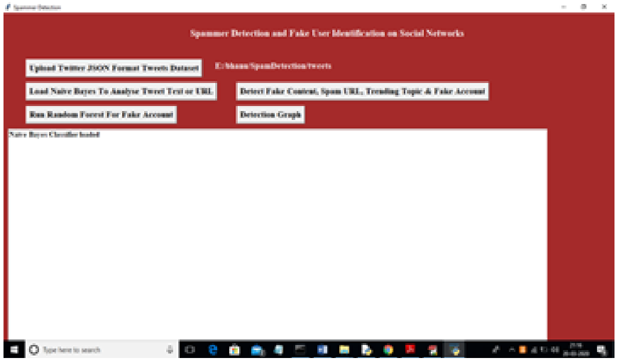

Spammer Detection and Fake User Identification on Social Networks
Upload Twitter JSON Format Tweets Dataset
Load Naive Bayes To Analyse Tweet Text or URL
Detect Fake Content, Spam URL, Trending Topic & Fake Account
Run Random Forest For Fake Account
Detection Graph
Select Folder
Select Folder
Cancel
Type here to search

You can see that I uploaded a folder called "tweets" that has tweets in JSON format from various users in the display screen above. Right now, relate to the open button to start reading tweets.


Spammer Detection
Spammer Detection and Fake User Identification on Social Networks
Upload Twitter JSON Format Tweets Dataset
E:/bhanu/SpamDetection/tweets
Load Naive Bayes To Analyse Tweet Text or URL
Detect Fake Content, Spam URL, Trending Topic & Fake Account
Run Random Forest For Fake Account
Detection Graph
Type here to search

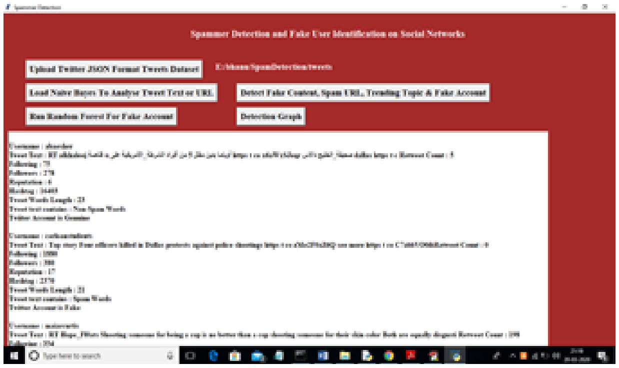
```
E:/bhanu/SpamDetection/tweets loaded
```

On the screen up above, we are able to see each single person's loaded tweet. Select "Load Naive Bayes To Analyze Tweet Text or URL" to bring up the catogorized called Naive Bayes.


Spammer Detection
Spammer Detection and Fake User Identification on Social Networks
Upload Twitter JSON Format Tweets Dataset
E:/bhanu/SpamDetection/tweets
Load Naive Bayes To Analyse Tweet Text or URL
Detect Fake Content, Spam URL, Trending Topic & Fake Account
Run Random Forest For Fake Account
Detection Graph
Type here to search

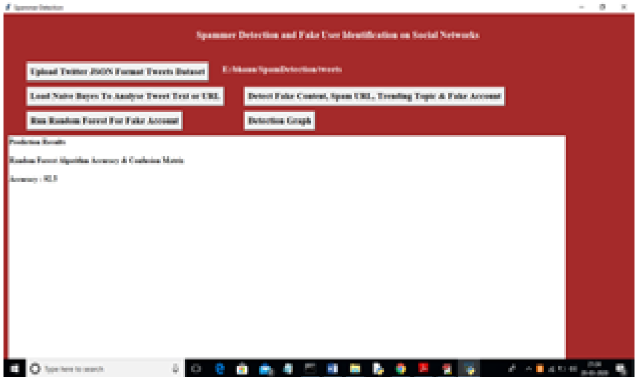
```
Naive Bayes Classifier loaded
```

Select "Detect Fake satisfied, Spam URL, Trending Topic & Fake Account" to apply the Naive Bayes classifier and the opposite methods mentioned before to look at every tweet for unsolicited mail URLs, fraudulent accounts, and faux content material. Up there, you may see that the Naive Bayes classifier has already been loaded.


Spammer Detection
Spammer Detection and Fake User Identification on Social Networks
Upload Twitter JSON Format Tweets Dataset
E:/bhanu/SpamDetection/tweets
Load Naive Bayes To Analyse Tweet Text or URL
Detect Fake Content, Spam URL, Trending Topic & Fake Account
Run Random Forest For Fake Account
Detection Graph
Type here to search

```
Username : [illegible]
Tweet Text : RT [illegible] Retweet Count : 5
Following : 75
Followers : 278
Reputation : 4
Hashtag : 16403
Tweet Words Length : 23
Tweet text contains : Non Spam Words
Twitter Account is Genuine

Username : [illegible]
Tweet Text : Top story Four officers killed in Dallas protests against police shootings https://t.co/[illegible] see more https://t.co/[illegible]Retweet Count : 0
Following : 1500
Followers : 388
Reputation : 17
Hashtag : 2378
Tweet Words Length : 21
Tweet text contains : Spam Words
Twitter Account is Fake

Username : [illegible]
Tweet Text : RT [illegible] Retweet Count : 198
Following : 254
```

On order to discover spam tweets, all of the characteristics from the dataset are retrieved and evaluated at the display that you may see above. The information proven in every tweet record includes things like TWEET TEXT, FOLLOWERS, FOLLOWING, and so on., in conjunction with information on whether the explanation is actual or no longer and if the tweet text includes junk mail or not. The textual content box up there has a line between every file fee. Click the "Run Random Forest Prediction" button to educate a random woodland classifier the usage of the traits of the obtained tweets. Incoming tweets from faux or unsolicited mail bills can be predicted or detected by the usage of this method. You may also see the info of each tweet by scrolling down above the textual content box.

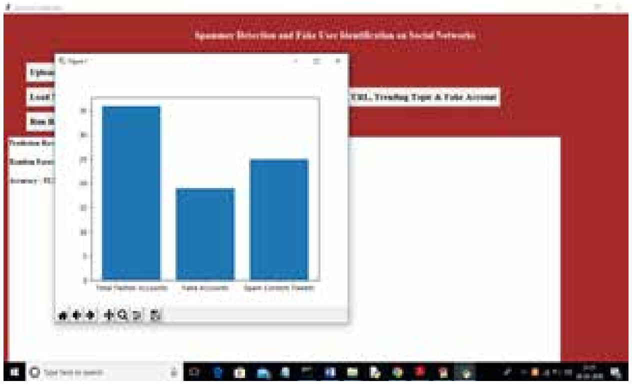

Spammer Detection and Fake User Identification on Social Networks
Upload Twitter JSON Format Tweets Dataset
Load Naive Bayes To Analyse Tweet Text or URL
Detect Fake Content, Spam URL, Trending Topic & Fake Account
Run Random Forest For Fake Account
Detection Graph
Type here to search

Prediction Results

Random Forest Algorithm Accuracy & Confusion Matrix

Accuracy : 92.5

Connect the "Detection Graph" button to get a graph that shows every possible combination of spam, bogus accounts, and tweets. In the aforementioned graphic, we discovered that the random forest prediction had a 92% accuracy rate.


Spammer Detection and Fake User Identification on Social Networks
URL, Trending Topic & Fake Account
Total Twitter Accounts
Fake Accounts
Spam Content Tweets
Type here to search

The x-axis of the graph above shows the overall wide variety of tweets, while the y-axis shows the matter of bogus bills and tweets the use of spamming language.

CONCLUSION

Methods for identifying Twitter spammers were the focus of this study. In calculation, we provided a taxonomy of Twitter spam finding approach, classifying them according to features including false user discovery, fraud detection in warm subjects, link detection, and fake content detection. To compare the offered approaches, we used a number of criteria, such as those pertaining to individuals, online content, charts, frameworks, and time. Additionally, the techniques were compared according to the datasets they used and the objectives they were designed to accomplish. Scientists should find it simpler to find information about novel Twitter spam detection approaches in one location thanks to the current assessment. Even though robust and effective techniques for spam detection and false user identification on Twitter have been developed, there are still some gaps that require in-depth research. The issues are emphasised in a concise manner as follows: The topic of misleading information appearing on social networks needs to be investigated since false news can have serious negative impacts on people's lives and on society as a whole. Further investigation into the sources of rumors on social media platforms is a worthwhile and significant field of study. While some effort has been done in the past to track down the sources of rumors using analytical methods, more advanced strategies, such those depending on social media, might be more successful.

Evaluation of Functions

There are still some research gaps that need to be filled, despite the fact that efficient and successful techniques for identifying bogus users and detecting spam on Twitter contain be urbanized. Many of the troubles have elements of the next: Studying how to spot bogus content on social media networks is crucial since fake news has serious negative effects on people's lives as well as society at large. Determining the source of rumors on social media is an additional crucial issue that requires research. Even while some recent research uses analytical ways to discover the source of rumors, more complicated strategies, like those depending on social networks, may be worn due to their demonstrated efficiency.

ACKNOWLEDGMENT

We express our gratitude to CMR Technical Campus for funding this paper, "FAKE PROFILE IDENTIFICATION IN SOCIAL NETWORK USING MACHINE LEARNING AND NLP," which provided us with great tools and help finishing our homework. For their wise counsel and direction in all aspects of our work, I would like to thank our chairman, director, deans, head of the department, computer science and engineering department, mentor, and teaching and non-teaching faculty members.

REFERENCES

1. Social Networks Analysis and Mining (ASONAM) 2018 Aug 28 (pp. 1191-1198). IEEE.

2. Pakaya FN, Ibrohim MO, Budi I. Malicious Gheewala S, Patel R. ML based Twitter Spam account detection: a review. In2018 Second International Conference on Computing Methodologies and Communication (ICCMC) 2018 Feb 15 (pp. 79-84). IEEE.

3. Kaliyar RK. Fake news detection using a deep neural network. In2018 4th International Conference on Computing Communication and Automation (ICCCA) 2018 Dec 14 (pp. 1-7). IEEE.

4. Erşahin B, Aktaş Ö, Kılınç D, Akyol C. Twitter fake account detection. In2017 International Conference on Computer Science and Engineering (UBMK) 2017 Oct 5 (pp. 388- 392). IEEE.

5. Gupta A, Kaushal R. Towards detecting fake user accounts in Facebook. In2017 ISEA Asia Security and Privacy (ISEASP) 2017 (pp. 1-6). IEEE.

6. Alom Z, Carminati B, Ferrari E. Detecting spam accounts on Twitter. In2018 IEEE/ACM International Conference on Advances in Account Detection on Twitter Based on Tweet Account Features using Machine Learning. In2019 Fourth International Conference on Informatics and Computing (ICIC) 2019 Oct 16 (pp. 1-5). IEEE.

7. Jardaneh G, Abdelhaq H, Buzz M, Johnson D. Classifying Arabic tweets based on credibility using content and user features. In2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT) 2019 Apr 9 (pp. 596-601). IEEE.

8. Harjule P, Sharma A, Chouhan S, Joshi S. Reliability of News. In2020 3rd International Conference on Emerging Technologies in Computer Engineering: ML and Internet of Things (ICETCE) 2020 Feb 7 (pp. 165-170). IEEE.

9. Dr.C K Gomathy, Article: A Study on the recent Advancements in Online Surveying , International Journal of Emerging technologies and Innovative Research (JETIR) Volume 5 | Issue 11 | ISSN : 2349-5162, P.No:327-331, Nov-2018

10. B. Erçahin, Ö. Akta³, D. Kilinç, and C. Akyol, ``Twitter fake accountdetection,'' in Proc. Int. Conf. Comput. Sci. Eng. (UBMK), Oct. 2017,pp. 388_392.

Malicious URL Detection Based on Machine Learning

TSS Bharath, Bondugulapati Srinivas Rao,
Dasyam Akshay Kumar
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

Ranjith Reddy
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ ranjith.kssr5@gmail.com

ABSTRACT

At present, the risk of community statistics insecurity is rapidly increasing both in quantity and severity. These days, cyberpunks specifically employ methods for managing human vulnerabilities and striking end-to-stop generation. Social engineering, phishing, pharming, and numerous more tactics are examples of these tactics. Lying to clients using risky Attire Source Locators (Links) is one step in executing these tactics. As such, detecting fraudulent URLs is becoming a very popular hobby. Numerous clinical investigations have undoubtedly revealed a variety of methods for finding harmful URLs that are entirely dependent on artificial intelligence and deep learning approaches. In this work, we propose a harmful URL discovery approach that leverages machine learning approaches based on our recommended link attributes and behaviors. Moreover, big-data innovation is also employed to enhance the capability of identifying dangerous relationships based on common behaviors. Put another way, a maker learning algorithm, a vast data cutting-edge period, and a new set of URL attributes and behaviors make up the encouraged detecting gadget. The results of the experiment demonstrate that the suggested link characteristics and behavior can significantly improve the ability to identify rogue URLs. It is claimed that the offered technique could be viewed as an optimized and superior applied solution for unfavorable URL discovery.

KEYWORDS: *URL, ML, Hyper link, Trademarks, Phishing.*

INTRODUCTION

Internet resources are referred to by their Attire Source Locator (URL). In their presentation on the features and two fundamental parts of the connection, Sahoo et al. [1] detailed the following: the procedure identifier, which specifies the protocol to be used, and the source name, which includes the domain name or IP address of the source. As you can see, each link adheres to a specific syntax and structure. In an attempt to trick users into sharing their malicious URL, attackers frequently attempt to change different elements of the URL structure. URLs that pose a risk to users are known as hazardous URLs. The URLs in question will lead visitors to malicious websites, phishing attempts, or other malicious code installations, as well as to unwelcome and potentially harmful content. Furthermore, malicious URLs may be concealed in seemingly safe download links, allowing them to swiftly proliferate via shared documents and messages on public networks. Several attack tactics, such as Drive-by Download, Phishing and Social Engineering, and Spam, use malicious URLs [2, 3, 4].

Attacks using the distributing malicious link technique rank first among the ten most common attack methods in 2019, according to statistics provided in [5]. As a result, the number of assaults and the level of risk are both increased by the three main methods of URL propagation: dangerous links, botnet URLs, and phishing URLs.

The figures demonstrating an increase in the variety of harmful URL distributions over the course of years demonstrate the necessity of researching and putting into practice strategies to identify and stop these dangerous Links.

There are two current trends in the field of hazardous URL discovery: one that relies on indications or rules, and the other that uses behavioural analysis methods to find dangerous links. 1 and 2 Using a set of rules or a collection of pens, a method may quickly and accurately identify malicious URLs. But new malicious links that aren't in the set of established indications or restrictions won't be picked up by this approach. In order to identify malicious URLs based on their actions, one strategy involves using artificial intelligence or deep learning formulae. In this research, we classify Links according to their attributes using AI techniques. Also included in the study is a novel method for removing URL attributes.

Our study use machine learning algorithms to categorise URLs according to the properties and behaviours of Links. These new functionalities are derived from URLs' fixed and dynamic actions. The primary contribution of the research is the set of newly proposed characteristics. All parts of the harmful link finding system are machine learning techniques. Both Support vector machine (SVM) and random forest (RF) are used as supervised device learning techniques.

EXISTING SYSTEM

Destructive URL Discovery based on Trademarks

Extensive research on harmful URL identification using trademark sets has been conducted and implemented for quite some time [6, 7, 8]. In most cases, lists of known harmful links are used in these types of investigations. A data source query is executed every time a new link is visited. By default, URLs are considered safe; however, if they are blacklisted, they are considered harmful and a warning is generated. The main problem with this approach is that it will be very tough to discover new harmful links that aren't already on the provided list.

Discovering Malicious URLs using AI

When it comes to harmful link detection approaches, you may use one of three AI formulas: monitored knowing, not being watched understanding, or semisupervised learning. The discovery methodologies are built around the habits of links. Several harmful URL systems that rely on machine learning formulae have been examined in [1]. The following formulae are examples of maker learning algorithms: Support Vector Machines, Logistic Regression, Decision Trees, Ensembles, Online Discovery, etc. This work makes use of both the RF and SVM formulae. The results of the speculation will reveal the accuracy of the two algorithms with varied combinations of parameters.

There are two main groups that may be defined by Links' behaviours and traits: the fixed and the lively. Methods for examining and eliminating Lexical, Web content, host, and popularity-based fixed actions of Links were detailed in the aforementioned research papers [9, 10, 11]. These studies use SVM and Online Understanding formulae as their device learning algorithms. Using dynamic URL activities for harmful URL discovery is possible in [12, 13]. Both static and dynamic patterns of behaviour are used to derive connection properties in this research. Personality and semantic teams, the Irregular group on the internet and host-based teams, and the Associated team are some of the distinctive groupings that are explored.

Negative aspects

Artificial intelligence formula choice is not implemented to the system. Neither URL Associate Removal nor Selection is being implemented by the system.

EXPLANATION OF WORK

Links are identified in the proposed system using artificial intelligence algorithms according to their roles and behaviours. These characteristics are novel to the literature and are extracted from dynamic and static URL behaviours.

The study's primary payoff is those newly proposed qualities. Every component of the harmful URL finding system includes AI formulae. Two machine learning methods, Support vector machine (SVM) and Random forest (RF), are used for monitoring purposes.

Benefits

Our newly selected criteria for harmful URL detection are well-suited to be employed by the algorithms that have been suggested. Although not our primary concern, SVM and RF are used in the proposed study to demonstrate the overall detection system's outstanding performance. Various more algorithms like Naïve Bayes, Choice trees, k-nearest neighbours, semantic networks, and others are enticed to be executed by visitors.

IMPLEMENTATION

Supplier of Services

A valid username and password are required for the Service Provider to get right of entry to this module. Upon a hit login, he may be capable of do positive duties, like logging in, View Datasets at URLs and Train and Test Datasets, Check out the bar chart showing the skilled and tested accuracy of the URL datasets. You also can see the effects of the training and testing, in addition to the prediction of the URL type and the ratio of the URL kind. Get the Forecasted Data Sets, Analyse the Type Ratio of URLs, See Who Is Online From Afar

Monitor and Permit Users

This section lets in the administrator to get a whole rundown of all registered users. Here, the administrator may see the user's information (call, electronic mail, and address) and furnish them access.

Individual Working Remotely

Numerous users (n) are found in this module. Before doing any actions, the person is required to sign up. Details could be entered into the database after a user registers. After he has efficiently registered, he's going to want to log in using the permitted credentials. Upon a hit login, customers will be able to do moves which include registering and logging in, predicting URL types, and seeing their profiles.

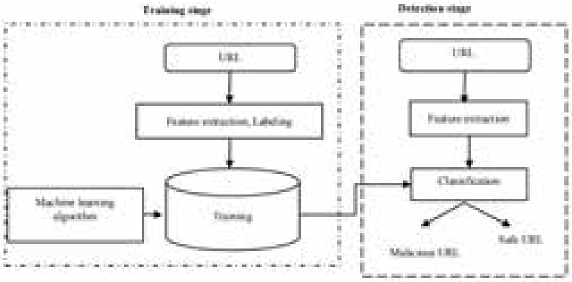


Fig.1. System architecture

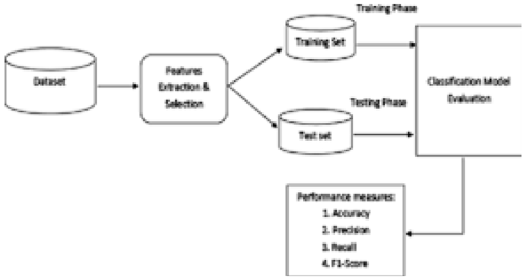
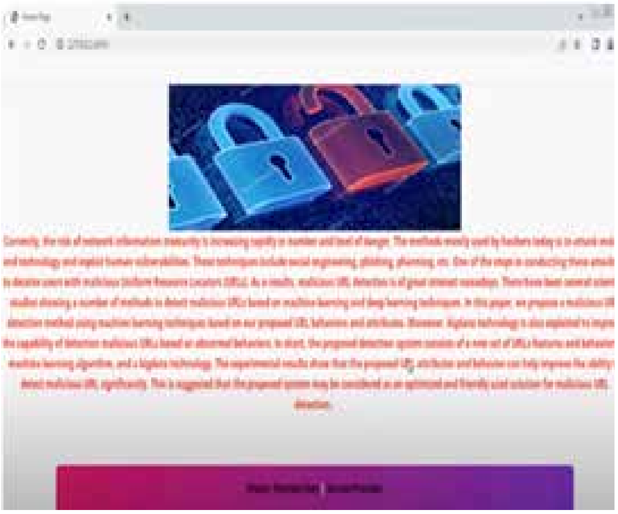
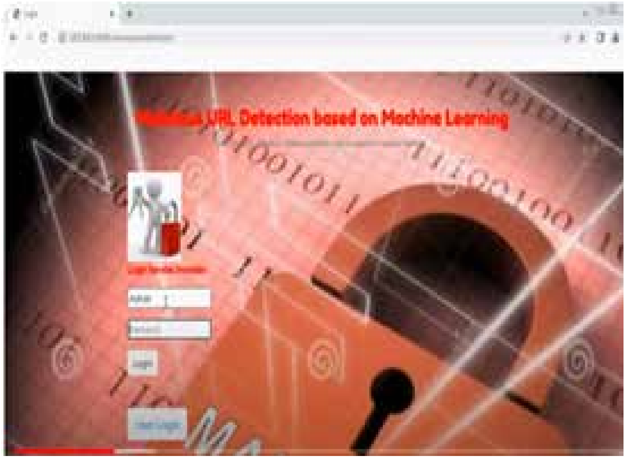


Fig.2. Output results.



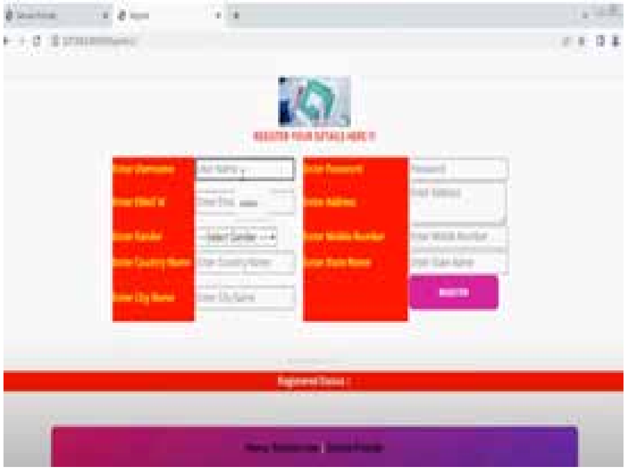
Currently, the risk of network information insecurity is increasing rapidly in number and level of danger. The methods mostly used by hackers today is to attack end-to-end technology and exploit human vulnerabilities. These techniques include social engineering, phishing, pharming, etc. One of the steps in conducting these attacks is to deceive users with malicious Uniform Resource Locators (URLs). As a results, malicious URL detection is of great interest nowadays. There have been several scientific studies showing a number of methods to detect malicious URLs based on machine learning and deep learning techniques. In this paper, we propose a malicious URL detection method using machine learning techniques based on our proposed URL behaviors and attributes. Moreover, bigdata technology is also exploited to improve the capability of detection malicious URLs based on abnormal behaviors. In short, the proposed detection system consists of a new set of URLs features and behaviors, a machine learning algorithm, and a bigdata technology. The experimental results show that the proposed URL attributes and behavior can help improve the ability to detect malicious URL significantly. This is suggested that the proposed system may be considered as an optimized and friendly used solution for malicious URL detection.

Fig.3. Web server page



Malicious URL Detection based on Machine Learning

Fig.4. Admin page



REGISTER YOUR DETAILS HERE !!

Enter Username
Enter EMail Id
Enter Gender
Enter Country Name
Enter City Name
Enter Password
Enter Address
Enter Mobile Number
Enter State Name

Registered Status :

Fig.5. New user registration page.

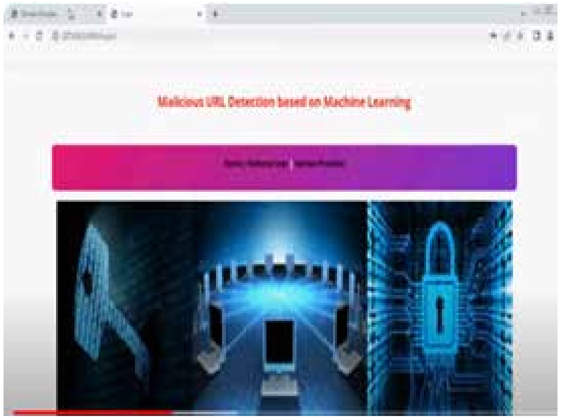


Fig.6. User page details

Fig.7. Upload dataset details

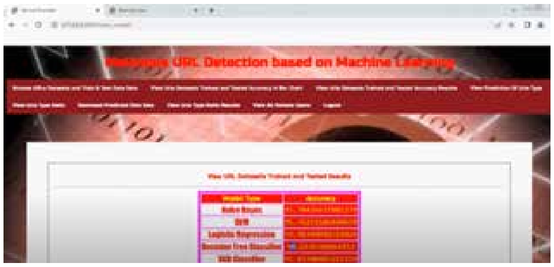


Fig.8. Accuracy output

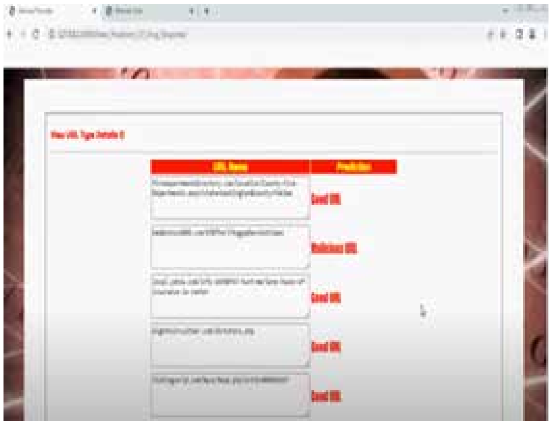


Fig.9. Final output

CONCLUSION

In this research, we provide an AI-based method for detecting harmful links. Tables V and VI demonstrate the performance of the suggested extracted attributes, which is based on empirical evidence. Unlike many other common publications, this one does not seek to build large datasets to improve the system's accuracy or leverage unique properties. The system's processing speed and accuracy are determined by the combination of characteristics that are simple to compute with technologies that analyse vast amounts of data. Information security systems may make use of the study's findings in their use of new technology for detailed security. A free tool to detect malicious URLs on browsers was developed using the findings of this article.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "MALICIOUS URL DETECTION BASED ON MACHINE LEARNING", which gave us excellent resources and assistance to complete our assignment. I want to express my gratitude to our chairman, director, deans, head of the department, computer science and engineering department, guide, and teaching and non-teaching faculty members for their insightful advice and direction in all facets of our work.

REFERENCES

1. D. Sahoo, C. Liu, S.C.H. Hoi, "Malicious URL Detection using MachineLearning: A Survey". CoRR, abs/1701.07179, 2017.

2. M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection: a literaturesurvey," IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp.2091–2121, 2013.

3. M. Cova, C. Kruegel, and G. Vigna, "Detection and analysis of drivebydownloadattacks and malicious javascript code," in Proceedings of the19th international conference on World wide web. ACM, 2010, pp. 281–290.

4. R. Heartfield and G. Loukas, "A taxonomy of attacks and a survey ofdefence mechanisms for semantic social engineering attacks," ACMComputing Surveys (CSUR), vol. 48, no. 3, p. 37, 2015

5. Internet Security Threat Report (ISTR) 2019–Symantec.https://www.symantec.com/content/dam/

symantec/docs/reports/istr-24-2019-en.pdf [Last accessed 10/2019].

6. S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong, and C. Zhang,“An empirical analysis of phishing blacklists,” in Proceedings of SixthConference on Email and Anti-Spam (CEAS), 2009.

7. C. Seifert, I. Welch, and P. Komisarczuk, “Identification of maliciousweb pages with static heuristics,” in Telecommunication Networks andApplications Conference, 2008. ATNAC 2008. Australasian. IEEE,2008, pp. 91–96.

8. S. Sinha, M. Bailey, and F. Jahanian, “Shades of grey: On theeffectiveness of reputation-based “blacklists”,” in Malicious andUnwanted Software, 2008. MALWARE 2008. 3rd InternationalConference on. IEEE, 2008, pp. 57–64.

9. J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, “Identifying suspiciousurls: an application of large-scale online learning,” in Proceedings of the26th Annual International Conference on Machine Learning. ACM,2009, pp. 681–688.

10. B. Eshete, A. Villafiorita, and K. Weldemariam, “Binspect: Holisticanalysis and detection of malicious web pages,” in Security and Privacyin Communication Networks. Springer, 2013, pp. 149–166.

11. S. Purkait, “Phishing counter measures and their effectiveness– literaturereview,” Information Management & Computer Security, vol. 20, no. 5,pp. 382–420, 2012.

A Machine Learning Approach for Rainfall Estimation Integrating Heterogeneous Data Sources

Lalith Singh, Munigoti Tharun, Narella Keerthi
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

G Pavan Kumar
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ pavangurram.reddy@gmail.com

ABSTRACT

Reducing risks associated with floods and landslides caused by heavy rains requires an accurate rainfall estimate at individual locations, which is no easy task. To get straight dimensions of rainfall intensity in these parameters, dense networks of sensing devices called rain gauges (RGs) are often used. In order to estimate the precipitation field throughout the whole rate-of-interest region, these dimensions are often added using spatial interpolation methods. However, these approaches are computationally expensive, and more data must be integrated to improve the assessment of the interest variable in unknown components. This work proposes a machine learning-based method to these problems; it uses an ensemble-based classifier for rainfall assessment and can combine data from several remote sensing dimensions. In cases where RGs are unavailable, the proposed method provides an accurate rainfall estimate, integrates disparate data sets by leveraging RGs' high quantitative precision with the spatial pattern recognition provided by radars and satellites, and uses less computational resources than interpolation methods. The experimental results for real data from the Italian region of Calabria demonstrate a significant improvement when compared to a well-known method for rainfall estimation, Kriging with external drift (KED), in terms of both the chance of discovery (0.58 versus 0.48) and mean-square error (0.11 versus 0.15).

KEYWORDS: *RG, KED, Rain fall estimation, Network, Classifier.*

INTRODUCTION

For many uses of hydro logical effect modelling, such as flood risk mitigation, river container management, disintegration modelling, and others, an accurate rainfall estimate is essential. Rain gauges (RGs) are used to directly measure the length and intensity of rainfalls at specific locations in order to achieve this goal. We use interpolation methods that are based on the data recorded by these RGs to estimate rain occurrences in areas that are not covered by them. The Kriging geo analytical approach [1, 2] is among the most utilised and recognised in the area, however there have been many recommendations for variants of these methods in literature. During severe convective weather occurrences, it is of the utmost importance to accurately recreate the rain's field in space. In example, thin RGs could miss extremely localised heavy precipitation from convective thunderstorms, and floods might form even when no rain is falling [3]. In order to overcome this problem, there is a recent trend in the literature to combine different rainfall data sources in order to get a more accurate estimate via the use of interpolation methods. [4]

One of the most popular ways to overcome the limitation of the widely-used regular Kriging (ALRIGHT)—which can only utilise one source of data as input—was Kriging with external drift (KED) [5], [6]. Unlike alright, KED may take into account secondary information, and it permits the insertion of a random region. The main problem is that these methods need a lot of resources and are computationally expensive.

Applying methods from the field of artificial intelligence (ML) is an alternative approach. But these methods come with a host of challenges that must be handled:

class imbalance, missing features galore, and the need to work gradually as new data becomes available. To fix these problems, people usually employ predetermined procedures. In ensemble [7], a classification approach is used to find previously unknown situations by combining several models that have been trained using different category formulae or data samples. Using an ensemble standard instead of a single classification design eliminates out-of-balance classes, reduces error variance and bias, and outperforms the single design case. In particular, difficulties with the rainfall estimate and with the monitoring of meteorological (extreme) events may be addressed using ensemble-based methodologies. These techniques may also capture nonlinear interactions, such as those between sensing unit data, cloud structures, and rainfall estimation, among others. Using an ordered probabilistic ensemble classifier (HPEC) for rainfall estimate, this research presents ML-based methods to overcome the primary difficulties of rains estimation. Using an under-tasting method to handle the out-of-balance classes problem typical of this situation, the proposed approach integrates data from multiple sources (e.g., RGs, radars, and satellites) and allows for accurate rainfall evaluation in situations where RGs are unavailable.

When it rains in a certain location that might be prone to landslides or floods, our method is an effective service for real-life scenarios like that of a Department of Civil Protection (DCP) officer. Data from the DCP is used in the experimental evaluation; the data pertains to the southern Italian region of Calabria. The diverse orography and striking climatic irregularities of Calabria make it an ideal testing site.

We may sum up our contributions by saying that they comply.

1) Three different data sources are used to provide more precise rainfall event estimates: RGs, radar, and Meteosat.

2) An ordered probabilistic set technique is suggested after contrasting several categorization approaches on an actual case relating to Calabria, a southern city in Italy.

3) We compare several ML-based methods that are exclusively trained on historical data with a frequently used interpolation approach in the hydrological sector, which is KED.

EXISTING SYSTEM

Even though the goal of this study is to develop a run-off assessment, an existing system is based upon the established standard including the job, which employs a probabilistic ensemble and blends two sources of data (i.e., rain gauges and radar), similar to our job. The results of the runoff hydrologic models are then mixed using a specific approach in order to isolate a single runoff hydrograph. The findings of the speculation show that the hydrologic models are correct, which may help with making better decisions during flood warnings. In order to get a probabilistic geographical assessment of daily rainfall using rain gauges, Frei and Isotta [13] outline a method. Depending on the observations, the final design represents a collection of possible fields that may be characterised as a Bayesian anticipating distribution evaluating the unpredictability due to the data coming from the station network. By analysing a real-life case study in the European Alps, we can see that the method can accurately predict the hydrological partitioning of the area.

A fascinating study of daily rainfalls for Australia and several parts of South and East Asia, based only on high-resolution evaluations, is proposed in [4]. Simply said, the average of the studies done for each resource will provide you the chosen version. In terms of global accuracy, the authors stress that the set technique is better than the individual components of the design. More information from other rainfall products may be recorded by the suggested model as well. The latter two tasks demonstrate that set procedures may guarantee outstanding outcomes in a rainfall estimate situation, since they both use a set system to provide more accurate forecasts. The adopted mix methodologies are simple, and a variety of heterogeneous data sources are not taken into account, which is different from our task.

Based on RG-only data and the integrated RG-radar product as a benchmark, Chiaravalloti et al. [6] studied the performance of three newly developed satellite-based items: IMERG, SM2RASC, and a smart combination of the two. The items were located in Calabria. A better quality satellite rainfall product may be obtained by combining IMERG with SM2RASC, and experiments show that IMERG performs well at temporal resolutions greater than 6 hours. The bulk of the alternative approaches use information gathered from many sources, such as radars and satellite channels. Several of them rely on finding appropriate models that use data to find the best specifications for these versions, which in turn change the relationship between clouds' optical and micro physical habitats. Where Different tasks use analytical methods to distinguish the designs. the years 19–21, for instance, rainfall assessments based on satellite multi spectral data are given by means of a Bayesian estimate, and techniques that use radar data as input generate recommended price quotations.

In their system, which uses RG observations and satellite data and an interpolation approach based on the Kriging technique, Verdin et al. [23] also use Bayesian estimates to approximately determine the design requirements. Although these approaches may provide interesting results, they need a very sensitive stage of parameters estimation for the specific version, which means that their flexibility and efficacy are sometimes compromised as a side consequence. Because of the very nonlinear nature of the relationships between sensing unit data, cloud features, and rainfall projections, more adaptable methods based on ML algorithms have recently been investigated. In [4], for example, ANNs combined with assistance vector makers handle the difficulty of recognising convective events and nearby stormy places. Unlike our study, the data collecting process does not include the use of RG measures in training the algorithm; instead, it involves refining information coming from the optical channels of the multi spectral instrument aboard of Meteosat Second Generation (MSG) satellites. In their proposal for an SVM-based rainfall assessment method, Sehad et al. [5] combine input data from multiple spectral networks on MSG and construct two models, one for daylight and one for nighttime.

Results are evaluated against similar ANN-based methods, with the sole purpose of validating the technique using arbitrary forest (RF) and RGs. In [6], we learn about another ANN-based approach; here, radar data is used as a suggestion to identify stormy pixels in a picture matrix. Using data from multispectral networks on MSG satellites, Kuhnlein et al. [7] estimate rainfall rates using RFs, and they also use the ensemble approach.

Negative aspects

No hierarchical probabilistic ensemble classifier (HPEC) is used in the system to predict when it will rain.

The system uses ANNs, or artificial neural networks, a method of forecasting that does not guarantee a perfect prediction.

PROPOSED SYSTEM

For example, our method provides a dependable answer for real-world problems, such as when a DCP officer wants to assess the rainfall in a specific area that might be prone to landslides or floods. Actual data pertaining to the southern Italian region of Calabria, provided by the DCP, is used in the experimental analysis. With its complex orography and very variable environment, Calabria makes for a great testing ground. You may sum up our payments by saying that they comply.

1) Rainfall event prices are improved by combining three diverse data sources: RGs, radar, and Meteosat.

2) Various categorization methods are evaluated using a real-world example involving the southern Italian region of Calabria, and the ordered probabilistic set strategy is suggested.

3) We compare several ML-based algorithms that are exclusively trained on historical data with a popular interpolation approach in the hydrological domain, namely KED.

Benefits of system:

- To address the issue of class imbalance, the proposed system uses an under sampling method and pre-processes raw data to make it evaluation-ready.
- By evaluating and training using effective ML Classifiers, the proposed system created an Effect of Incorporating RG, Satellite, and Radar Measurements.

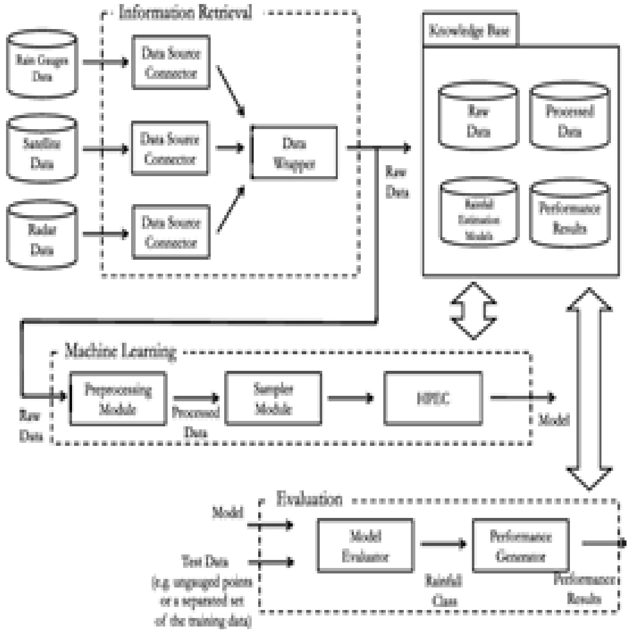
IMPLEMENTATION

Company

The Provider must provide a valid username and password in order to access this module. Once he logs on, he'll be able to perform things like check out the Data Sets and Training and Evaluation, Checked and Trained Accuracy Shown in a Bar Chart, Learned and Tested Accuracy Results Revealed, Discover the Rain Approximated Predicted Type Proportion, View the Rain Estimated Predicted Type Information, Install and Download Anticipated Data Sets, Visual Rainfall Estimated Anticipated Kind Ratio Results, Check Out Every Remote User may be seen and licensed. Here the administrator may see a complete roster of all registered users. User names, email addresses, and physical addresses are all viewable by the admin, who may also approve or disapprove users.

Working from a far

Here, you'll find n different types of users. Before proceeding with any activities, customers are required to register. Customers' details will be securely stored in the database upon registration. Upon successful enrollment, he will be prompted to log in using the authorised credentials. After logging in, users will be able to do things like see their profile, get a rain quote, and more.



CONCLUSION

The spatial rainfall area estimation will be done using an ML-based method. With the utilisation of heterogeneous information resources including RGs, radars, and satellites, this method may estimate rainfall in areas without RGs while still taking use of the spatial pattern recognition provided by these other sources. Using an HPEC, the design may be developed to approximately measure the intensity of rainstorm occurrences after a pre-processing step, after which a random attire under tasting approach is used. The two-tiered architecture of this ensemble begins with the training of a set of RF classifiers; the second tier makes use of a probabilistic steel income earner to combine the estimated probabilities provided by the base classifiers in accordance with a stacking schema. The Division of Civil Protection provides real data, and speculative findings show that it performs far better than Kriging with outer drift, a widely used and well-known method for rain prediction. In instance, the set strategy performs better at identifying when it will rain. Compared to the numbers obtained by KED (0.48 and0.15, respectively), HPEC's CASE (0.58) and MSE (0.11) stages are clearly much superior. In terms of the final two courses, which stand in for instances of severe rainfall, HPEC is computationally much more efficient than the Kriging method, although the difference between the two is not statistically significant (with respect to F-measure).

In reality, when evaluating a large number of elements, the technique becomes computationally costly due to the fact that the complexity of the Kriging approach is cubic in the variety of the cases . Actually, there is a square complexity in the ML algorithms (here meaning RF). Furthermore, ensemble methods are highly parallelized and salable. Because of this, we think our solution has some useful benefits for this problem. In addition, when looking at how different data sources are integrated, it seems that all of them contribute to the strategy's great success. Removing the RG information improves the formula's performance and makes all the actions more sensible. Although the degradation is less noticeable when only one type of data is removed, the lowest value of the MSE (0.11) is obtained when all data is used, proving that all data resources must be utilised for substantially better outcomes. We want to evaluate the method on a longer time frame in future work to account for outcomes caused by annual and seasonal fluctuation, and we are also considering the prospect of gradually improving the flexible set version using the additional data. In addition, we want to use time series analysis to determine the relative contributions of the various radar and Meteosat features in order to evaluate the algorithm's performance in identifying locally concentrated heavy rainfall events.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "A MACHINE LEARNING APPROACH FOR RAINFALL ESTIMATION INTEGRATING HETEROGENEOUS DATA SOURCES", which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. J. E. Ball and K. C. Luk, "Modeling spatial variability of rainfall over acatchment," J. Hydrologic Eng., vol. 3, no. 2, pp. 122–130, Apr. 1998.
2. S. Ly, C. Charles, and A. Degré, "Different methods for spatial interpolation of rainfall data for operational hydrology and hydrological modeling at watershed scale. a review," Biotechnologie, Agronomie, Société et Environnement, vol. 17, no. 2, p. 392, 2013.
3. H. S. Wheater et al., "Spatial-temporal rainfall fields: Modelling and statistical aspects," Hydrol. Earth Syst. Sci., vol. 4, no. 4, pp. 581–601, Dec. 2000.
4. J. L. McKee and A. D. Binns, "A review of gauge–radar merging methods for quantitative precipitation estimation in hydrology," Can. Water Resour. J./Revue Canadienne des Ressources Hydriques, vol. 41, nos. 1–2, pp. 186–203, 2016.
5. F. Cecinati, O. Wani, and M. A. Rico-Ramirez, "Comparing approaches to deal with non-gaussianity of rainfall data in Kriging-based radargauge rainfall merging," Water Resour. Res., vol. 53, no. 11, pp. 8999–9018, Nov. 2017. [6] H. Wackernagel, Multivariate Geostatistics: An Introduction With Applications. Berlin, Germany: Springer, 2003.
7. L. Breiman, "Bagging predictors," Mach. Learn., vol. 24, no. 2, pp. 123–140, Aug. 1996.
8. B. J. E. Schroeter, Artificial Neural Networks in Precipitation Now-Casting: An Australian Case Study. Cham, Switzerland: Springer, 2016, pp. 325–339.
9. X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in Proc. 28th Int. Conf. Neural Inf. Process. Syst., vol. 1, Dec. 2015, pp. 802–810.
10. W.-C. Hong, "Rainfall forecasting by technological machine learning models," Appl. Math. Comput., vol. 200, no. 1, pp. 41–57, Jun. 2008.

Predicting Drug-Drug Interactions based on Integrated Similarity and Semi-Supervised Learning

M. Paul Abhishek Emmanuel, Palem Pullarao, Bhugolla Nandu
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

D. Sandhya Rani
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ davu.sandhya@gmail.com

ABSTRACT

When one medicine's pharmacological outcomes are modulated by another, this phenomenon is called a drug-drug interplay (DDI). Negative DDIs result in severe medicinal drug responses, which may be fatal for patients or motive the drugs to be eliminated from the market. In contrast, nice DDIs frequently decorate patients' therapeutic outcomes. Drug discovery and contamination remedies now rely heavily on DDI identity. Here, we gift DDI-IS-SL, a brand new technique for DDI prediction that mixes semi-supervised studying with incorporated similarity. DDI-IS-SL makes use of the cosine similarity method to determine how comparable medicinal drugs are primarily based on their functions by integrating facts from the medication' chemicals, biology, and phenotype. Drug similarity as measured by using the Gaussian Interaction Profile kernel is likewise decided on the use of known DDIs. To determine the ratings for the potential of interactions between drugs, a semi-supervised mastering method known as the Regularised Least Squares classifier is used. When as compared to different processes, DDI-IS-SL demonstrates superior prediction ability in five-fold, 10-fold, and denote drug validation. On top of that, DDI-IS-SL has a faster common calculation time as compared to its competition. Case studies conclude by way of presenting more evidence of DDI-IS-SL's effectiveness in real-world scenarios.

KEYWORDS: *DDI-IS-SL, DDI, Drug chemical, Semi supervised learning.*

INTRODUCTION

It is common practice to give a person two or more drugs at once since each drug has its own unique pharmacological impact [1]. These groups, which are also known as drug-drug interactions (DDIs), may be advantageous or detrimental to efficiency depending on the results seen by professionals [2]. More effective treatments and less human suffering can be provided by positive DDIs. Still, the majority of adverse response occurrences originate from undesired DDIs [3]. Serious cases may lead to the drug market pulling medications and, in the worst case scenario, a client treated with many prescriptions dying. At the present day, multi-drug treatments are widely used to treat a variety of diseases or complicated situations, including cancer cells [4]. Reducing pain, increasing therapeutic efficacy, and raising overall survival rates are the initial goals of multi-medication therapy. In addition to the budgetary issue, the therapy's efficacy has been compromised due to the development of unwanted DDIs brought about by the increased usage of substances in the synergistic treatment. Pharmaceuticals including lipid-lowering drugs, macrolides, and oral anti fungal medicines are often used in combination therapies, and new investigations have shown that these drugs are very likely to interact with one other [5]. There have been pharmaceutic, pharmacokinetic (PK), and pharmacodynamic studies conducted on DDIs. In most cases, pharmaceutic DDIs arise as a consequence of chemical conflicts between many medications. A diamagnetic interaction (PK) is the impact of one medication on the absorption, distribution, or metabolism of another drug in the patient's body; this interaction is often associated with adverse reactions [6]. PD interactions may occur when two or more drugs have an effect on the same receptor, location, or physiological system; these effects can be additive or additive harmful to patients. Prior investigations have actually assumed DDIs using a large number of PK and PD communications [7]. Patients need physicians with extensive knowledge of people, antimicrobial medications, and microscopic microbes in order to make informed treatment decisions. The availability of additional diagnostic tools, pharmaceuticals, and medical professionals is always expanding thanks to the daily publication of new research [8]. This makes it more challenging than ever before for professionals to prescribe a course of therapy or medication to a patient based on their symptoms and health history. Reviews on products have become an integral part of the buying process for almost all products due to the meteoric rise of the internet and e-commerce. Worldwide, consumers have become used to researching products online and reading reviews before making a purchase [9]. The placement of healthcare or healing medications has seldom been discussed in previous study, which mostly focused on the shopping sector and its assumptions and proposals. People are increasingly seeking online diagnoses as a result of growing health concerns. For example, a Seat An American Proving ground research from 2013 found that 35% of consumers looked for ways to improve their health and well being online, and that 60% of individuals looked for information on health-related topics on the internet. a medicine that kills bacteria Having a recommended system in place is crucial for both professionals and people when it comes to building knowledge about medications for particular health concerns [10].

SURVEY OF RESEARCH

In recent times, several computer methods have been created to predict future DDIs, all based on AI ideas. Drug adverse event profiles are the mainstays of the signal finding method that Tatonetti et al. used to assume DDIs [1]. An INDI (INferring Drug Interactions) structure was developed to anticipate DDIs. This structure took into account medication chemical similarities, negative effects similarities, protein interaction similarities, and target sequence similarities. It employed two categories of medication communications: potential CYP (Cytochrome P450)-related DDIs and non-CYP-related DDIs, or NCRDs. [2] in

Cricotinib was used in a PBPK (physiologically based pharmacokinetic) methodology to predict DDIs when combined with ketoconazole or rifampin. Special DDIs were also found by using text-mining and reasoning algorithms based on drug metabolic process features [3]. Vilar et al. estimated DDIs by comparing medications based on their molecular fingerprint and molecular framework similarities.

Vilar et al. strengthened a technique suitable for extensive data in order to deduce distinct DDIs using 2D and 3D molecular structures, communication patterns, target and side-effect similarities, and so on. Cheng et al. [4] presented a computer method for DDI prediction based on medical phenotypic, restorative, chemical, and genetic characteristics as well as an AI model. Li et al. developed a computational approach to determine the mix efficiency of drugs using a Bayesian network design [5] based on the medicine's molecular and phenotypic similarities. A computer technique for DDI forecasting was proposed by Liu et al. [6] using a random woodland model. This method incorporates chemical interactions, protein communications across medication targets, and target enrichment of KEGG pathways. In order to extract the relevant aspects of drugs, our approach used a feature choice methodology.

By using the chemical-protein interactive, which provided a web server (referred to as DDICPI), Luo et al. developed a computational method for DDI anticipation. [7] By combining networks of several pharmaceutical similarities and known DDs, Sridhar et al. were able to use a PSL (Probabilistic Soft Logic) technique based on structural probabilistic soft logic to predict unique DDIs. Takako et al. used a logistic regression variant to predict potential DDIs using 2D pharmaceutical architecture similarities [8]. Combining ratings based on targets and enzymes improves its prediction performance even more.

Ferdousi et al. presented a computational method for DDI forecasting based on inner product based similarity measures (IPSMs). This approach also made use of the similarity between medicines and the key biological components that make them up, such as drug targets, enzymes, transporters, and carriers. Additionally, NLLSS (Network-based Laplacian regularised Least Square Synergistic medication mix forecast) was proposed to anticipate hidden collaborating medication combinations, on the premise that collaborating effects with drugs are usually comparable and vice versa; however, it is unable to predict DDIs for novel drugs. [9]

EXISTING SYSTEM

A plethora of device mastering-primarily based computational tactics for DDI prediction have emerged in recent years. Tavonatti et al.'s signal discovery method derives DDIs from drug unfavourable occasion profiles, that are the maximum critical pharmaceutical residences. By integrating shared capabilities in terms of chemical composition, detrimental outcomes, protein-protein interactions, and target sequences, a powerful medicinal drug may be advanced. For the reason of DDI prediction, the INDI (Inferring remedy Interactions) framework used two types of pharmaceutical interactions: people who may be related to CYP (Cytochrome P450) and people that are not.

PROPOSED SYSTEM

We develop a computer method (DDI-IS-SL) to forecast DDIs in this work by combining the pharmacological, organic, and phenotype aspects of drugs. Medications' chemical structures, target communications, enzymes, transportation, pathways, indications, side effects, off-side affects, and understood DDIs are all part of the drug data set. We begin by building a high-dimensional binary vector using these medication data elements in order to use the cosine similarity approach to find the medicines' attribute similarity. We also calculate the bit similarity of medications' Gaussian Interaction Profiles (GIPs) [8] using recognised DDIs. Function and GIP similarity are the building blocks of medication similarity. After that, DDI prediction is done using an RLS classifier [9]. We also use the node-based medication network diffusion method to find the relational early ratings of new pharmaceuticals that do not communicate with any existing medications. Consequently, our method may predict possible DDIs for both well-known and novel drugs. We carefully evaluate our method's and competing techniques' forecast efficiency using 5-fold cross validation, 10-fold cross recognition, and indicate validation. One way to measure the effectiveness of computational methods is by looking at their area under the ROC curve, or AUC. When compared to other completion methods, ours has a higher AUC. Particularly in 5-fold cross recognition, our technique outperforms the contemporary L1E with an AUC of 0.9691, outperforming it by a significant margin. In addition, our method outperforms L1E's best result (0.9599) in the 10-fold cross validation, with an AUC value of 0.9745. With an area under the curve (AUC) of 0.9292—higher than the best result of many other methods (WAE, weighted typical set approach, 0.9073)—our method also achieves the greatest prediction efficiency in de novo medicine identification. Furthermore, when compared to other competing ways, our method has a higher running effectiveness based on the comparison of the usual running time. Lastly, case study verification results show that DDI-IS-SL is a trustworthy computational method for predicting new DDIs, and they confirm our approach's prediction capabilities in real-world applications.

WORKING METHODOLOGY

In order to achieve the goal of DDI forecasting, the system's many components work together. The first part is in charge of gathering and cleaning up the data. This involves scouring many sources, such drug databases and scientific literature, for information on the medications, their chemical residential or commercial qualities, and their well-known communications. After that, the data is stabilised, cleaned, and converted into a format that is perfect for further analysis. The similarity module, which follows, compares drug sets according to their chemical qualities, including molecular structure, socioeconomic residential features, and pharmacological outcomes. The similarity scores are calculated by this component using a number of similarity actions, including the Tanimoto coefficient and the Euclidean distance. Next, the category designs are trained utilising the pre-processed data and similarity ratings using the supervised knowing component. For DDI forecasting, we have used five classification ML formulas: decision tree, logistic regression, k-nearest neighbour, random woodland, and assistance vector device. Previous research projects' results and the formulae' capacity to handle complex, high-dimensional data are the deciding factors in their selection. To evaluate the efficacy of the category designs, the system is examined using a number of measures, including recall, accuracy, precision, and F1-score. In order to ensure that the models are successful and can generalise, they are evaluated on both the training and testing data. Healthcare providers may benefit from the proposed system's ability to shed light on potential DDIs and aid in informed drug prescription decisions. Electronic health records and medication databases may also benefit from the system's ability to provide real-time warnings and notifications about potential DDIs.

Predicting drug-drug interactions based on integrated similarity and semi-supervised learning
Sign In
SERVICE PROVIDER
REGISTER

Fig. 1. Home page

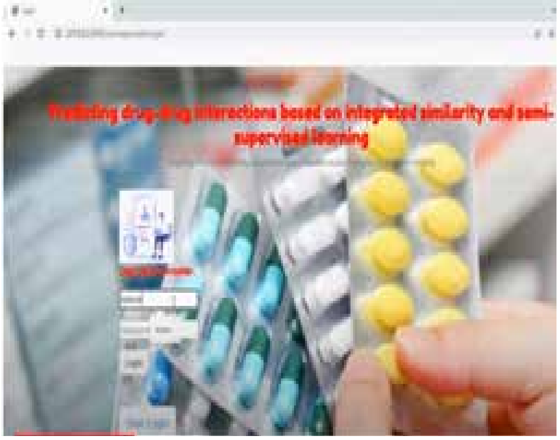
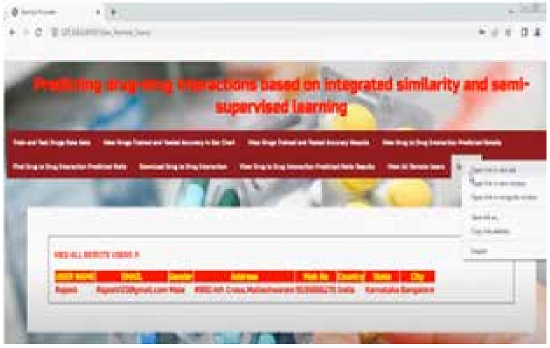

Predicting drug-drug interactions based on integrated similarity and semi-supervised learning

Fig.2. Admin page


Predicting drug-drug interactions based on integrated similarity and semi-supervised learning

VIEW ALL REMOTE USERS !!!

USER NAME	EMAIL	Gender	Address	Mob No	Country	State	City
Rajesh	Rajesh123@gmail.com	Male	#892 4th Cross,Malleshwaram	[illegible]	India	Karnataka	Bangalore

Fig. 3. User details

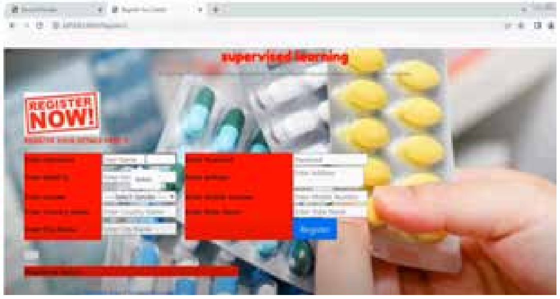
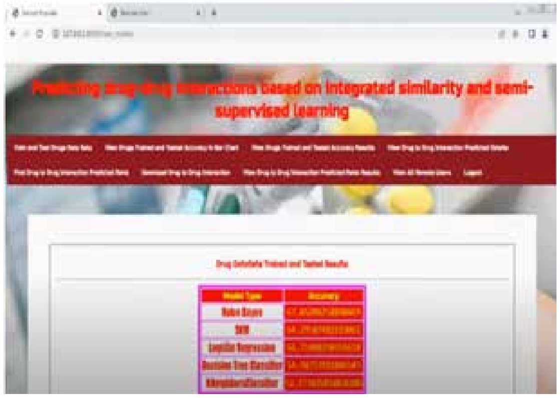


Fig.4. New user registration

Predicting drug-drug interactions based on integrated similarity and semi-supervised learning
Sign In
SERVICE PROVIDER
REGISTER

Fig.5. Sign in details


Predicting drug-drug interactions based on integrated similarity and semi-supervised learning

Drug DataSets Trained and Tested Results

Model Type	Accuracy
Naive Bayes	[illegible]
SVM	[illegible]
Logistic Regression	[illegible]
Decision Tree Classifier	[illegible]
KNeighborsClassifier	[illegible]

Fig.6. Output graphs.

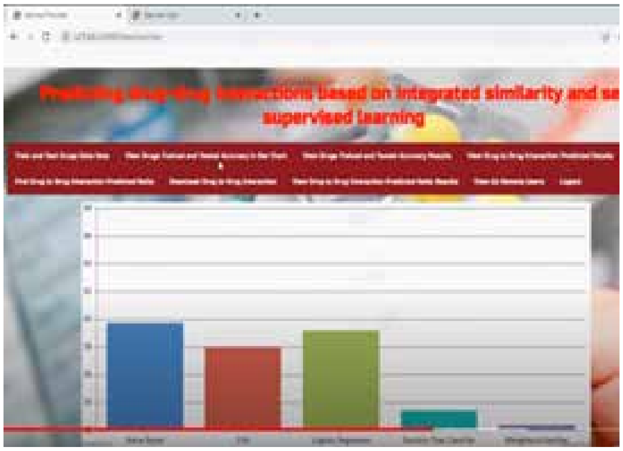
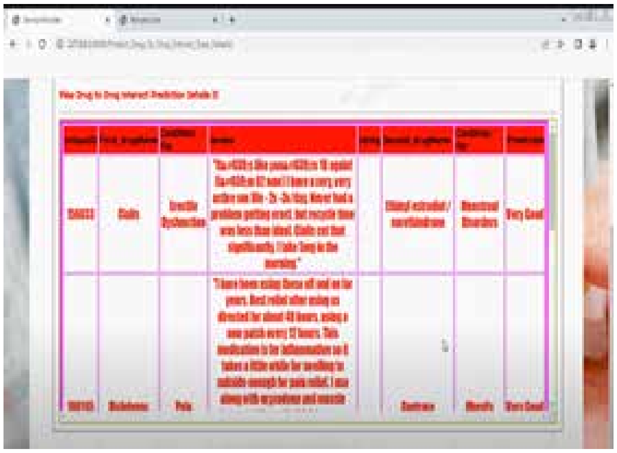


Fig. 7. Output graphs



[illegible]	First_drugName	Condition [illegible]	review	rating	Second_drugName	Condition [illegible]	Prediction
[illegible]	[illegible]	Erectile Dysfunction	[illegible]		Ethinyl estradiol / norethindrone	Menstrual Disorders	Very Good
[illegible]	Diclofenac	Pain	[illegible]		[illegible]	[illegible]	Very Good

Fig. 8. Quality indication

CONCLUSION

In order to increase the treatment's efficacy and lessen patients' suffering, multi-drug therapies have become more popular, especially for complicated diseases like cancer. However, side effects from multi-drug treatments have also been noted, which might lead to serious health issues or even death. Consequently, minimising the issue of medicine breakthroughs and contributing to improved therapy of diseases are both helped by discovering drug-drug communications. Particularly pressing is the need to create novel computational approaches to DDI determination. A novel computational method for inferring DDIs is proposed in this article (DDI- IS SL). Data on the chemical, organic, and phenotype properties of drugs are all part of DDI-IS-SL. Drugs' chemical bases are stored in the PubChem database, which uses 2D binary fingerprints (0 and 1) as its basis. Medications' organic properties include their target communications, enzymes, transporters, and routes. Medications, their side effects, and the negative consequences of medication withdrawal are all part of the phenotype data of pharmaceuticals. A high-dimensional binary feature vector is built using this data for every single medication. Subsequently, we ascertain the cosine action's attribute similarity to pharmaceuticals. Furthermore, we determine the GIP similarity of medications using recognised DDIs. The meaning of the similarity between pharmaceutical attributes and medication GIP is used to get the final medicine similarity. A semi-supervised learning version (RLS) is then used to calculate the likelihood scores of medication pairings. When compared to competing methods, DDI-IS-SL achieves significantly greater prediction efficiency in the 5-fold cross recognition and 10-fold cross validation.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "PREDICTING DRUG-DRUG INTERACTIONS BASED ON INTEGRATED SIMILARITY AND SEMI-SUPERVISED LEARNING", which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. D. Quinn and R. Day, "Drug interactions of clinical importance," Drug safety, vol. 12, no. 6, pp. 393–452, 1995.

2. T. Prueksaritanont, X. Chu, C. Gibson, D. Cui, K. L. Yee, J. Ballard,T. Cabalu, and J.Hochman, "Drug–drug interaction studies: Regulatory guidance and an industry perspective," The AAPS journal,m vol. 15, no. 3, pp. 629–645, 2013.

3. H. Kusuhara, "How far should we go? perspective of drug- drug interaction studies in drug development," Drug metabolism and pharmacokinetics, vol. 29, no. 3, pp. 227– 228, 2014.

4. N. R. Crowther, A. M. Holbrook, R. Kenwright, and M. Kenwright, “Drug interactions among commonly used medications. Chart simplifies data from critical literature review.” Canadian Family Physician, vol. 43, p. 1972, 1997.

5. R. Nahta, M.-C. Hung, and F. J. Esteva, “The her-2- targeting antibodies trastuzumab and pertuzumab synergistically inhibit the survival of breast cancer cells,” Cancer research, vol .64, no. 7, pp. 2343–2346, 2004.

6. T.-C. Chou, “Drug combination studies and their synergyquantification using the chou-talalay method,” Cancer research, vol. 70, no. 2, pp. 440–446, 2010.

7. K. Venkata Krishnan, L. L. von Moltke, R. Obach, and D. J. Greenblatt, “Drug metabolism and drug interactions: application and clinical value of in vitro models,” Current drug metabolism, vol. 4, no. 5, pp. 423–459, 2003.

8. P. J. Neuvonen, M. Niemi, and J. T. Backman, “Drug interactions with lipid-lowering drugs: mechanisms and clinical relevance,” Clinical Pharmacology & Therapeutics, vol. 80, no. 6, pp. 565–581, 2006.

9. Y. B¨ ottiger, K. Laine, M. L. Andersson, T. Korhonen, B. Molin, M.-L. Ovesj¨ o, T. Tirkkonen, A. Rane, L. L. Gustafsson, and B. Eiermann, “Sfinxła drug-drug interaction database designed for clinical decision support systems,” European journal of clinical pharmacology, vol. 65, no. 6, pp. 627–633, 2009.

10. M. P. Pai, D. M. Graci, and G. W. Amsden, “Macrolide drug interactions: an update,” Annals of Pharmacotherapy,vol. 34, no. 4, pp. 495–513, 2000.

11. J. Kuhlmann andW.M¨ uck, “Clinical-pharmacological strategies to assess drug interaction potential during drugdevelopment,” Drug safety, vol. 24, no. 10, pp. 715– 725, 2001.

12. S. Preskorn and S. Werder, “Detrimental antidepressant drug– drug interactions: Are they clinically relevant?” Neuropsychopharmacology, vol. 31, no. 8, pp. 1605–1612, 2006.

13. D. Sridhar, S. Fakhraei, and L. Getoor, “A probabilistic approach for collective similarity-based drug–drug interaction prediction,” Bioinformatics, vol. 32, no. 20, pp. 3175–3182, 2016.

14. S. Ekins and S. A. Wrighton, “Application of in silico approaches to predicting drug–drug interactions,” Journal of pharmacological and toxicological methods, vol. 45, no. 1, pp.65–69, 2001.

15. G. Jin, H. Zhao, X. Zhou, and S. T. Wong, “An enhanced petri-net model to predict synergistic effects of pairwise drug combinations from gene microarray data,” Bioinformatics, vol. 27, no. 13, pp. i310–i316, 2011.

Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms

Edula Saketha, Gampa Bhavitha,
Boddu Bharath Kumar
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

Raheem Unnisa
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ raheemaswcet.17@gmail.com

ABSTRACT

There is still a significant risk that people and financial institutions throughout the globe are vulnerable to bank card fraud. Thanks to the explosion of high-tech tools in recent years, con artists have been able to develop more complex schemes to carry out their illicit business dealings. This study finds a way to tackle this always changing problem by using deep learning algorithms and current artificial intelligence to the problem of credit card theft. In order to analyse various detection strategies, this research makes use of a comprehensive dataset that includes both legitimate and fraudulent credit card transactions. We evaluate their efficacy in spotting fraudulent actions using a diverse array of AI and deep learning models, including, but not limited to, Random Woodland, Support Vector Machines, Slope Boosting, and Convolutional Neural Networks (CNNs). Compared to more traditional device detection algorithms, our experimental results show that deep understanding techniques, and CNNs in particular, achieve better accuracy and fraud discovery rates. There are trade-offs between version complexity and performance, which we also consider when looking at the inter portability of different versions. To optimise the efficiency of the formulae, this study investigates the significance of attribute engineering, dimensional reduction, and hyper criterion tuning. In addition, we find predefined approaches, like stacking and boosting, to use the strength of many designs and improve overall fraud detection capabilities.

KEYWORDS: *CNN, DL,ML, Fraud detection, High efficiency.*

INTRODUCTION

The way we conduct monetary transactions has been transformed by the widespread use of electronic payment systems and the pervasive usage of bank cards for both online and offline purchases. Although there are many positive aspects to this ease, it has also exposed banks and their consumers to a growing threat: bank card fraud [1]. Credit card fraud is still a major issue that costs a lot of money every year and has a negative impact on the economy [2]. Traditional rule-based systems for fraud detection have failed miserably in response to the dynamic nature of fraudsters' methods. Because they are based on previously established criteria and thresholds, these systems are ill-equipped to identify complex and unique forms of fraud [3]. A promising new approach to overcoming this challenge is the integration of AI with deep understanding formulae. These advanced algorithms can learn from data and adjust their approach accordingly, making them a proactive and dynamic tool in the fight against bank card fraud. This project aims to investigate the potential use of cutting-edge AI and deep learning algorithms for detecting bank card fraud [4]. We want to develop more accurate and reliable versions of scams detection that can spot fraudulent trades in real-time by using the power of artificial intelligence.

Several important questions are sought to be answered by the research study:

Calculus Proficiency: What are the differences and similarities between deep-finding algorithms and sophisticated machine learning algorithms when it comes to detecting bank card fraud? The capacity of deep learning models to capture intricate patterns makes them superior to more traditional methods of equipment discovery [5].

The interoperability of designs: Although deep learning models have proven effective in many domains, it is sometimes difficult to understand how they make decisions since they are seen as black-box models [6]. When it comes to detecting frauds, how can we strike a compromise between the need for accuracy and the need of version analysis capabilities?

Personality and Attribute Enhancement: How might hyper parameter tuning, feature design, and dimensional reduction improve the efficacy of fraud detection algorithms? How can we make these models work best when released to the public?

Develop Plans: Would it be possible to use ensemble tactics like stacking and improving to improve bank card fraudulence detection systems by combining their strengths?

By thoroughly analysing these questions, our study seeks to provide financial institutions, companies, and the larger community with insights into developing long-lasting and dependable methods for detecting credit card fraud. Using cutting-edge machine learning and deep learning techniques, our goal is to protect customers in an increasingly digitised financial world from the financial and reputational harm caused by credit card theft [8].

SURVEY OF RESEARCH

In regards to the investigation of credit card fraud using state-of-the-art algorithms for deep learning and machine learning:

Title: Data and Technique Study on Strategies for Detecting Charge Card Scams Observation Perspective Thanks to Aditya Dharmadhikari, Samyak Shah, and Vanshika Bhardwaj for your great work! The year 2021 (2021 With an emphasis on data-centric and technique-centric aspects, the study provides a comprehensive review of ways for detecting bank card fraud. Methods combining machine learning and deep discovery are covered as well as more traditional techniques [9].

Methods for the Discovery of Charge Card Fraud: A Study Mohammed Qahtan Alqahtani and colleagues wrote. Month: 2019 Various strategies for detecting credit card fraud, such as rule-based systems, statistical methodologies, and AI formulae, are explored in this review. It describes the benefits and drawbacks of each approach and stresses the need of cutting-edge designs like deep understanding [10].

Deep Learning for Identifying Credit Card Scams: A Critical Review (Chengyu Qiang et al., 2018). Year: 2020 clarify This study provides a thorough examination of deep learning designs, structures, and efficiency compared to traditional methodologies, with a focus on their use in charge card fraudulence detection. Also covered are the challenges and directions for future research in this area.

Investigation on Payment System Fraud Detection Written by Shubham Atal and colleagues. The research, which takes place in 2019, examines a variety of methods for detecting settlement fraud, including bank card fraud. Analysis of anomaly detection in fraud detection systems, data-driven methodologies, and machine learning architectures are all part of what it looks at.

Deep Knowledge for Detecting Credit Card Fraud: A Comparative Study Lau, Hongyu, et al. Research compares deep learning algorithms for charge card fraud detection in the year 2020. As such, it assesses the practicality and efficacy of various semantic network architectures.

Credit Card Fraud Discovery Methods: A Synopsis from an Informational and Methodological Perspective A. K. Sharma and colleagues wrote it. The year 2020 This paper provides a comprehensive overview of approaches for detecting credit card fraud, with an emphasis on data preprocessing, feature selection, and machine learning formulae. Insights on the challenges and opportunities in this field are provided.

Deep Learning for the Determination of Credit Card Fraud: A Survey Written by: NhatHai Phan and colleagues. This study delves into the use of deep learning in detecting credit card fraud in the year 2019. The paper discusses the development, advantages, and limitations of deep finding versions in handling complicated fraudulence detection tasks.

Together, these literature reviews include a wealth of information on cutting-edge methods for detecting credit card frauds, such as those that use deep learning formulae and artificial intelligence, and they shed light on important topics for researchers and professionals in this field.

EXISTING SYSTEM

ML is quite modular, with multiple sub fields that each handle certain discovery tasks. However, ML discovery has several types of structures. A solution for CCF, like random forest (RF), is provided by the ML technique. The random forest sets the choice tree. A lot of scientists use the RF method. By combining RF with network analysis, we can integrate the model. The acronym APATE describes this method. Scientists have access to a variety of ML techniques, including supervised and unsupervised learning. Common machine learning algorithms used for CCF identification include LR, ANN, DT, SVM, and NB. The researcher may build robust discovery classifiers by integrating these tactics with set methodologies. An artificial semantic network is a network that has been constructed by linking several nerve cells and nodes. There are several layers that make up a feed-forward perception multilayer, including input, output, and hidden layers. The input nodes for the exploratory variables are shown in the first layer. These input layers are multiplied with an exact weight, and then each hidden layer node is transferred with a certain bias, and the sum is taken.

PROPOSED SYSTEM

Worldwide, banks and consumers are still very much at risk from credit card frauds. Scammers have been able to develop more sophisticated ways of carrying out fraudulent transactions due to the development of technological technology. This research delves at the use of cutting-edge AI and deep learning algorithms for credit card fraud detection, in an effort to tackle this always changing dilemma. In order to test various discovery strategies, this study makes use of a large dataset that contains both legitimate and fraudulent charge card transactions. We test the efficacy of several AI and deep learning models in detecting fraudulent tasks. These models include, but are not limited to, Random Woodland, Assistance Vector Machine, Gradient Boosting, and Convolutional Neural Networks (CNNs).

WORKING METHODOLOGY

In today's digital economy, credit card theft is becoming an increasingly big problem for consumers and financial institutions alike. Constant innovation and use of new, sophisticated strategies by fraudsters to compromise payment systems causes huge losses and damages people's faith in online purchases. When it comes to spotting complex and unique forms of fraud, traditional rule-based solutions fall short. This highlights the critical need of quickly creating reliable and flexible fraud detection systems.

An issue that needs solving is the development and deployment of a system for the identification of credit card fraud that makes use of cutting-edge algorithms in machine learning and deep learning. The following important problems are what this system is trying to solve:

Flexibility and Detection in Real-Time: Due to the massive volume of credit card transactions, real-time fraud detection is essential for avoiding fraudulent charges. The system must be able to process a large number of transactions quickly and accurately without causing any noticeable delays.

The capacity to change and adjust one's strategy in response to changing fraud patterns is a hallmark of modern fraudsters. In order to detect previously unseen fraudulent behaviours, the detection system has to be able to learn and adapt to new patterns of fraud.

Due to the rarity of credit card fraud in comparison to actual purchases, databases including this information are skewed. Accurate fraud detection and non-biased model performance are dependent on the system fixing class imbalance problems.

Model Interpretability: Although deep learning models are frequently regarded as black-box models, they are capable of producing great prediction performance. To ensure that financial institutions can comprehend and rely on the system's judgements, it is important that the system finds a happy medium between model accuracy and inter pretability.

Feature Engineering and Data Preprocessing: The success of a model relies heavily on the accuracy and timeliness of its features. Methods for reducing noise and extracting useful information from transaction data should be part of the system.

Algorithms for detecting credit card fraud should be fine-tuned using optimisation and hyper parameter tuning techniques for the system to attain optimal performance.

Techniques for Ensembles: The use of ensemble approaches to combine the strengths of several deep learning and machine learning models has the potential to improve fraud detection skills in general. An efficient investigation of ensemble approaches should be carried out by the system.

Implementing a scalable and cost-effective fraud detection system into financial institutions' infrastructure is crucial for keeping up with the constant flow of transactions.

By resolving these issues, we can create a credit card fraud detection system that protects customers and financial institutions in an ever-changing digital financial environment against fraudulent transactions in a way that is accurate, quick, and adaptable.

IMPLEMENTATION

There are a number of steps involved in implementing a Credit Card Fraud Detection system that makes use of cutting-edge ML and DL algorithms. Here is a brief rundown of the application process:

Collecting and Preprocessing Data

Get a database with all the credit card transactions that have ever happened. It must include both legitimate and fraudulent transactions. Take care of missing values, standardise characteristics, and deal with course discrepancies by either oversampling the minority class or under sampling the bulk class during pre-processing.

Engineering with Attributes

Make useful features out of the transaction data, like the amount of the deal, the time of day, the specifics of the merchant, and the cardholder's history. To minimise the amount of features while maintaining crucial information, you may want to look into dimensional reduce methodologies like Principal Element Evaluation (PCA).

Partition the dataset into three distinct sets: one for training, one for validation, and one for testing. The designs are educated using the training collection, hyper parameters are adjusted from the validation set, and final assessment is conducted using the examination set.

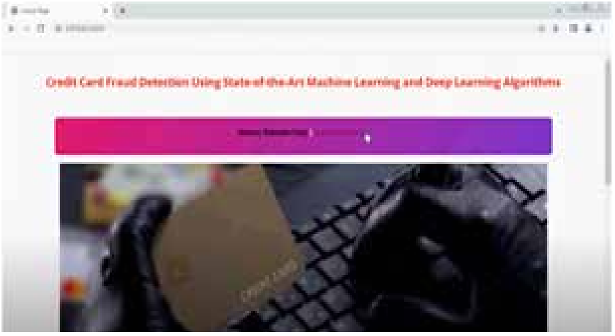
Choice of Version: Pick the Best ML and DL Algorithms for Detecting Credit Card Scams. Slope Boosting, Random Woodland, and Assistance Vector Device are common ML algorithms. It is also possible to think about DL algorithms like RNNs and CNNs.

Algorithm Improvement and Training

Using the training data and appropriate hyper parameters, train the selected models. Find the best values for the hyper parameters by improving the models using techniques like grid search or random search. Put in place measures for early halting to ensure appropriate stopping.

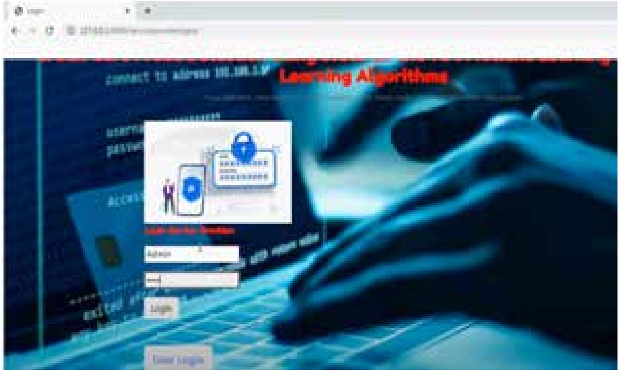
Implementing a system to identify charge card fraud is an ongoing process that requires constant vigilance and adaption to new threats. Maintaining efficacy in fighting fraudulence requires regular upgrades and modifications to the software and system architecture.

Fig. 1. Home page



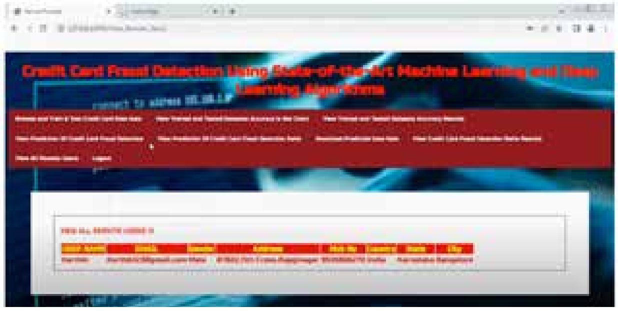
Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms

Fig. 2. Login page



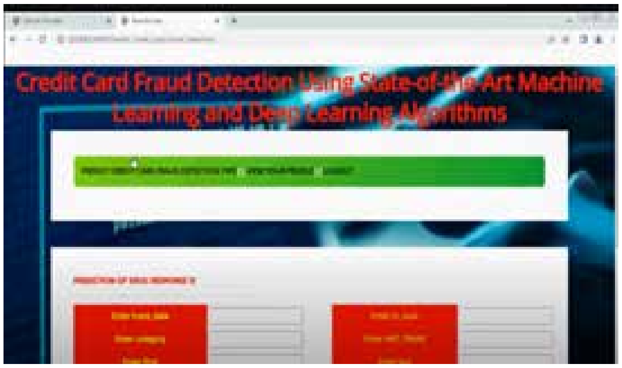
Learning Algorithms

Fig. 3. Admin login page



Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms

Fig. 4. User details



Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms

Fig. 5. Prediction page.

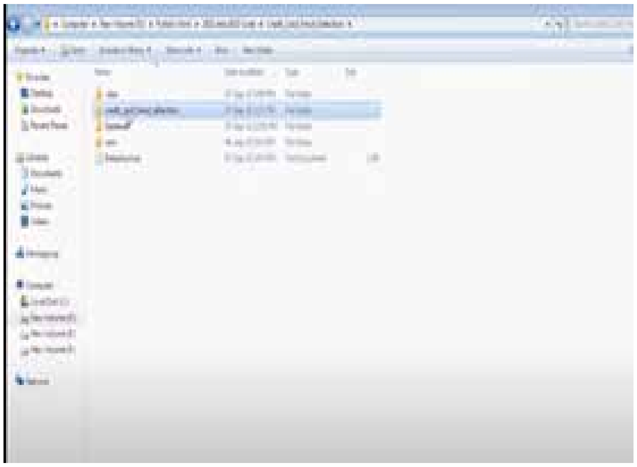
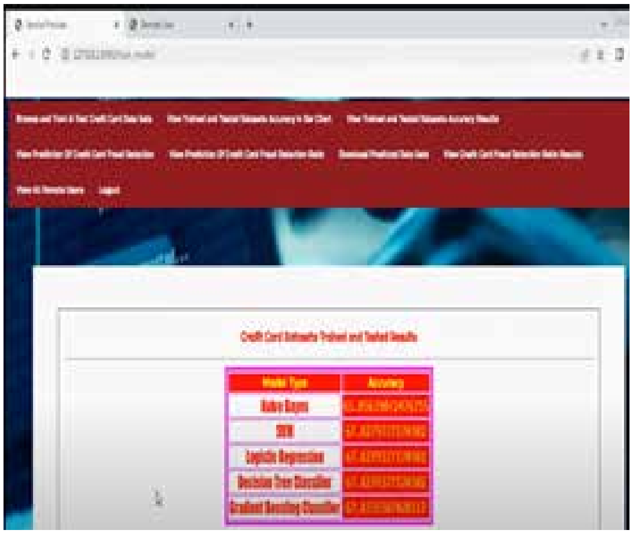


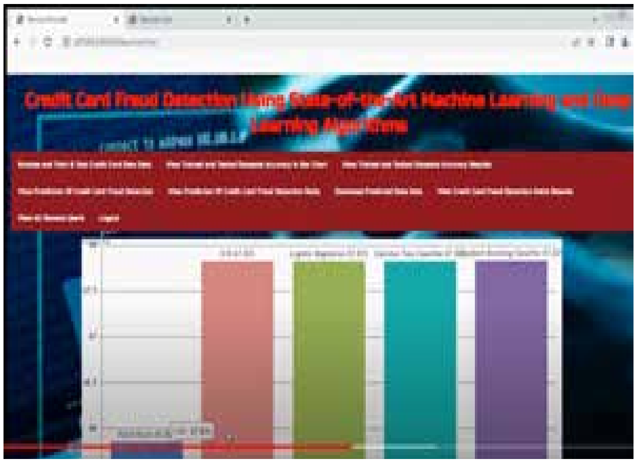
Fig. 6. Dataset upload page.



Credit Card Datasets Trained and Tested Results

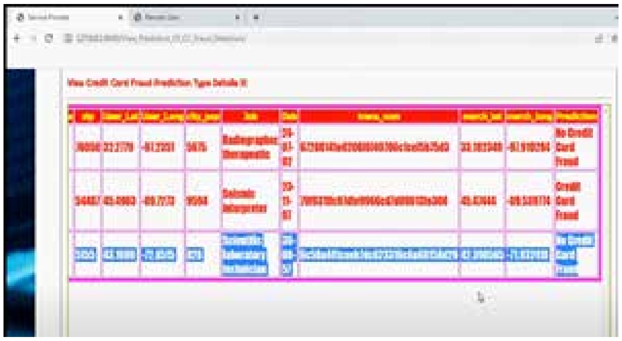
Model Type	Accuracy
Naive Bayes	[illegible]
SVM	[illegible]
Logistic Regression	[illegible]
Decision Tree Classifier	[illegible]
Gradient Boosting Classifier	[illegible]

Fig. 7. Algorithms applied.



Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms

Fig. 8. Prediction Graphs.



View Credit Card Fraud Prediction Type Details !!!

[illegible]	User_Lat	User_Long	city_pop	Job	Date	trans_num	merch_lat	merch_long	Prediction
[illegible]	22.2779	-81.2331	5675	Radiographer, therapeutic	[illegible]	[illegible]	[illegible]	[illegible]	No Credit Card Fraud
[illegible]	45.0983	-89.7773	9568	Seismic interpreter	[illegible]	[illegible]	[illegible]	[illegible]	Credit Card Fraud
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	No Credit Card Fraud

Fig. 9. Final output

CONCLUSION

Ultimately, protecting the financial interests of both institutions and consumers may be achieved by the application of sophisticated Artificial Intelligence (ML) and Deep Knowing (DL) formulae to the identification of bank card fraud. According to the results of this study, these cutting-edge methods use a robust defence against the ever-evolving tactics used by con artists. After much trial and error, it has been shown that DL architectures, and CNNs in particular, have great promise for improving the accuracy of fraud detection. Finding a happy medium between model inter pretability and efficiency is still an important factor to think about.

To maximise the formulae for real-world application, this research has also shown how important it is to do things like data pre treatment, function design, and hyper parameter tweaking. When combining the robustness of several models, set approaches have shown to be effective. Implementing these methods in real-time systems has the ability to significantly reduce credit card fraud as the digital economy evolves. While guaranteeing compliance with data personal privacy standards, these systems can adapt to new risks, handle large transaction numbers, and provide practical insights. In a nutshell, this research contributes useful insights and practical recommendations for the development of long-term, adaptable systems to identify credit card fraud. We are committed to improving financial safety and increasing trust in electronic settlement systems, and we are pursuing these sophisticated ways to do just that.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled “CREDIT CARD FRAUD DETECTION USING STATE-OF-THE-ART MACHINE LEARNING AND DEEP LEARNING ALGORITHMS”, which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. Ribeiro, A. H., Santos, C. H., & Papa, J. P. (2019). Credit card fraud detection: a realistic modeling and a novel learning strategy. Expert Systems with Applications, 135, 281-298.
2. Dal Pozzolo, A., Boracchi, G., Caelen, O., & Bontempi, G. (2015). Credit card fraud detection: a realistic modeling and a novel learning strategy. IEEE transactions on neural networks and learning systems, 29(8), 3784-3797.
3. Zheng, Y., Yang, S., & Xie, J. (2014). Credit card fraud detection using Bayesian and neural networks. Expert Systems with Applications, 41(4), 4915-4924.
4. Phua, C., Lee, V., Smith, K., & Gayler, R. (2005). A comprehensive survey of data mining-based fraud detection research. arXiv preprint cs/0506067.
5. López-Rojas, E., Axelsson, S., & Niklasson, L. (2015). A study of the effect of imbalanced training data on convolutional neural networks for credit card fraud detection. Journal of computational science, 16, 171-178.
6. Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O’Reilly Media.
7. Chollet, F. (2017). Deep Learning with Python. Manning Publications.
8. Raschka, S., & Mirjalili, V. (2017). Python Machine Learning. Packt Publishing Ltd.
9. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
10. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.
11. Schapire, R. E. (1999). A brief introduction to boosting. In Proceedings of the sixteenth international joint conference on artificial intelligence (Vol. 2, pp. 1401-1406).
12. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

A Novel Approach for Credit Card Fraud Detection using Decision Tree and Random Forest Algorithms

Burre Vedika, Allenky Harshitha, Boda Ashok
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

Jonnadula Narasimharao
Associate Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ jonnadula.narasimharao@gmail.com

ABSTRACT

Around the world, new methods of organising came into view as a result of the proliferation of contemporary technologies. The use of credit cards is one example. However, a lot of problems arise in this system when it comes to charge card scams since there are a lot of holes in it. This has resulted in a significant loss for both the business and consumers who use bank cards. Due to concerns about individual privacy, there is a dearth of investigative training focusing on the examination of working credit card numbers in default. An effort to identify credit card fraud using formulae that use machine learning techniques is presented in the book. Credit card fraud detection utilising a Choice Tree and random forest are the two formulas employed here. By taking a look at some publicly available data, we can determine how well the design worked. After that, a real-life bank card facts team is examined. In addition, the data samples are enhanced with additional noise to further verify the systems' robustness. The first approach builds a tree against the individual's actions, and rip-offs will definitely be believed using this tree, which is relevant to the techniques used in the study. Second, we'll try to identify the culprit by making use of a user-task based forest that has already been constructed. The findings of the research prove beyond a reasonable doubt that the popular alternative method detects bank card fraud with sufficient accuracy.

KEYWORDS: *Secure data, CNN, Feedback, Fraud detection.*

INTRODUCTION

There is a daily increase in online buying. The usage of charge cards for online purchases of products and services is becoming more common. physical card, as opposed to digital cards, which are used for transactions conducted offline [1]. When paying using a physical card, the buyer actually hands over the card to the vendor. An assailant needs the credit card in order to make fraudulent purchases in this transaction [2]. It might cost the credit card company a lot of money if the cardholder doesn't notice they lost their card. In order to commit fraud in an online payment situation, all that is needed is a little amount of sensitive information (such as a protected code, card number, expiration date, etc.) [3]. Online or over the phone transactions will constitute the bulk of this acquisition strategy. A criminal needs just knowledge of the card details to perpetrate fraud in these types of purchases. The real cardholder usually has no idea that someone else has viewed or swiped his card details [4]. To uncover this kind of fraud, one must compare the investment habits of each card and spot any deviation from the "common" investing trends [7]. An attractive approach to lowering the cost of effective charge card fraud is scams identification based on the examination of current cardholder purchase data [5]. Due to the fact that people exhibit predictable behaviourist accounts, A collection of patterns that incorporate details like the average buying group, the amount of money spent, the duration since the last purchase, and other information can be used to represent each cardholder. A divergence from these patterns could jeopardize the system [6].

LITERATURE SURVEY

Canadian Bank Card Fraud Detection with the Application of Predictive Analytics and Existing Technology

This research paper examines a scorecard that is based on pertinent evaluation criteria, features, and capabilities of predictive analytics vendor solutions used to identify charge card frauds. Presented below is a comparative analysis of five Canadian credit card predictive analytics provider treatments. Presently, researchers are trying to compile a comprehensive inventory of risks, limitations, and hurdles associated with credit card fraud rub vendor remedies.

BLAST-SSAHA Combination for the Identification of Bank Card Frauds

Arun K. Majumdar, Shamik Sural, Amlan Kundu, and Suvasini Panigrahi are all members of the IEEE.

In this study, we suggest a two-step sequence structure. An account analyst first determines if the incoming purchase history of a particular bank card corresponds with the cardholder's past investment behavior. A variance analyzer (DA) examines the anomalous transactions once the account analyzer has mapped them out to see if they align with any dishonest patterns of behavior. The final decision regarding the type of purchase is based on the conclusions of these two analysts. To achieve online action time for PA and DA, we provide a novel way to combine the two sequence alignment techniques, BLAST and SSAHA.

A Study on a Model for Detecting Charge Card Scams Based on Distance Amount

Na Wang and Wen-Fang Yuan.

Charge card frauds are on the rise in China, paralleling the country's increasing use of credit cards and the number of people working in the industry. Bank risk management ultimately focuses on how to better identify and prevent credit card frauds. Using outlier mining for credit card fraud detection, this article proposes a model that takes into account the rarity and unusualness of fraudulent charges in purchase data and uses outlier detection based on distance sum. Based on the results of the experiments, this version is both practical and effective in detecting credit card frauds.

Using Support Vector Machines and Decision Trees to Detect Fraud in Bank Card Systems

Worldwide, frauds are proliferating in response to the expanding frontiers of online commerce, causing victims to lose substantial sums of money. Bank card fraud is now the main source of financial losses; it affects both individual traders and consumers. The approaches that are discussed here are used to identify bank card fraud. They include decision trees, genetic formulae, a meta-learning methodology, semantic networks, and HMM. The problem of deceit detection is being tackled by considering systems that use the expert system principles of Support Vector Machine (SVM) and decision trees. Economic losses may be reduced to a greater degree by using this hybrid technique.

Surveillance of Machine Learning for the Determination of Bank Card Scams

We propose an SVM based method with heavy bit involvement in this thesis; this method goes beyond only investment accounts and incorporates a wide range of customer profile aspects. In addition to reducing the FP and FN costs, the simulation results demonstrate an improvement in the TP and TN rates, which stand for real positive and real unfavourable, respectively.

EXISTING SYSTEM

The model changed into trained the use of simply the two characteristics that had the highest variance, in line with the ok-way algorithm implementation. The model is configured with two clusters: one for non-fraud and one for fraud. We also attempted an expansion of hyper parameter settings, however they have been all pretty ineffective. The consequences have been, in addition, unaffected with the aid of lowering the records' dimensionless (i.e., making it less complex than 2 dimensions).

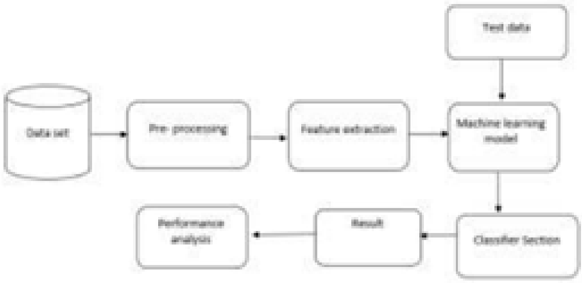


Fig.1. Architecture model

PROPOSED SYSTEM

We need to apply a Kaggle dataset to train a machine learning version which can as it should be perceive times of credit card robbery. We decided to apply a logistic regression version after doing initial records exploration since it constantly produces the best satisfactory accuracy reports. Logistic regression, as it excels at growing binary classifications. The challenge become implemented using the Python learns library. Kaggle datasets had been utilised for credit card fraud detection. Data was classified as either "no fraud" or "fraud" the use of pandas to create a records body. Matplotlib became used to plot the fraud and non-fraud statistics. Train_test_split became used for information extraction, dividing arrays or matrices into random teach and test subsets. A gadget mastering set of rules called Logistic Regression changed into used for fraud detection, and the ensuing prediction score becomes revealed consistent with the set of rules's predictions. The ultimate step changed into to plotting the Confusion matrix on each of the real and foretasted records units.

Credit card fraud records were sourced from the Kaggle website and uploaded as a dataset in this module.

To prepare the obtained statistics for generation of the educating and test models, we need to get rid of null values, pointless rows, and columns. The next step is to divide the statistics in half of: 80% for education and 20% for checking out.

To collect correct effects, we need to teach the RF set of rules the use of schooling records after which check it with take a look at data. This is known as going for walks the Random Forest set of rules.

With the assistance of random wooded area, we are able to perceive the telltale symptoms of credit card fraud in our check statistics.

We can display the easy transactions and the fraudulent ones with the usage of a graph.

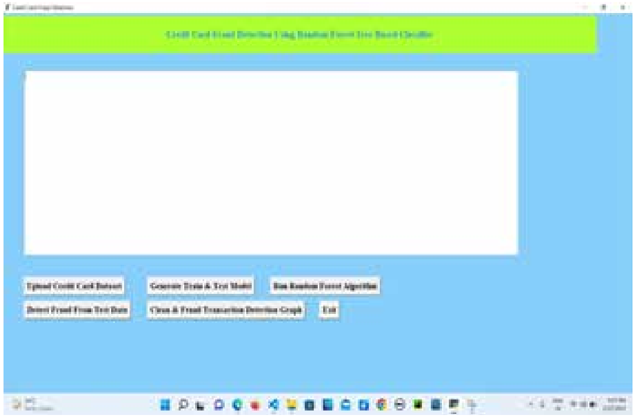
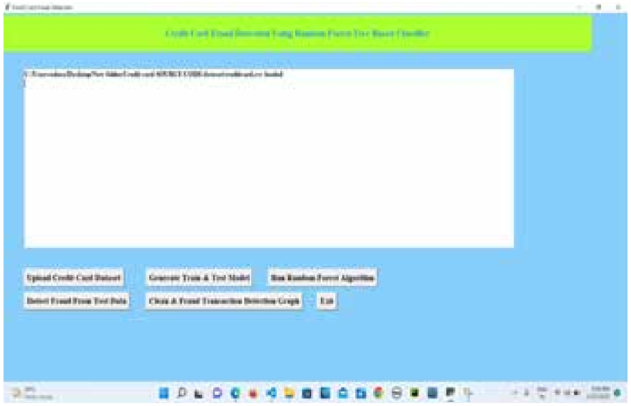

Credit Card Fraud Detection Using Random Forest Tree Based Classifier
Upload Credit Card Dataset
Generate Train & Test Model
Run Random Forest Algorithm
Detect Fraud From Test Data
Clean & Fraud Transaction Detection Graph
Exit

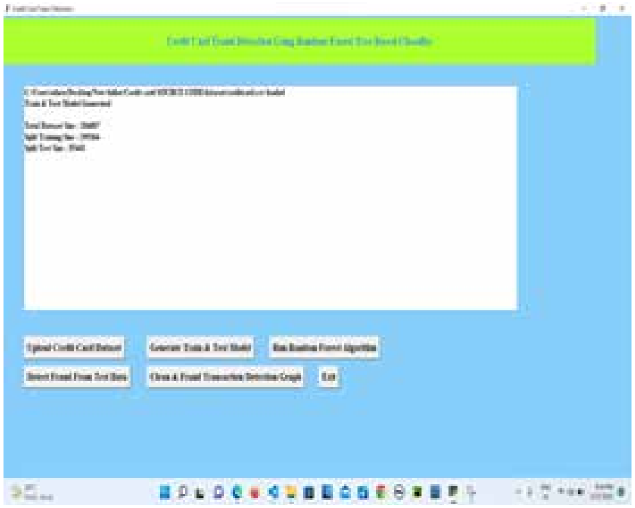
Fig 2 Home page


Credit Card Fraud Detection Using Random Forest Tree Based Classifier
Upload Credit Card Dataset
Generate Train & Test Model
Run Random Forest Algorithm
Detect Fraud From Test Data
Clean & Fraud Transaction Detection Graph
Exit

[illegible] loaded

Fig. 3. Upload credit card data set

Our credit score card fraud records came from the Kaggle platform, which you may get admission to by importing our facts set.


Credit Card Fraud Detection Using Random Forest Tree Based Classifier
Upload Credit Card Dataset
Generate Train & Test Model
Run Random Forest Algorithm
Detect Fraud From Test Data
Clean & Fraud Transaction Detection Graph
Exit

[illegible] loaded
Train & Test Model Generated

Total Dataset Size : [illegible]
Split Training Size : [illegible]
Split Test Size : [illegible]

Fig. 4. Generate Train & Test Model

Generate train & test model: The gathered data needs to be pre-processed in order to remove null values, unnecessary rows, and unnecessary columns. The data must then be divided into two sections: an 80% training portion and a 20% testing portion.

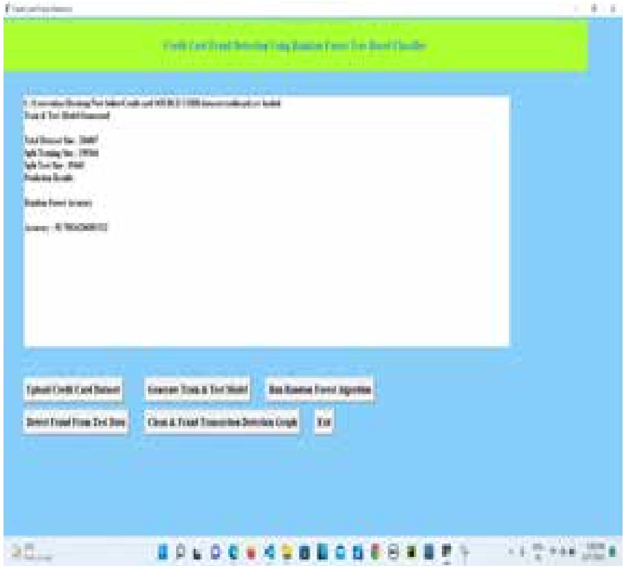

Credit Card Fraud Detection Using Random Forest Tree Based Classifier
Upload Credit Card Dataset
Generate Train & Test Model
Run Random Forest Algorithm
Detect Fraud From Test Data
Clean & Fraud Transaction Detection Graph
Exit

Fig 5 Run Random Forest Algorithm

Run Random Forest Algorithm: In order to achieve accuracy, we need to train the RF algorithm using training data and then test it using test data.

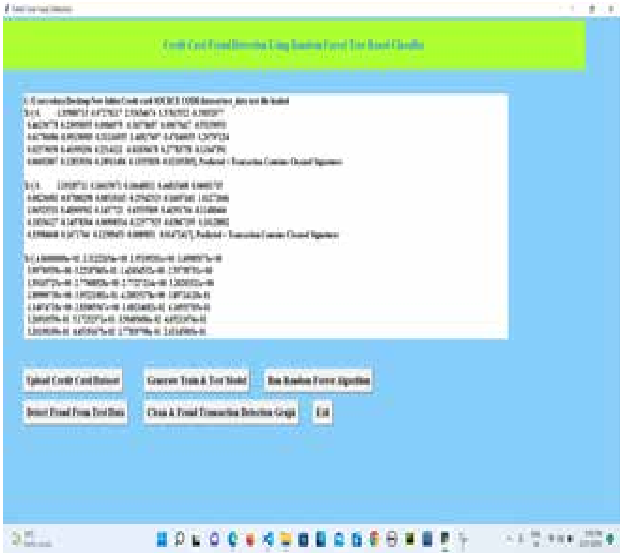

Credit Card Fraud Detection Using Random Forest Tree Based Classifier
Upload Credit Card Dataset
Generate Train & Test Model
Run Random Forest Algorithm
Detect Fraud From Test Data
Clean & Fraud Transaction Detection Graph
Exit

Fig 6 Detect Fraud from Test Data

Detect Fraud from Test Data:The signs of fraud may be detected by using random forest.

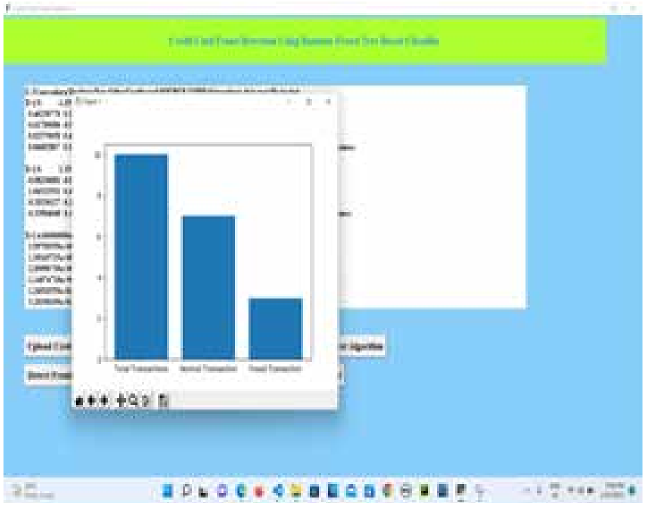

Credit Card Fraud Detection Using Random Forest Tree Based Classifier
Total Transactions
Normal Transaction
Fraud Transaction

Fig 5 clean and fraud graph

CONCLUSION

While a bigger amount of training data would undoubtedly improve the Random Woodland algorithm's performance, it will almost likely slow down the screening and application processes. It might also be helpful to use extra pre-processing procedures. Even though SVM produces great results, the method may have performed even better with more data pre treatment since it still handles the imbalanced dataset problem and requires further pre-processing.

ENHANCEMENTS TO COME

We want to address these issues in our future endeavours. It is necessary to improve the random forest algorithm. The voting method, for instance, treats all basic classifiers as having the same weight, even if some of them could be more important than others. Consequently, we also strive to improve this method in certain ways.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "A Novel Approach for Credit Card Fraud Detection using Decision Tree and Random Forest Algorithms", which gave us excellent resources and assistance to complete our assignment. I truly appreciate the helpful advice and direction that our Chairman, Director, Deans, Head of the Department, Department of Computer Science and Engineering, Guide, and

Teaching and Non-Teaching Faculty Members have provided for every facet of our job.

REFERENCES

1. Salazar, Addison, et al. “Automatic credit card fraud detection based on non-linear signal processing.” Security Technology (ICCST), 2012 IEEE International Carnahan Conference on. IEEE, 2012.
2. Delamaire, Linda, H. A. H. Abdou, and John Pointon. “Credit card fraud and detection techniques: a review.” Banks and Bank systems 4.2 (2009): 57-68.
3. Quinlan, J. Ross. “Induction of decision trees.” Machine learning 1.1 (1986): 81-106.
4. Quinlan, J. R. (1987). “Simplifying decision trees”. International Journal of Man-Machine Studies. 27 (3): 221. doi:10.1016/S0020-7373(87)80053-6.
5. K. Karimi and H.J. Hamilton (2011), “Generation and Interpretation of Temporal Decision Rules”, International Journal of Computer Information Systems and Industrial Management Applications, Volume 3.
6. Aggarwal, Charu C. “Outlier analysis.” Data mining. Springer International Publishing, 2015.
7. Salazar, Addisson, Gonzalo Safont, and Luis Vergara. “Surrogate techniques for testing fraud detection algorithms in credit card operations.” Security Technology (ICCST), 2014 International Carnahan Conference on. IEEE, 2014.
8. Ogwueleka, Francisca Nonyelum. “Data mining application in credit card fraud detection system.” Journal of Engineering Science and Technology 6.3 (2011): 311-322.

Phishing Website Detection using Machine Learning Algorithms

Bokkena Srihitha, Peddauppari Sai Kiran
Kukka Sunil
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

Raj Kumar Patra
Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ rajkumarpatra.cse@cmrtc.ac.in

ABSTRACT

A phishing attack is the simplest method of stealing sensitive information from unsuspecting consumers. Phishers aim to steal sensitive information, such as login credentials and checking account data. People concerned with cyber security are now trying to find reliable and consistent methods of detecting phishing websites. Through the elimination and evaluation of various functionalities of legitimate and malicious links, this study addresses machine finding technology as it pertains to the detection of phishing links. Phishing websites may be identified using algorithms such as Decision Tree, Assistance vector builder, and random forest. Using light gbm and svm formulas, the article aims to identify phishing links.

KEYWORDS: *URL, SVM, Light GBM, Cyber security, Phishing website.*

INTRODUCTION

The expansion of the net's talents during the last several decades has revolutionised the way we live. Communication, education, enterprise conditioning, and commerce are all heavily reliant on it. The net is a treasure trove of know-how which can aid in non-public, organisational, financial, and societal development [1]. The internet allows us to get right of entry to a wealth of statistics at any time, from any area in the globe, and it also makes it less difficult to offer a huge kind of offerings on-line [2]. The time period "phishing" refers back to the exercise of sending a seemingly legitimate e mail or journeying malicious web sites to be able to trick unsuspecting recipients into divulging sensitive data which includes passwords, social safety numbers, financial institution account details, date of birth, and credit card numbers. Hundreds of thousands of online drug customers all over the world are victims of phishing attacks [3]. Associations and people have suffered huge losses because of phishing tries on their personal and private records. It turns out that detecting the phishing try is no picnic. By changing some characters in the URL to similar Unicode characters, for example, this assault may additionally anticipate a complicated form that fools even the most vigilant customers [5]. A careless implementation, which includes substituting an IP address for a site name, is one of the downsides. Still, some studies [6] have tried to pick out phishing attacks through the use of AI and information mining techniques, with a popular fee of 99.62% is the best done. Due to their complicated processing and high battery use, such structures aren't best for tiny gadgets like smartphones. They want entire HTML pages or at the least HTML hyperlinks, tags, and web pages as entry factors [7]. Some of these systems employ photo processing to accomplish reputation, that's a JavaScript detail. In assessment to competing structures, our reputation algorithm requires only six absolutely extracted residences from the URL as enter, making it a less useful resource in depth in terms of each CPU and reminiscence. Following an outline of the relevant field studies, this article will get into the specifics of the URLs used by our gadget for popularity [8]. Aside from that, we'll go over our popular technique. Then, within the sensible phase, we'll position it via its paces and show you the effects. As a remaining step in countering the phishing assault, we will list the benefits and effects of our device [9] [10].

Project Objectives

Phishers purpose to scour borrow sensitive information together with login credentials and economic account records. Experts in cyber defence at the moment are looking for reliable and regular detection methods for phishing websites. Using machine-gaining knowledge to extract and examine distinctive components of proper and phishing URLs, this exam deals with the identification of phishing URLs. Phishing web sites may be recognized by the use of algorithms which include Decision Tree, Random Forest, and Support Vector Machine. By comparing every set of rules's accuracy rate, fake superb charge, and false terrible price, this observe targets to discover phishing URLs and narrow down the first-rate system studying technique.

LITERATURE SURVEY

A kernel-primarily based method to categorization changed into recommended via Rashmi Karnik et al. We classify phishing in this way. When it comes to identifying malicious and phishing websites, our approach achieves an estimated 95% accuracy.

By comparing Google Safe browsers with a supervised Machine Learning machine, Andrei Butnaru et al. Had been capable of stopping phishing assaults based on revolutionary combined phishing assaults.

One of the simplest strategies for detecting those dangerous works turned into advised via Vahid Shahrivari et al. And is Machine Learning. The motive for this is due to the fact gadget studying algorithms are capable of stumbling on the general public of phishing attacks because of their shared traits. To expect which websites will be phishing, this makes use of a huge, wide variety of classifiers based totally on machine mastering. The ability to build adaptable fashions for specific jobs, like phishing detection, is the key gain of gadget mastering. Machine getting to know fashions can be a effective weapon inside the fight against phishing because it is a categorization venture.

A method for recognizing phishing web sites using the heaping version changed into provided with the aid of Ammara Zamir et al. In order to assess the functions of phishing, one might also utilise feature selection techniques together with records benefit, gain ratio, Relief-F, and recursive feature elimination (RFE). Two characteristics are formed by means of combining the best and worst features. Several device studying strategies, such as neural community [NN] and random forest [RF], use bagging in primary thing evaluation. To enhance classification accuracy, two heaping representations are used: heaping1 (RF + NN + Bagging) and heaping2 (kNN + RF + Bagging).

Deep neural networks (DNNs), convolution neural networks (CNNs), long brief-term reminiscences (LSTMs), and gated recurrent devices (GRUs) were counseled by way of researchers Ali Selamat et al. Of their have a look at on phishing detection. Comprehensive experiments had been carried out to research the impact of parameter adjustment on the overall performance accuracy of the deep learning models with a purpose to examine the behaviour of those architectures. The models show various tiers of accuracy.

A machine mastering-based totally URL identification technique changed into counseled with the aid of Ashit Kumar Dutta. To stumble on a phishing URL, an RNN is used. There are 7,900 malicious sites and five,800 valid ones used for assessment. This method's results are better than the ones of greater cutting-edge techniques.

To become aware of phishing domains, which fluctuate from legitimate ones, Atharva Deshpande et al. Counseled using a combination of system learning algorithms and herbal language processing strategies. In order to discover phishing web sites, Ms. Sophia Shikalgar et al. Advised a hard and fast of strategies and device studying classifiers that use A hybrid device learning strategy uses several classifiers to improve prediction accuracy. Classifiers fluctuate in how they categorise facts and the way they function. Takes use of an unstructured statistics set of URLs that comprises 2,905 URLs.

In an effort to tackle this quandary, Nureni Ayofe Azeez et al. Tried to remedy huge issues. The primary concern is the identification of questionable URLs on social media and the subsequent mitigation of risk for internet users because of such URLs. The AdaBoost, Gradient Boost, random woodland, Linear SVM, decision tree, and Naïve Bayes classifier are the six gadget learning algorithms which are adapted for education the usage of traits retrieved from the social network and for further processing. Data was culled from 532,403 postings in all. Finally, the models will be trained on 87,083 postings. When compared to different strategies, AdaBoost's ninety five% accuracy and 97% precision are the fine. Dear Ademola, Machine getting to know was cautioned through Philip Abidoye and Boniface Kabaso as a way for correctly classifying datasets so that you can stumble on traits of phishing URLs that cybercriminals may additionally make the most.

In order to classify phishing attacks, R. Kiruthiga and D. Akila offered a new method of identifying phishing websites the usage of system gaining knowledge of strategies. Additionally, it gives a way for appropriately detecting phishing e-mail attacks through the usage of gadget studying and natural language processing.

EXISTING SYSTEM

One type of on-line fraud is called "phishing," and it entails the sender of misleading communications posing as a valid enterprise or business enterprise. Upon beginning the attached file or clicking on the provided URL, the recipient's private statistics can be stolen or a virus can be set up on their system. Phishing attacks was once disbursed with the aid of massive junk mail operations that randomly focused big corporations of individuals. The goal became to maximise the number of people inflamed by way of a hyperlink or record. This sort of attack can be detected in a selection of ways. Machine learning is one approach. The consumer will enter URLs right into a gadget studying version, in an effort to then determine whether or not the URL is phishing or now not and document the result. To categorise those URLs, one may use a number of system gaining knowledge of methods, which include assist vector machines, neural networks, selection bushes, random forests, XG raise, and many others. Specifically, the Random Forest and Decision Tree classifiers are addressed inside the counseled approach. Using Random Forest and selection tree classifiers, respectively, the recommended approach correctly recognized 87.Zero% of phishing URLs and 82.4% of valid URLs.

PROPOSED SYSTEM

In recent decades, phishing attacks have become both more numerous and more sophisticated. This has led to corresponding advancements in the methods used to prevent the detection of phishing attacks, which pose complex challenges to the privacy and security of smart system users. In order to identify phishing websites and maintain the security of smart systems, this study employs LightGBM and domain properties to propose a machine-learning-based strategy. There is a trait where several formulae for domain name creation are consistent, and that is domain name features, sometimes called proportion. To begin, we utilise the proposed discovery model to extract domain characteristics, which include character-level functions and domain name information, for the provided website. In order to make the version more accurate, the characteristics are filtered before being utilised for categorization. According to the results of the experimental comparisons, the discovery design that is suggested employs two types of features for training and is much more effective than the design that just uses one kind of attribute. Also, the suggested technique is suitable for the real-time identification of many phishing websites and has higher detection accuracy than existing approaches.

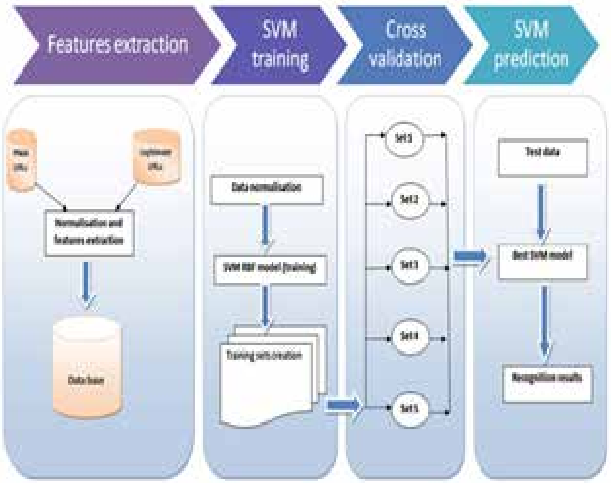


Fig.1. Phishing website process.

METHODOLOGY

Classifiers used in machine learning to envision phishing are expected to be covered in this field. We want to lay out our plan for detecting phishing websites here. There are two sections to this: one deals with classifiers and the other with our suggested system.

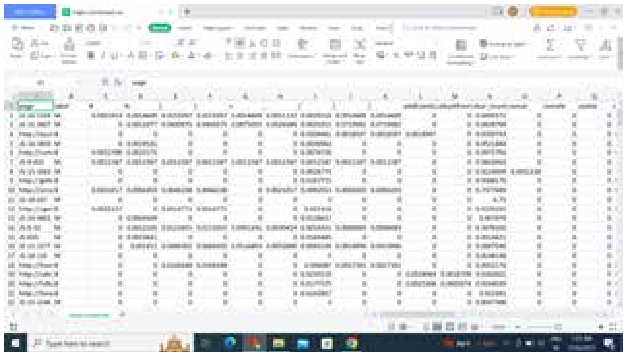
Classifiers and methods for artificial intelligence to evaluate the phishing website The problem of recognising and avoiding phishing websites is complex and energetic. The use of machine learning to provide automated results has become widespread. There are several other ways that phishers might attack, including via speech, websites, malware, and dispatch. Finding phishing links on websites using the Hybrid Algorithm Approach is the main focus of this article. The system's accuracy and quotation rate are both enhanced by the combination of several classifiers. Any classification method may be employed, depending on the application and the dataset. We can't tell whether algorithms are universal or not since they all have diverse uses.

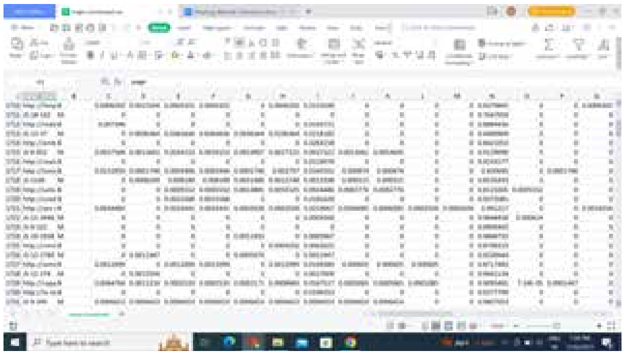
Machine for Supporting Vectors (SVM): This is also one of the supervised and user-friendly categorization formulae. Classification applications are selected, while it may also be utilised for regression. SVMs are distinct from other category formulae because they determine the decision boundary by calculating the distance between the closest data points of all classes. Decision boundaries produced by support vector machines (SVMs) are the best classifiers for maximum margin active aeroplanes or optimal margins. Distinctions in the courses, which are data established points in different aeroplanes, form the basis of the categorization.

Data Collection

Cybercriminals still use phishing as one of the most prominent methods to steal our financial and personal information. Scammers have found the perfect setting to conduct targeted phishing attacks because to our growing reliance on the internet to perform many of our daily services. Now more than ever, phishing assaults are complex and difficult to detect. Intel conducted a research and found that almost all security specialists fail to distinguish between legitimate and phishing emails.

With 87 retrieved traits, the supplied dataset has 11430 URLs. Phishing detection structures that depend upon machine gaining knowledge of will discover this dataset useful for benchmarking their performance. The capabilities come from 3 assets: fifty six are derived from the syntax and structure of URLs, 24 are derived from the content material of the web sites that use them, and seven are retrieved via queries to different services. Half of the URLs in the dataset are phishing tries, at the same time as the alternative half are true.





IMPLEMENTATION

Here we will go over the specific procedures that were followed during the experiment. We will go over the methodical process that was used to examine the data and forecast the phishing attempt. We have made use of URL-only unstructured data. Some 110,664 URLs were retrieved from the web. Which includes both legitimate and phishing URLs, with the former accounting for the vast majority of the links found.

1. On one hand, we have the Phish tank website's unstructured data, which consists of URLs.
2. To begin with, unstructured data is transformed into eight characteristics during preprocessing. The characteristics include the following: IP address, phishing phrase, length of subdomain, suspicious character, number of slashes, prefix/suffix, and URL length.
3. The third step is to build a structured dataset and

send it to the different classifiers together with binary values (0,1) for each feature.

4. Fourth, we train two separate classifiers, SVM and light gbm, and then we compare their accuracy to see which one performs better.
5. Next, the classifier uses the training data to determine if the provided URL is a phishing attempt; if it is not, it displays an error message and the browser visits the requested website.
6. When we compared the accuracy of other classifiers, we discovered that light gbm provided the highest level of accuracy.
7. The execution procedure screen shots are below.

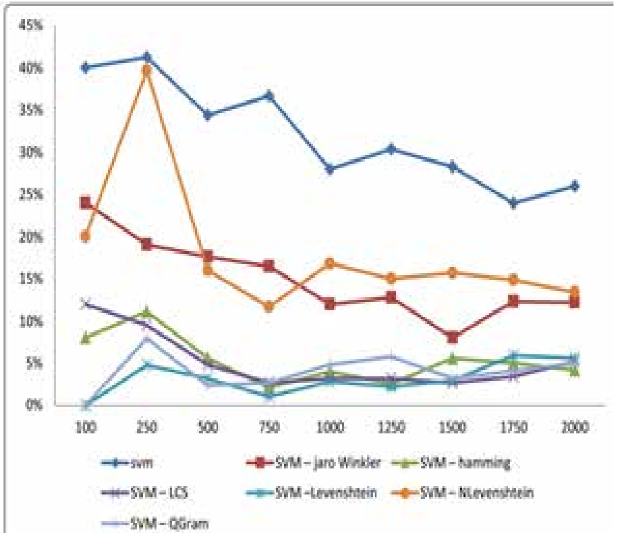


Fig.2. Graphical representation

Based on our testing, we've proven that the use of the similarity distance would possibly help become aware of phishing websites. Adding the distance of similarity substantially multiplied our detection machine's identity fee in 3 out of four checks. Similarly, out of all our experiments, the one the usage of Probabilistic Neural Networks had the lowest detection fee of phishing websites, which is the one instance wherein similarity did not definitely affect the fee of identification. Since our gadget's identity fees become more advantageous with the aid of 21.8% when the hamming distance changed into used as an entrance function, This has an impact on is most important in tests carried out using the SVM approach.

CONCLUSION

By using gadget gaining knowledge of technologies, this article seeks to improve detection strategies for phishing web sites. Our detection accuracy was ninety seven.14% with the lowest false fantastic fee when we used the random forest approach. Using additional data as schooling data also improves classifier performance, in line with the effects. For better phishing internet site detection in the future, a hybrid approach could be utilised, combining the random forest method of machine studying with a blacklist method.

Analysis of Features

Without gathering data pertaining to consumer privateers, such as network traffic, the traits of the domain name used here can also only be retrieved with the aid of using known sequences of domain names. Based on the means of acquisition, there are sorts of area called functions: those relating to the characters used inside the area name and those bearing on the records at the domain name. You can get the area name statistics features from the applicable website or different query websites, however you may get the domain name character capabilities through a local function-extraction method without ever having to visit the internet site.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "PHISHING WEBSITE DETECTION USING MACHINE LEARNING ALGORITHMS", which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department Of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. Ms. Sophiya Shikalgar, Mrs. Swati Narwane (2019), Detecting of URL based Phishing Attack using Machine Learning. (vol. 8 Issue 11, November – 2019)
2. Rashmi Karnik, Dr. Gayathri M Bhandari, Support Vector Machine Based Malware and Phishing Website Detection.

3. Arun Kulkarni, Leonard L. Brown, III2 , Phishing Websites Detection using Machine Learning (vol. 10, No. 7,2019)

4. R. Kiruthiga, D. Akila, Phishing Websites Detection using Machine Learning.

5. Ademola Philip Abidoye, Boniface Kabaso, Hybrid Machine Learning: A Tool to detect Phishing Attacks in Communication Networks. (vol. 11 No. 6,2020)

6. Andrei Butnaru, Alexios Mylonas and Nikolaos Pitropakis, Article Towards Lightweight URL-Based Phishing Detection.13 June 2021

7. Ashit Kumar Dutta (2021), Detecting phishing websites using machine learning technique. Oct 11 2021

8. Nguyet Quang Do, Ali Selamat, Ondrej Krejcar, Takeru Yokoi and Hamido Fujita (2021) Phishing Webpage Classification via Deep Learning-Based Algorithms: An Empirical study.

9. Ammara Zamir, Hikmat Ullah Khan and Tassawar Iqbal, Phishing website detection using diverse machine learning algorithms.

10. Vahid Shahrivari, Mohammad Mahdi Darabi and Mohammad Izadi (2020), Phishing Detection Using Machine Learning Techniques.

11. A. A. Orunsolu, A. S. Sodiya and A.T. Akinwale (2019), A predictive model for phishing detection.

12. Wong, R. K. K. (2019). An Empirical Study on Performance Server Analysis and URL Phishing Prevention to Improve System Management Through Machine Learning. In Economics of Grids, Clouds, Systems, and Services: 15th International Conference, GECON 2018, Pisa, Italy, September 18-20, 2018, Proceedings (Vol. 11113, p. 199). Springer.

13. Desai, A., Jatakia, J., Naik, R., & Raul, N. (2017, May). Malicious web content detection using machine leaning. In 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 1432-1436). IEEE.

Heart Disease Identification Method in E-Healthcare using Machine Learning Classification

Kadari Sai Chandu, Korutla Mounica,
Mahadev Srinath Goud
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

DTV Dharmajee Rao
Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ dtvdrao@gmail.com

ABSTRACT

In this publication, we conducted a study on cardiovascular disease using data analytics. Predicting the occurrence of cardiovascular disease is an emerging field of study, especially as more data becomes accessible. A number of researchers have looked at it using different methods. In order to identify and anticipate illness victims, we used data analytics. We started by doing some preliminary processing on data sets of varying sizes utilizing three data analytics methods (Choice tree, Random forest, SVM, and KNN) to identify the most important attributes according to the connection matrix. Because of this, we are intelligent to estimate the consistency and accuracy of each method. Using clinical criteria, the datasets are classified. Our method use a data mining category approach to examine such factors. The optimal design showed the greatest degree of accuracy for heart disease when the datasets were examined in Python using machine learning methods.

KEYWORDS: *Decision trees, Random Forest, SVM, and KNN.*

INTRODUCTION

Of all the potentially fatal diseases, cardiac arrest is by far the most common. In order to learn more about heart patients, their symptoms, and the progression of their problems [1], doctors conduct several surveys on cardiac diseases. Nowadays, potentially fatal cardiac arrests are very prevalent [2]. A hint as to what lay ahead was given by a number of signs. Research in the medical field have made outstanding use of technical advances to raise healthcare standards. Recent developments in medical technology have made it possible to provide patients with more precise diagnoses and prognoses [3]. In order to accurately prepare for heart ailments, machine learning might be a great choice for you. As a result, three different formulae will be used. The method combines a logistic regression decision tree with an arbitrary forest. On top of that, these three methods consistently and regularly produce better outcomes [4]. Technological progress is making forecasts more comprehensible. Nowadays, people all across the globe work really hard to become famous and rich, and they live lavish lifestyles as a result. In the midst of their hectic schedules, people often neglect their health [5]. As a consequence, their diets and ways of life have changed. Conditions like high blood pressure, diabetes mellitus, and countless others are more likely to develop in young people whose lives are already fraught with stress and anxiety. The progression of cardiovascular disease may occur to any of these reasons [6].

This study used Python to run a battery of categorization and clustering algorithms on the UCI repository's cardiovascular illness dataset. The primary objective is to find all possible combinations of features and test them against various formulae. Afterwards, the best approach out of all the methods is chosen for early-stage cardiovascular disease prediction [7]. It would be much easier to identify and classify the illness if the three algorithms, Decision tree, Random forest, and Logistic Regressionwere applied [8]. The version is trained and classified using a dataset. The disease was predicted using the most precise and effective method once the design was trained [9].

Declaration of Issue

Some of the risk factors for the condition were found to include high blood pressure, total cholesterol, LDL cholesterol, and HDL cholesterol. Predictions for coronary heart disease in 12 years included 383 men and 227 women [9]. It was shown that using the constant variables itself brought the accuracy of this category technique up to par with CHD prediction. Because cardiac problems cannot be predicted with a higher learning rate or more accuracy in their early phases, the current algorithms can only predict them with a 93% accuracy [10].

LITERATURE SURVEY

In order to provide an HDPM more effectively, several studies have confirmed an improvement in cardiac disease clinical prognosis based entirely on tool learning versions. Researchers have achieved a great deal in analysing the efficacy of forecast modifications using two publicly available coronary heart disease datasets, Statlog and Cleveland. A cardiac problem medical decision support system based on chaotic firefly approach and challenging units-based top-notch reduction (CFARS-AR) was developed by Long et al. for the Statlog dataset (2015). In order to classify the illness, units were used to reduce the shape of capacity, and health problems firefly components were used. After that, in comparison to several fantastic variants, like NB, SVM, and ANN, the installed version has become significantly different.

According to Nahato et al. (2015), RS-BPNN is the result of combining BPNN with rough devices-primarily based entirely on trends. Based on the selected characteristics, the recommended RS-BPNN achieved an accuracy of around ninety.Just 4%. A number of efficiency indicators were compared by Dwivedi (2018) across six AI variants: ANN, SVM, LR, precise enough-next-door neighbour (kNN), classification tree, and NB. According to the results, LR outperformed many other styles when it came to accuracy (85%, 89%, 81%, and 85%, respectively), phase of degree of sensitivity, location of information, and precision.

Using a combination of several AI architectures (okay-NN, DT, NB, LR, SVM, Semantic Network (NN)), and a strategy for identifying excellent traits, Amin et al. (2019) were able to complete comparative assessment. Results showed that the hybrid (combining NB and LR) with selected traits achieved an acceptable level of accuracy (87.4%). Researchers have routinely used the Cleveland cardiac health concerns dataset to provide predictive patterns.

A hybrid prediction format was developed by Verma et al. (2016) using K-method clustering, MLP, piece swam optimisation (PSO), and correlation characteristic element (CFS). As a result, the recommended hybrid model achieved an accuracy of up to 90.28 percent.

In a comparative study, Haq et al. (2018) [6] compared a hybrid model that combined various attribute alternative methods with expert system styles. The model included remedy, minimal-redundancy maximal-relevance (mRMR), the very least outright contraction and alternative operator (LASSO), LR, kNN, ANN, SVM, DT, NB, and RF. According to their findings, the fundamental efficiency of the styles is impacted by the loss in capabilities. Compared to other combinations used in the studies, the study found that LR-primarily based on full device uncovering collection of guidelines (MLA) with Relief-based on a combination of the two produced the highest accuracy (up to 89%).

A style based on SVM beauty layout and inferred Fisher rating specific need gadgets (MFSFSA) was suggested by Saqlain et al. (2019). The attributes that were selected were determined by the Fisher score that was better than the desired score. Following that, SVM checked and determined the MCC using a record approach using the selected feature portion. Study after study found that combining FSFSA with SVM produces accuracy, sensitivity, and uniqueness as high as 81.19%, 79.99%, and 88.50%, respectively.

A hybrid approach combining NB, BN, RF, and MLP was proposed by Latha and Jeeva (2019). Accuracy levels as high as 85.4% were achieved by the proposed version. In order to enhance the clinical analysis technique, Ali et al. (2019) [5] advocated for loaded SVMs.

The first SVM was used to filter out irrelevant features, and the second one to predict the occurrence of cardiac problems. The results showed that compared to various variants and prior study, the supported format achieved much higher trendy average efficiency. A hybrid RF/immediate model (HRFLM) was developed by Mohan et al. (2019) to improve the HDPM's performance. In general, they found that the recommended method achieved precision, accuracy, sensitivity level, f-diploma, and strength as high as 88.1 percent, 90.1 percent, 90.8 percent, 90.1 percent, and 82.6 percent.

Gupta et al. (2020) have established a device data form using RF-based fully MLA and variable analysis of combined truths (FAMD). To determine the relevant abilities, the FAMD was used, and to guess the current situation, the RF was used. Using a degree of uniqueness, degree of sensitivity, and precision of up to 96.34%, 89.28%, and 96.76%, respectively, the recommended technique outperformed other variations in terms of previous research study results.

EXISTING SYSTEM

Infections associated to the heart, or cardiovascular diseases (CVDs), have emerged as the most deadly ailment in India and the rest of the world in recent years. Globally, they are the primary cause of a staggering amount of deaths. Consequently, in order to analyze these illnesses and provide the right treatment in a timely manner, a reliable, accurate, and accessible framework is needed. AI techniques and algorithms are currently used to numerous clinical datasets in order to automate the processing of large and complex amounts of data. a while ago, a number of scientists have begun utilizing some AI methods to support the medical community and the professionals studying heart-related illnesses.

PROPOSED SYSTEM

In terms of global mortality, cardiovascular disease ranks first. Forecasting is difficult for doctors as it calls for a higher degree of predictive competence and skill. There may be a knowledge deficit, but data is abounding in the healthcare industry. Despite the abundance of data available online from healthcare systems, effective analytic tools are lacking, making it difficult to uncover hidden trends. Clinical effectiveness, together with cost and waiting time reduction, will unquestionably be enhanced by an automated method. This programme tries to foretell when a disease could strike by using data collected from Kaggle. Finding hidden patterns and estimating the visibility worth on a range are the goals of applying data mining algorithms to the dataset. It is quite difficult to assess and analyse using conventional methods because of the massive amounts of data required to forecast heart status. Our goal is to find a reliable method for predicting the occurrence of cardiovascular disease.

METHODOLOGY

Information Asset

The dataset used in this study to anticipate cardiac status was obtained from the UCI Artificial Intelligence database. It is possible to run machine learning algorithms on the databases that comprise UCI. This data collection is authentic. Including the proper 14 scientific factors, the dataset consists of 300 data instances. The dataset's scientific description is based on tests performed to diagnose cardiac issues, such as the severity of hypertension, the kind of chest pain, the results of electrocardiograms, and others.

Formula Summary

Here we'll go over the two main algorithms that this system uses: i) the decision tree classification formula the Support Vector Machine (SVM) algorithm.

III. The K-next-door-neighbors method.

IV. The random forest.

The hope is that this will lead to a diagnosis of heart disease. Records in the datasets are divided into two groups: training and examination. Following data pre-processing, methodologies for data extraction from categories such as uninformed Bayes and decision trees were used. This section displays the final results of the categorization designs that were created using Python shows. To get the results, we need both training datasets and test sets of data.

RESULTS EXPLANATION

As part of our project, we have integrated a user interface (UI) with a database so that people can register and access their cardiac risk assessment results. Even without creating an account and logging in, users may still use the fast forecast button to anticipate a cardiac issue. The following step is for the consumer to input 13 characteristics, including personal information like age, gender, cholesterol level, etc. It will also train the dataset after delving into the data, and a design will be created according to the individual's specifics. The proportions of examination and training are 25% and75%, respectively. By choosing the develop design and heart disease threat percentage on our user interface (UI), individuals may access alternatives for symptom management and prevention when the percentage is more than 60%, indicating the presence of a potential heart issue. Users may view their past prediction data on our website by logging in using the same login details as when they predicted before. If the design anticipates a cardiac issue, the person might choose to communicate with the client directly by phone or email. Every piece of personal information is stored in a database and may be accessed at any time.

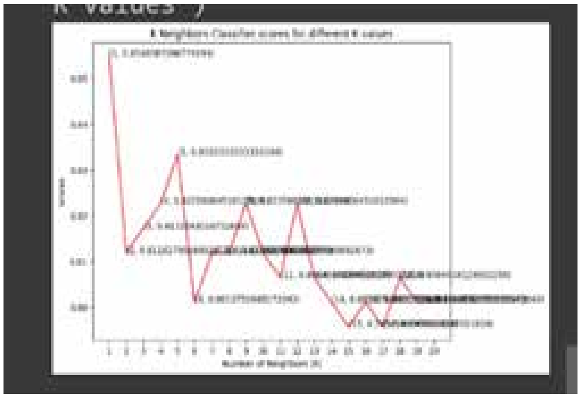


Fig. 1. K neighbors algorithm

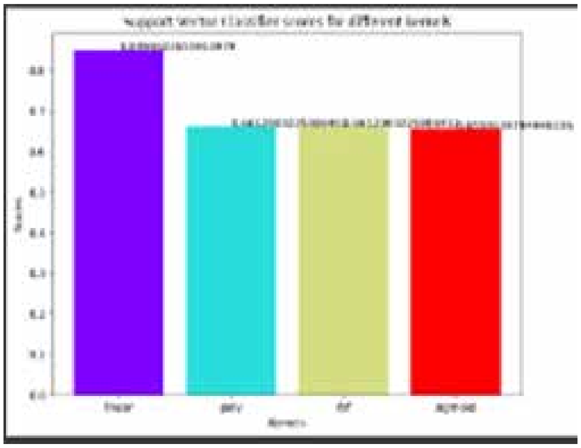


Fig. 2. SVM classifier

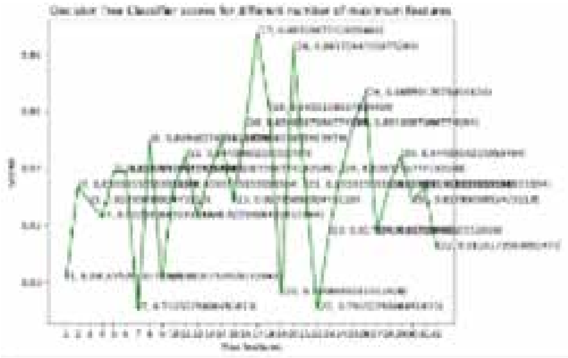


Fig. 3. Decision tree algorithm

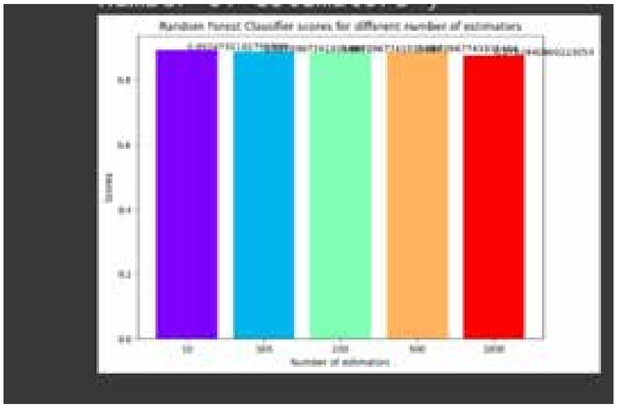


Fig.4. Random forest classifier.

The UCI repository's heart disorder dataset is processed using Python the usage of some of clustering and class algorithms. The number one objective is to discover and target all viable combos of attributes through the use of diverse algorithms. Next, the technique that plays the excellent in predicting the onset of heart disease at an early degree is selected.

CONCLUSION

As the number of fatalities caused by heart disease rises, the need of develop a system that can accurately and reliably predict heart problems grows. The primary goal of the research was to identify the most efficient ML algorithm for the detection of cardiac problems. This project uses the UCI maker learning repository dataset to compare the accuracy ratings of Logistic Regression, Random Woodland, and K Neighbours for cardiac issue prediction. This study's results show that the logistic regression algorithm is one of the most dependable algorithms for predicting cardiovascular disease, with an accuracy score of 89%. The accuracy of maker discovery algorithms is affected by the dataset that is utilised for training and screening. Cardiac disease prediction may make use of a variety of alternative devices finding approaches. Furthermore, logistic regression is an effective tool for dealing with binary classification issues, such as the prediction of cardiac problems. It is possible that decision trees may under perform randomly generated forests. It is also possible to apply set approaches and synthetic semantic networks on the collected data. The outcomes may be improved by comparing and contrasting.

ACKNOWLEDGMENT

We express our gratitude to CMR Technical Campus for providing excellent facilities and assistance to enable us to complete our work on the article "HEART DISEASE IDENTIFICATION METHOD USING MACHINE LEARNING CLASSIFICATION IN E-HEALTHCARE." I want to express my gratitude to our chairman, director, deans, head of the department, computer science and engineering department, guide, and teaching and non-teaching faculty members for their insightful advice and direction in all facets of our work.

REFERENCES

1. Rairikar, A., Kulkarni, V., Sabale, V., Kale, H., &Lamgunde, A. (2017, June). Heart disease prediction using data mining techniques. In 2017 International Conference on Intelligent Computing and Control (I2C2) (pp. 1-8). IEEE.
2. Gandhi, Monika, and Shailendra Narayan Singh. "Predictions in heart disease using techniques of data mining." In 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), pp. 520-525. IEEE, 2015.
3. Chen, A. H., Huang, S. Y., Hong, P. S., Cheng, C. H., & Lin, E. J. (2011, September). HDPS: Heart disease prediction system. In 2011 Computing in Cardiology (pp. 557-560). IEEE.
4. Aldallal, A., & Al-Moosa, A. A. A. (2018, September). Using Data Mining Techniques to Predict Diabetes and Heart Diseases. In 2018 4th International Conference on Frontiers of Signal Processing (ICFSP) (pp. 150-154). IEEE.
5. Sultana, Marjia, Afrin Haider, and Mohammad Shorif Uddin. "Analysis of data mining techniques for heart disease prediction." In 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), pp. 1-5. IEEE, 2016.
6. AlEssa, Ali Radhi, and Christian Bach. "Data Mining and Warehousing." American Society for Engineering Education (ASEE Zone 1) Journal (2014).
7. Shetty, Deeraj, KishorRit, Sohail Shaikh, and Nikita Patil. "Diabetes disease prediction using data mining." In 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pp. 1-5. IEEE, 2017.
8. Methaila, Aditya, Prince Kansal, Himanshu Arya, and Pankaj Kumar. "Early heart disease prediction using data mining techniques." Computer Science & Information Technology Journal (2014): 53-59.
9. Dewan, Ankita, and Meghna Sharma. "Prediction of heart disease using a hybrid technique in data mining classification." In 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 704-706. IEEE, 2015.

A Comparative Study on Fake Job Post Prediction Using Different Data Mining Techniques

Batta Gopi Basava Raju, Tanneeru Kiran, Pavurala Teja
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

Suma S
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ sn.suma05@gmail.com

ABSTRACT

Posting ads for open positions has become ubiquitous in the contemporary era, thanks to the proliferation of online networks and communication tools. Consequently, everyone will be quite worried about the phoney job posting prediction assignment. fake job posing predictions is fraught with difficulty, much like other classification problems. A variety of data mining and classification algorithms, including KNN, decision trees, support vector machines, naive bayes, random forest, multi layer perception, and deep neural networks, are suggested in this research as ways to determine the authenticity of a job posting. We conducted experiments using the 18000-sample Employment Scam Aegean Dataset (EMSCAD). This classification challenge is well-suited to deep neural networks as classifiers. In order to construct this deep neural network classifier, we have used three thick layers. In predicting a fake job posting, the trained classifier demonstrates about 98% DNN classification accuracy.

KEYWORDS: *EMSCAD, DNN, RF, KNN, DT, SVM.*

INTRODUCTION

These days, job-seekers have a lot of options when it comes to fresh and varied jobs because of technological and industry advancements [1]. The advertisements for these work supplies allow job searchers to find their possibilities based on factors like availability, qualifications, experience, appropriateness, etc [2]. The effect of the internet and social media has grown to the point that it affects the hiring process. The marketing of an employment process is crucial to its effective completion, and social media has a huge impact on this [3]. More and more opportunities to disseminate job listings have emerged as a result of social media and online advertising. On the contrary, the percentage of fraudulent task postings, which irritate job applicants, has increased due to the rapid expansion of the ability to share work blog posts [4]. People may confidently express their interest in fresh job ads without worrying about the security or authenticity of their personal, academic, or professional information. Consequently, gaining people's ideas and trust via legitimate professional postings on social and electronic media is quite challenging [5]. The world around us is filled with technologies that simplify and standardise our lives, but they should not create a dangerous environment for professional life. This would be a great breakthrough for recruiting new employees if job postings could be adequately vetted to avoid inappropriate postings. False task postings make it difficult for candidates to find better work, which causes them to spend a lot of time. Human resource management issues may be better addressed with the use of an automated system that can detect and prevent fake job articles [6].

Job Scam: False Task Posting

A job rip-off occurs when an online job advertisement is deceitful and only interested in collecting personal and professional details from prospective applicants rather than really hiring them. Scammers often try to fraudulently acquire funds from job applicants. More than 67% of people who hunt for employment online are vulnerable to scams because they are unaware of the dangers of clicking on fake job ads [2]. This information comes from a recent poll conducted by Activity Fraudulence in the UK. In the United Kingdom, nearly 700,000 people have lost more than $500,000 due to employment fraud. Over the previous two years, the research found an almost 3000% growth in the UK [2]. Fraudsters prey on students and recent graduates because they often want to gain a guaranteed employment, which requires them to pay more. Because fraudsters alter their techniques of labour rip-off so often, cyber crime evasion and security measures cease to be effective in reducing this crime [7].

Typical Forms of Employment Fraud

Phishers create fake job adverts to steal people's personal information, such as their date of birth, social security number, bank account data, insurance information, and income tax details. When fraudsters ask for money using excuses like administrative fees, information security examination costs, monitoring fees, etc., they are committing a front money scam. Scammers pose as legitimate businesses and submit false documents purporting to be pre-employment background checks, such as driver licences, bank statements, and personal information.

When they trick students into depositing funds into their accounts and then transferring it back, they are committing a prohibited money weighing fraud [2]. The 'cash' option allows you to operate without paying taxes. In order to trick people into giving them their money, scammers often create phoney websites that appear like legitimate businesses or banks. Instead of meeting in person, many employment scammers prefer to lure victims in via email. In an effort to pass themselves off as recruiting agencies or talent scouts, they often aim for professional networking sites like Linkedln [8]. Typically, they will make an effort to portray their company account or websites in the most reasonable light possible to the job applicant. No matter what style of work scam they use, their goal is always to trick unwary candidates into giving up personal information so that they may steal their jobs or their money. [6, 7]

II EXISTING SYSTEM

You can tell whether a job posting is legit or not with the help of several investigations. Verifying online job ads for fraud is an effective use of a wide range of research methods. False online job advertisements were identified by Vidros [1] et al. as job fraudsters. They found data on a large number of legitimate and well-known companies, as well as those that made fraudulent job ads or vacancy notifications for malicious purposes. They tried out several classification techniques on the EMSCAD dataset, such as naive bayes, random woodland, absolutely no R, one R, and so on. On this dataset, the Random Forest Classifier performed the best, with a classification accuracy of 89.5%. When applied to the dataset, they found logistic regression to be woefully insufficient. After they stabilised the dataset, they tested one R classifier, and it performed well. Using a plethora of prominent classifiers, they set out to identify problems with the ORF (Online Recruitment Fraudulence) version and provide solutions.

In order to detect online job fraud, Alghamdi [2] et al. suggested a concept. They put the AI system through its paces on the EMSCAD dataset. There were three stages to their service for this dataset: data pre-processing, attribute selection, and classifier-based fraud detection. To ensure that the fundamental message pattern remained intact, they eliminated noise and html tags from the data during the pre-processing stage. To effectively reduce the quantity of traits, they used the attribute choosing approach. In order to find fake job blog posts in the test data, we used a set classifier that took use of arbitrary forest and a Support Vector Device for feature selection. It seemed that the random forest classifier was a tree-structured model that, with the aid of the bulk voting approach, operated as a set classifier. The accuracy of this classifier in identifying false job postings was 97.4 percent.

The authors Huynh et al. [3] recommended using pre-trained deep neural network architectures with text datasets, such as Text CNN, Bi-GRU-LSTM CNN, and Bi-GRU CNN. They were tasked with organising a dataset of IT tasks. Using TextCNN, which includes a convolutional layer, a merging layer, and a fully connected layer, they trained the IT task dataset. This layout used convolution and merging layers to train data.

Then, the fully linked layer received the compressed experienced weights. As a category approach, this design made advantage of the softness feature. They also used set classifiers that used the majority voting approach to increase category accuracy, such as Bi-GRULSTM CNN and Bi-GRU CNN. They found that Bi-GRU-LSTM CNN had a classification accuracy of 70% and TextCNN had a precision of 66%. A set classifier with a 72.4% accuracy rate performed the best on this classification challenge.

Using text processing, Zhang et al. [4] proposed an automated model for fake detectors to analyse real and bogus information (including short articles, designers, and subjects). The PolitiFact site's Twitter account had provided them with a bespoke dataset including articles and other content. The suggested GDU diffusive unit model was trained using this dataset. As an automated fake detector design, this trained version worked effectively when information from various resources was received simultaneously.

Negative Aspects

1) Standard Machine Learning runs the system.
2) Looking through large data sets is not something the machine can handle.

PROPOSED SYSTEM

False task posts have been detected by the system using EMSCAD. There are 18,000 samples in this dataset, and the class tag is one of 18 attributes per row. Job ID, title, location, division, pay range, company profile, description, needs, benefits, telecommunication, logo, has questions, employment kind, necessary education and experience, sector, function, and deceptive are some of the elements. We have made use of only seven of these eighteen characteristics, all of which are swapped into categorical traits. Message worth is transformed into categorized worth for T telecommuting, has_company_logo, has_questions, employment_type, needed_experience, required_education, and misleading. As an example, the values for "employment_type" are substituted in the following way: 0 for "none," 1 for "permanent," 2 for "part-time," 3 for "others," 4 for "agreement," and 5 for "short-lived." The major goal of transforming these attributes into categorical form is to identify and categorise fake job advertising without involving message processing or natural language processing in any way. Only those classified features have been used in this study.

Benefits

1) The very precise and rapid EMSCAD technique that was suggested has been put into action.
2) Since the system accurately finds fake work articles, it creates incongruity for the task hunter to find their true job, leading to a significant loss of time. As a result, the system is highly efficient.

WORKING METHODOLOGY

Service Supplier

The Company must enter a valid user ID and password in order to access this section. He would be able to do actions like accessing the Train and Test Data Sets if he successfully logs in. See Trained and Tested Precision Results in a Bar Chart, Forecast Task Message Kind Details, Find the Proportion of Work Article Types, and See Trained and Examined Accuracy in a Bar Chart Get your hands on Educated Data Sets, Sort of Sight Job Message Prediction Outcomes, Check Out Every Remote User

Evaluate and Commend People

The admin may see a complete list of registered clients in this section. Here, the administrator may see the user's details (name, email, handle, etc.) and grant them access.

Customer on the Go

After all, there are n clients in this section. The client must check in before proceeding with any operations. All information provided by the user will be securely stored by the data provider after they register. He has to use a legitimate person's name and password to log in after registering. The following actions will be taken by the user after Login is hit: SIGHT YOUR ACCOUNT, PREDICT WORK MESSAGE FORECAST, ARTICLE TASK BLOG POST DATA SETS, and LOGIN AND REGISTER.

Fig.1. Home page

Fig.2. Login page

Fig.3. Upload page details

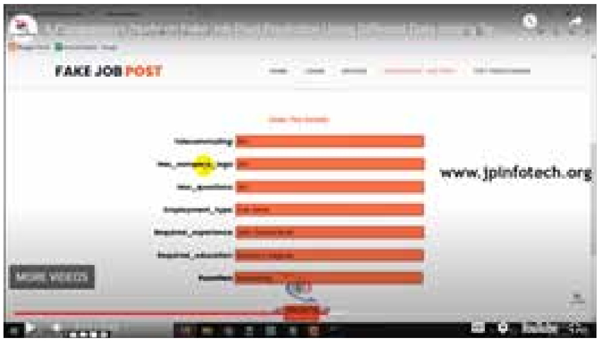


Fig.4. Output results

CONCLUSION

The identification of job fraud has recently become a major problem on a global scale. We have examined the effects of work scam in this article; this is a growing field of study that has made it difficult to identify fake job postings. The EMSCAD dataset, which contains authentic but fake job postings, has been tested by us. This study includes experiments with deep neural networks (DNN) and artificial intelligence algorithms (SVM, KNN, Naïve Bayes, Random Forest, and MLP). This paper presents the results of a research comparison between classic AI and classifiers based on deep understanding. We found that the Random Forest Classifier outperformed all other traditional equipment detecting techniques in terms of category accuracy, with 99% precision for DNN (layer 9), and 97.7% accuracy for Deep Semantic Network as a whole.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "A COMPARATIVE STUDY ON FAKE JOB POST PREDICTION USING DIFFERENT DATA MINING TECHNIQUES", which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. S. Vidros, C. Kolias , G. Kambourakis ,and L. Akoglu, "Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset", Future Internet 2017, 9, 6; doi:10.3390/fi9010006.

2. B. Alghamdi, F. Alharby, "An Intelligent Model for Online Recruitment Fraud Detection", Journal of Information Security, 2019, Vol 10, pp. 155176, https://doi.org/10.4236/iis.2019.103009 .

3. Tin Van Huynh1, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen1, and Anh Gia-Tuan Nguyen, "Job Prediction: From Deep Neural Network Models to Applications", RIVF International Conference on Computing and Communication Technologies (RIVF), 2020.

4. Jiawei Zhang, Bowen Dong, Philip S. Yu, "FAKEDETECTOR: Effective Fake News Detection

with Deep Diffusive Neural Network", IEEE 36th International Conference on Data Engineering (ICDE), 2020.

5. Scanlon, J.R. and Gerber, M.S., "Automatic Detection of Cyber Recruitment by Violent Extremists", Security Informatics, 3, 5, 2014, https://doi.org/10.1186/s13388-014-0005-5

6. Y. Kim, "Convolutional neural networks for sentence classification," arXiv Prepr. arXiv1408.5882, 2014.

7. T. Van Huynh, V. D. Nguyen, K. Van Nguyen, N. L.-T. Nguyen, and A.G.- T. Nguyen, "Hate Speech Detection on Vietnamese Social Media Text using the Bi-GRU-LSTM-CNN Model," arXiv Prepr. arXiv1911.03644, 2019.

8. P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, and H. Hao, "Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification," Neurocomputing, vol. 174, pp. 806814,2016.

9. C. Li, G. Zhan, and Z. Li, "News Text Classification Based on Improved BiLSTM-CNN," in 2018 9th International Conference on Information Technology in Medicine and Education (ITME), 2018, pp. 890-893.

10. K. R. Remya and J. S. Ramya, "Using weighted majority voting classifier combination for relation classification in biomedical texts," International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014, pp. 1205-1209.

Tweet Based Bot Detection Using Big Data

Ummadisetti Hariharan, Gogurla Pranay, Manthena Nikhila
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

R Sai Krishna
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ Saikrishna.it@cmrtc.ac.in

ABSTRACT

Among the many millions of users of social media platforms that enable micro blogging, Twitter stands out. Because of its popularity, Twitter has been the subject of many attacks, including the dissemination of false information, malicious links, and phishing scams. Because of their ability to launch massive attacks and manipulation campaigns, botnets based on tweets pose a significant threat to people. To address these concerns, researchers have used big data analytics techniques, particularly superficial and deep knowledge methodologies, to accurately differentiate between real people's accounts and automated bots using tweets. In this research, we review the current state of the art in tweet-based crawler detection methods and provide a taxonomy to classify them. Along with the performance outcomes, we also detail the shallow and deep learning approaches to bot discovery based on tweets. In the end, we explore the current state of affairs in the field of tweet-based robot discovery and discuss the challenges and unanswered questions.

KEYWORDS: *ML, DL, Large scale attacks, Twitter data, big data.*

INTRODUCTION

These days, people' go-to methods of connecting with one another is via various forms of social media. It is also mostly used by businesses to communicate with customers. The number of active social media users worldwide is 3.5 billion, according to [1]. Brand awareness and revenue may be enhanced when companies use social media platforms like Facebook, Twitter, LinkedIn, and many more. Of all the major social media sites, Twitter is among the most well-known. It boasts 340 million monthly active users who may communicate in a variety of ways and express their opinions on a wide variety of subjects.

Many different types of attacks might be directed at Twitter. For example, in July 2020, high-profile Twitter accounts were compromised in a spear phishing attack [2]. Additionally, dishonest people or businesses could appear as legitimate ones by creating fake accounts. A group of malevolent accounts known as "robot web" may also utilise Twitter; these accounts are not run by humans but by computer programme. Users' security is seriously jeopardised by social media crawlers that rely on tweets. The purpose of these bots is to disseminate spam, phishing links, and false material. They aren't utilised as bots to launch distributed denial of service (DDOS) assaults, but they might be used as command and control (C&C) facilities to plan such an attack [3, 4]. They can also communicate with human accounts to steal their credentials. The deployment of large-scale control initiatives to sway public opinion is another common purpose for these crawlers. While human users account for the remaining 48% of website traffic, botnets account for 52% (as shown in research [5]). Another important thing to remember is that some crawlers have more than 350,000 synthetic fans. The development of detection systems capable of accurately differentiating between Twitter bot accounts and human accounts is necessary for the management of the aforementioned challenges. Given that there are around 500 million tweets generated daily, or 6,000 tweets per second [6], Twitter data is one example of massive data.

For processing massive amounts of data, finding hidden patterns, and establishing correlations among data points, big data analytics has seen extensive usage across many domains [7] _ [11]. The examination of massive amounts of data greatly enhances expert system tactics. Because of their effectiveness in handling complex and diverse data, automatically understanding models, uncovering hidden patterns, identifying dependencies, and gaining insights from data assessments, deep learning and shallow (conventional) learning approaches have garnered significant attention from both academia and business.

For users' benefit, Twitter has made heavy use of AI to detect tweet referrals. In reality, Twitter data is fed into deep neural networks to recommend relevant online material to consumers, enhancing their overall experience with the system [12]. In the fight against offensive materials, artificial intelligence has become an indispensable tool. Expert system technologies, rather than people, were used to place almost 300,000 accounts on hold in 2017.

This study aims to provide many ways for discovering bots in tweets. These methods compare human and bot accounts using superficial and deep understanding tactics. In particular, this study primarily contributes to the following areas:

1) The most recent developments in AI techniques for robot identification based on tweets have been organised into a taxonomy.

2) The solutions up to the year 2020 are covered in a thorough analysis of shallow and deep finding approaches for bot discovery based on tweets.

3) We explore the challenges and unanswered questions related to robot discoveries using tweets.

LITERATURE SURVEY

A brief comparative assessment of the research work in the field of Twitter spam detection from 2009 to 2015 was provided by Kabakus and Kara [1]. Based on four broad categories, they detailed several finding methods: account-based, tweet-based, graph-based, and hybrid-based. Using information on the account's age, number of followers, and other acquired characteristics, the account-based methods were able to make use of the profile's metadata. The utilisation of functions such as the range and toughness of connections between persons for spam identification was discovered in graph-based approaches. But in tweet-based methods, the research mainly aimed at spam detection using the link and its acquired properties, such domain and size. The URLs that users submitted were checked for dangerous or benign content in order to identify spammers. In addition to this, the authors brought attention to features that were said to enhance spam detection but were actually disregarded.

In the realm of discovering spam users across platforms, Chakraborty et al. [4] conducted an additional relevant study. It was acknowledged by the authors that diverse platforms, like email, blog sites, or micro blogs, need alternative methods and features to accomplish precise finding. That is why the proposed approaches from 2011 to 2015 were classified according to the platform the dataset is housed on. Every single method team was subjected to the same platform-wide quality comparison.

The botnet hid its actual landing sites by using URL reduction services and redirects, as noted by Besel et al. [2]. They found that people registered landing pages on phishing websites, clicked on these URLs, and discovered the bot master who set up the Bursty botnet. They confirmed that the bot master may still control solutions pertaining to Twitter bots. The cyber crime method, the black marketplaces, and Twitter's cyberspace architecture are all covered in this investigation.

Current research is focused on discovering social botnets on Twitter, as summarised by Alothali et al. [5]. They rationally outlined the advantages and disadvantages of each recommended approach. There were three primary categories into which the methods were sorted: graph-based, maker learning-based, and group sourcing-based. Among the three methods, the group sourcing strategy which relies on human expertise to spot patterns, is said to be the most prone to mistakes. It was also found that among the most often used approaches for identifying social bots among Twitter users are artificial intelligence technologies and, more specifically, random forest classifiers.

In a comprehensive assessment, Latah [6] focused on the covert nature and detection tactics of harmful social bots. There was a thorough evaluation of graph-based, AI-based, and new techniques for discovery by the author. Furthermore, the article assessed the tactics' strengths and weaknesses as well as the crawlers' methods for evading detection. As a result, the report offered suggestions that may strengthen security measures against dangerous robots.

EXISTING SYSTEM

Presented below is a synopsis of the studies conducted on Twitter spam detection from 2009 to 2015. In four categories, they detailed several detection methods: account-based, tweet-based, graph-based, and hybrid-based. It was shown that account-based techniques make use of client account information, including fan count, compliance with count, and other acquired functions like account age. Contrarily, graph-based approaches revealed that spam detection took use of characteristics like range and stamina of connections between individuals. But in tweet-based methods, the research is primarily aimed at spam detection using URL and its acquired properties, including domain name and length. In order to identify a spammer, we evaluated and categorised all submitted URLs as either hazardous or harmless. Along with this, the authors brought attention to improvements that were proposed but never implemented, which would have improved the spam detection.

Negative aspects

1) Learning-based discovery strategies are not implemented in the system.
2) Currently, there isn't a plan to improve users' experience on Twitter by using deep semantic networks to identify relevant content for them.

PROPOSED SYSTEM

The suggested version uses the bidirectional approach, which allows for a better grasp of the whole message context by processing twitter phrases forward and backward for each layer. A public dataset called Cresci-2017 is used to train the design. This dataset includes 3,474 human accounts and 1,455 crawlers, totaling 11.4 million tweets. To suit the words embedding version, every tweet was pre-processed and tokenism before training. In order to convert text into mathematical vectors that the network could understand, a pre-trained GloVe version was used. A three-layer architecture was used to process the vectors; the starting value of the lowering failure layer was set at 0.5. They used two separate screening datasets, one with 1,982 accounts and the other with 928 accounts, to test their version, and the results were 93% and 95% accurate, respectively. In order to differentiate between bots and people, RTbust describes a new deep learning model that uses retweet patterns. The authors studied both crawler and human behaviour patterns before creating the version. The data revealed a distinct pattern of retweeting based on the time stamp, which was then classified into four distinct patterns. First, there's the droplet pattern, which represents typical users for whom a reasonable amount of time elapses between posting a tweet and retweeting it. Prospective crawlers' suspicious and fast retweeting trend led to the three remaining patterns.

The Benefits

1) A number of state-of-the-art techniques were provided by the proposed system, which significantly improved the efficiency of spam detection.

Having the ability to find bots based on tweets makes the system considerably more efficient.

WORKING METHODOLOGY

Assistance vector makers are a subset of supervised learning techniques that are under the umbrella of regression and category machine learning. Support vector machines (SVMs) are a set of machine learning methods that may be applied to data in order to find patterns within it. Offered A collection of training data that has to be sorted. Data is usually partitioned into training and testing sets while performing a categorization job. It has been use in a broad variety of practical applications, including the following: data classification, micro-array gene expression data analysis, photo sorting and items finding, tone recognition, handwritten message discrimination, and number discovery.

One set that stands out is random woodland, which uses a majority vote to compile the results of many choice trees. The idea of ensemble knowing is to pool the output of several classifiers into a single neighborhood-wide decision. Using the bootstrap method to build starting data, each decision tree in the forest is built by picking distinct samples. Afterwards, the committee uses the route with the most diverse votes as its estimate based on the decisions made by different individual branches. The RF approach uses the boot bagging mix technique with CART algorithms to generate trees. Data is immediately divided into two categories: training and test. A bootstrap (resampled and tested) technique is used to choose samples from the training data. Some of these samples will definitely form trees, while others will not.

IMPLEMENTATION

Company

The Company must log in using a valid individual phone number and password in this section. He will be able to access the Training and Test Data Sets and perform some basic operations if he successfully logs in. Keep an eye on all remote users, see trained and evaluated precision in a bar chart, see educated and tested precision results, see predicted tweet kindness, find out tweet kindness proportion on data sets, download and install trained data sets, view tweet kindness ratio results, and more.

People who have View and Licence privileges.

The admin may see a complete list of registered users in this section. This allows the admin to see user information such as name, email, and address, and allows the admin to authorise users.

Solo Traveller

Customer number n are found in this section. Everyone has to sign up before they can do anything. Personal information may be entered into the database after a person registers. He has to log in using his legitimate personal phone number and password when he hits the enrol button. Customers will do minor operations such as viewing their profile, predicting the kind of tweet they will send, and updating their Twitter information sets after login is successful.

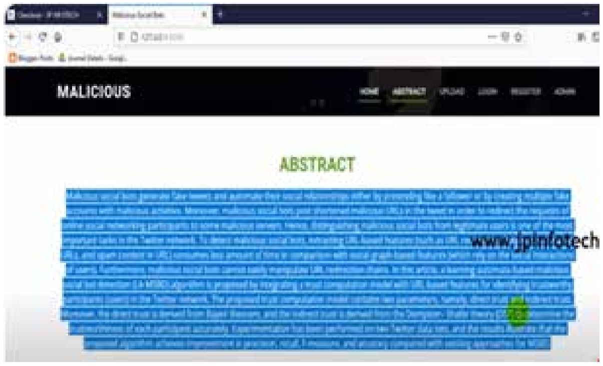


Fig.1.Home Page

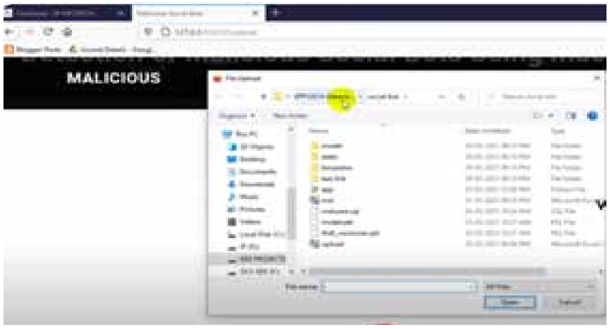


Fig.2. Upload the data set details

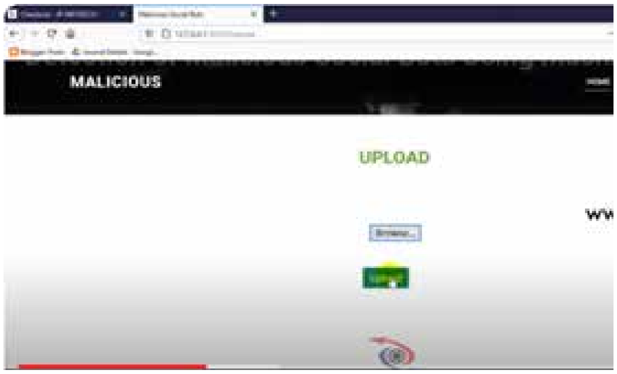


Fig.3. Dataset uploaded

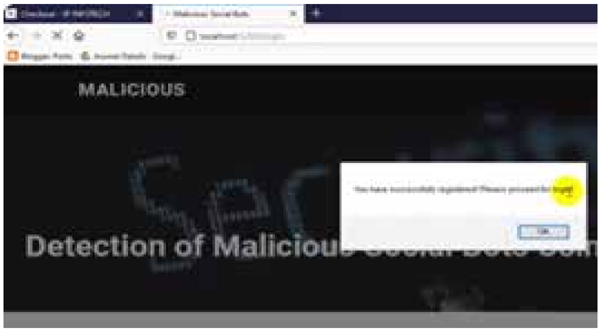


Fig. 4. Registration completed

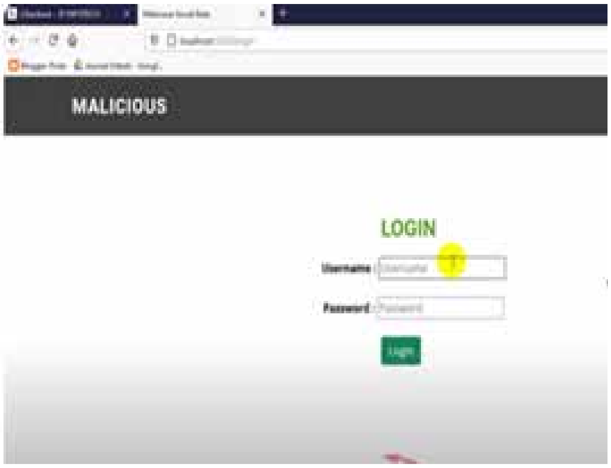
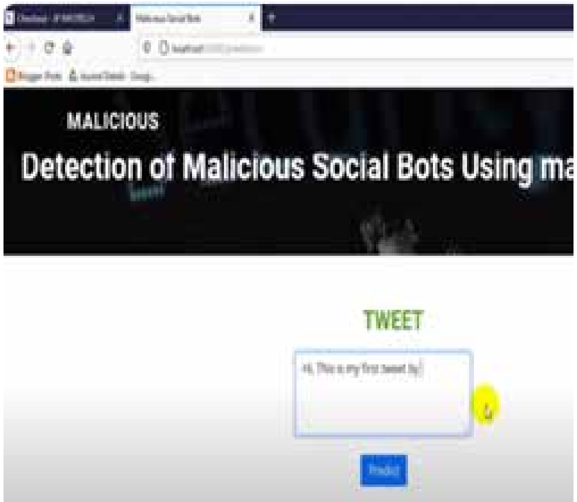


Fig.5. Login page



Hi, This is my first tweet by

Fig.6. Input tweeter data

CONCLUSION

One of the most widely used social media platforms, Twitter facilitates connections between people and helps businesses reach out to clients. A botnet that relies on tweets might compromise Twitter by creating fake accounts to launch massive assaults and exert control over many projects. In our analysis, we have concentrated on big data analytics, namely shallow and deep learning, to counteract botnets that employ tweets and accurately differentiate between real users' accounts and those used by automated bots. We have reviewed the literature and provided a taxonomy that categorises the state-of-the-art robot discovery approaches using tweets around the year 2020. Also included are descriptions of the shallow and deep finding approaches, together with their performance results, for tweet-based crawler discovery. At last, we discussed the outstanding issues and future study obstacles and gave our solutions.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "TWEET BASED BOT DETECTION USING BIG DATA", which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department Of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. M. Mohsin. (2020). 10 Social Media Statistics You Need to Know in 2021. [Online]. Available: https://www.oberlo.com/blog/social-mediamarketing-statistics

2. I. Arghire. (2020). Twitter Hack: 24 Hours From Phishing Employees to Hijacking Accounts. https://www.securityweek.com/twitter-hack-24-hours-phishing-employees-hijacking-accounts

3. The Rise of Social Media Botnets. Accessed: Feb. 21, 2021. [Online].Available: https://www.darkreading.com/attacks-breaches/the-rise-ofsocial-media-botnets/a/d-id/1321177

4. M. Imran, M. H. Durad, F. A. Khan, and A. Derhab, ``Toward an optimalsolution against denial of service attacks in software de_ned networks,"Future Gener. Comput. Syst., vol. 92, pp. 444_453, Mar. 2019.

5. M. S. Savell. (2018). Protect Your Company's Reputation From Threats by Social Bots. [Online]. Available: https://zignallabs.com/blog/protect-yourcompanys-reputation-from-threats-by-social-bots/

6. S. Aslam. (2021). Twitter by the Numbers: Stats, Demographics &Fun Facts. [Online]. Available: https://www.omnicoreagency.com/twitterstatistics/

7. A. Aldweesh, A. Derhab, and A. Z. Emam, ``Deep learning approaches foranomaly-based intrusion detection systems: A survey, taxonomy, and openissues," Knowl.-Based Syst., vol. 189, Feb. 2020, Art. no. 105124.

8. S. Mahdavifar and A. A. Ghorbani, ``Application of deep learning to cybersecurity:A survey,'' Neurocomputing, vol. 347, pp. 149_176, Jun. 2019.

9. E. B. Karbab, M. Debbabi, A. Derhab, and D. Mouheb, ``MalDozer: Automaticframework for Android malware detection using deep learning,''Digit. Invest., vol. 24, pp. S48_S59, Mar. 2018.

10. F. A. Khan, A. Gumaei, A. Derhab, and A. Hussain, ``A novel twostagedeep learning model for ef_cient network intrusion detection,'' IEEEAccess, vol. 7, pp. 30373_30385, 2019.

11. A. Derhab, A. Aldweesh, A. Z. Emam, and F. A. Khan, ``Intrusion detectionsystem for Internet of Things based on temporal convolution neuralnetwork and ef_cient feature engineering,'' Wireless Commun. MobileComput., vol. 2020, pp. 1_16, Dec. 2020.

12. B. Marr. (2020). How Twitter Uses Big Data and Arti_ cialIntelligence (AI). [Online]. Available: https://www.bernardmarr.com/default.asp?contentID=1373

13. A. T. Kabakus and R. Kara, ``A survey of spam detection methods onTwitter,'' Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 3, pp. 29_38, 2017.

14. M. Chakraborty, S. Pal, R. Pramanik, and C. R. Chowdary, ``Recent developmentsin social spam detection and combating techniques: A survey,''Inf. Process. Manage., vol. 52, no. 6, pp. 1053_1073, Nov. 2016.

15. E. Alothali, N. Zaki, E. A. Mohamed, and H. Alashwal, ``Detecting socialbots on Twitter: A literature review,'' in Proc. Int. Conf. Innov. Inf. Technol.(IIT), Nov. 2018, pp. 175_180.

A Student Attendance Management Method based on Crowd Sensing in Classroom Environment

Lingampalli Anugna
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

N Shweta, Rakshitha
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ shwetanashikar@gmail.com

ABSTRACT

One critical use case for smart cities is smart knowing settings, and one critical technique for ensuring high-quality learning is assessing students' attendance in class. Using a trainee attendance management technique called AMMoC (Participation Management Technique based upon Crowdsensing), this study proposes a solution to the current problems with course attendance verification, such as poor performance and easy to counterfeit. There are two parts to AMMoC: the setup phase and the testing phase. During the setup process, a teacher notifies the server that they need to track student attendance. Once the web server receives the request, it will send a request to the students to provide their location information. Once the web server receives all the necessary information from the trainees, it will generate the student place map. During authentication, the server verifies the accuracy of the location data by requesting a headcount from several pupils. The two parts that make up the authentication process are the job assignment and participation confirmation modules. As part of the work project module, AMMoC finds the improved sequence of subareas and verifies them using the Monte Carlo method. Then, it asks the verifies to tally the students in the lower area. In the end, the attendance confirmation module will check the data results. From what we can tell from our experiments, AMMoC is a great choice for presence checking applications in the classroom because of its fast speed, little impact on the learning environment, and outstanding anti-cheating performance.

KEYWORDS: *AMMoc, Server, Attendance, Count, Class room, High efficiency.*

INTRODUCTION

Given the widespread use of mobile phones, one of the most pressing concerns in the development of smart cities is how to create an environment conducive to mobile learning and interaction [1]. In today's educational institutions, mobile discovery is quickly becoming a crucial paradigm for knowledge. Many problems with traditional methods of classroom instruction may be addressed by incorporating mobile computing into the classroom (i.e., mobile education and learning) [2]. These problems include tedious course administration, late feedback from mentors, and insufficient communication between instructors and students. In recent years, mobile education has emerged as a major trend in the field of modern education [3].

One important metric for evaluating the quality of a training course is the percentage of students who actually attend. In order to investigate the correlation between the percentage of college students who actually show up to class and their performance in class, Lukkarinen et al. [1] used clustering and regression analysis. A high percentage of student engagement will increase the mentor's influence, and they discovered a positive correlation between the two. On top of that, absenteeism impacts not just students' individual performance but also the whole classroom environment [2]. An important aspect of college administration has always been presence.

The current method of taking attendance in classes is often done by hand and comes in two varieties: one that is unsupervised and one that is guided by the teacher. Without teacher guidance, students complete participation monitoring by passing a check-in form in class. However, sending out the forms will not only mess with the class schedule, but it will also lead to some phoney attendance. the third Teachers or aids in the classroom will call on each student by name to ensure they are present during the class participation contacting instructor supervision. This kind of roll-calling is utterly useless. A significant amount of class time will be used by the roll-calling procedure when there are a large number of pupils [4]. Our investigation into the steps involved in manually verifying participants' involvement led us to the conclusion that this is necessary to prevent trainees from accidentally checking their presence while they work on other activities. Parallelizing the presence checking technique is thus essential for improving the efficiency of presence monitoring. One service that may be implemented is the use of radio frequency identification (RFID) readers in classrooms. Students could simply show their RIFD cards to verify their attendance [5]. The plan's benefits to presence examination effectiveness are undeniable, but the plan's downsides are also obvious. The first major issue is the astronomical cost of releasing RFID class guests. Second, we still don't know for sure whether someone is signing in fraudulently since RFID scanners can't verify the cardholder's identification. Some novel approaches to these problems are developing in tandem with the proliferation of mobile smart gadgets. As an example, trainees may simply use their mobile devices to complete participation monitoring on apps that are relevant to attendance checking [6]. Though it still can't tell whether someone is logging in as an impostor, this technique cuts system installation costs in half. In order to finish the presence monitoring for others, students may simply bring their own phones into class. Some researchers have proposed bio-metric technologies including facial recognition, fingerprint scanning, and voice-print acknowledgment as a potential solution to this issue in the context of the participation monitoring system [7-9]. Because these biological qualities may be acquired using smartphones, which can cut expenditures, face recognition and voiceprint recognition are preferred for course presence inspecting systems. Biometric verification eliminates the problem of fraudulent attendance tracking, but it might put students' privacy at risk and their home security at risk [10].

We provide AMMOC, a smart presence monitoring approach, in this work. AMMOC doesn't have to collect students' biological traits or use supplementary equipment in the classroom. In order to complete attendance tracking, AMMOC only needs two Android applications, one for instructors and one for students [6]. The apps employ shared confirmation amongst trainees. The AMMOC classroom is divided into many sections, and trainers are assigned to check the student body in each section. Group picking up jobs are the building blocks of the confirmation procedure. the eleventh Within a certain time limit, trainees report their whereabouts to AMMOC as part of the presence verification process. Once AMMOC has the trainees' locations, it employs a sophisticated search algorithm to choose a subset of students to work on group sensing tasks that require them to report the total number of students in a given region, among other things. The results of the crowd-picking tasks submitted by the students will allow AMMOC to determine the veracity of the original data. The primary contributions to this work are as follows.

This article introduces a method for managing students' attendance in class that combines active reporting with a sample check of their location data. The method has a number of benefits, including minimal disruption and high efficiency in real-time [7]. To accurately choose the improved sub areas for participation verification, this study suggests a method that considers the worth of sub regions in relation to the remaining student body [8].

(3) To enhance the participation monitoring system's anti-cheating performance, this research proposes a below region creation approach based on unique randomness. This technique can fully identify the viable sub regions room.

SURVEY OF RESEARCH

Combining face recognition with the rapidly expanding contemporary technology known as the Internet of

Things (IoT) makes providing smart home systems easier, less complicated, and more trustworthy. Machine Learning, a subfield of Artificial Intelligence that has many potential applications, including facial recognition and the exploration of new markets and the optimisation of current ones. While humans struggle greatly with facial recognition, computers have no such issues. Consequently, they might be used in situations when a larger number of photos has to be included in face database entries. A training set, a face identification feature, and an image capturing component should make up the training established monitoring sub-system's conceptual style.We also saved the trained images in the face recognition folder after that was done. To put our Technique into action, here are a few crucial steps. The camera continuously records data in real-time while it is near the classroom entrance. It captures information from the live stream from the camera. We compared the captured photos to the ones that were in a folder while we were qualifying. If the two images are a good fit, the identified or found face will show the student's name and registration number. In order to detect face photo spoofing, this study employs an image structure analysis technique. Combining the Grey Level with the Regional Binary Pattern (LBP) is the suggested approach for evaluating structures. Using these features, line mapping is applied to faces. In this case, we convert photographs to greyscale using the grey-scale technique. It is possible to apply the circulation of an oriented variation on both faces and objects. Following the reduction of all images to grayscale, an integer was assigned to each pixel. The primary goal is to identify the area of the face that is darker. Eigenfaces are a way to encode and decode pictures and faces by removing unwanted features. A collection of faces with all the necessary characteristics can only be generated from a large number of images. Liveness detection refers to any method used to identify a spoof attempt by determining whether the biometric example's resource is a real person or an inaccurate depiction. In order to establish whether the resource is real-time or replicated, algorithms examine data obtained from biometric sensing equipment. A digital identification burglary known as "face spoofing" occurs when a person's identity is stolen via the use of a photograph or video of their face to impersonate their bio-metric data. While face spoofing may be used for many money-laundering schemes, it is most often employed to commit crimes related to financial institution identification fraud. This is the identical method utilised in banking identity frauds. Particularly in digital settings, identification spoofing has grown in popularity lately. Its primary function is to facilitate the opening of savings accounts and loan applications inside financial institutions.In order to open a bank account in a foreign country, identity spoofing schemes employ face unlock recognition technologies that rely on face spoofing, such as photos/ selfies, masks, and configurations. Using the data, we compute the variation and covariance matrices. An eigenvector is every single picture. This method is used to eliminate the functions of the eyes, nostrils, and mouths since they possess several other features. But these strategies should only succeed in passing the anti-spoofing face recognition or phishing detection tests that are necessary at certain points in the on boarding process, when signing contracts, when getting loans, or when purchasing new solutions. False positives, or spoofing, are a problem for face-based safety systems. In order to gain unauthorised access to and usage of a protected system, spoofing occurs when an attacker pretends to be a registered user.

EXISTING SYSTEM

Near Field Communication (NFC) and Radio Frequency Identification (RFID) are common components of the ID-based participation verification system. The acronym AMS stands for "presence administration and information service" and was proposed by Rjeib et al. [3]. In AMS, the RFID tag on each student's ID card is tied to their personal information and class schedule. Web app users may access and see all trainee information and attendance records stored in the database.

Their presence tracking technology, TouchIn, is based on nearfield communication (NFC) [14]. Ahmad et al. There are two primary components to TouchIn: the viewer system and the internet server unit. Students may complete the presence monitoring by touching the NFC viewers with their smartphones or student ID cards that include NFC tags. The ID-based attendance checking system was upgraded by Jacob et al. to include one-time password (OTP) technology. Once the NFC reader detects that a student has entered the class, the server will generate a random password and transmit it to the trainee's mobile device. After gathering the necessary data, students need to log in using the password that came with the app on their phone in order for the attendance tracking to be finished. Typically, trainees are recognised by bio-metric current technologies such as fingerprint recognition, face recognition, and others by bio-metric attendance inspection systems. Muchtar et al. created a fingerprint-based presence tracking system [20]. Each user may be identified on many fingerprint sensors, enhancing the performance of the presence monitoring, by using Arduino and Raspberry Pi to centrally handle the fingerprint data.

PROPOSED SYSTEM

The authors of this work provide AMMoC, a smart method for managing participation. AMMoC doesn't have to collect students' biological traits or use supplementary equipment in the classroom. Specifically, AMMoC uses common verification among students to finish participation monitoring, and it only requires two Android programmes on the mobile devices of instructors and students. AMMoC divides the classroom into several smaller areas and gives students the task of checking the student diversity in each of those areas. A number of group noticing duties make up the verification procedure. Students are given a certain amount of time to transmit their location information to AMMoC when the attendance verification process begins. Using an intelligent search algorithm, AMMoC picks many trainees to do crowd sensing duties, such as submitting the number of students in a given sub-region, after which it utilises this information to do other things. The results of the student-submitted crowd sensing tasks will be used by AMMoC to determine the factual information.

WORKING METHODOLOGY

Distributor of Solutions

To access this component, the Service Provider needs a valid login and password. When he gets there, he'll be able to do things like explore, study, and train data sets, among other things. Using a bar chart, you may see the outcomes of the precision checks and skillful analysis. In addition, you can see the predicted participation type and the percentage of each sort of participant. Get in touch with any distant workers, see the results of the attendance type percentage, and access qualifying data sets.

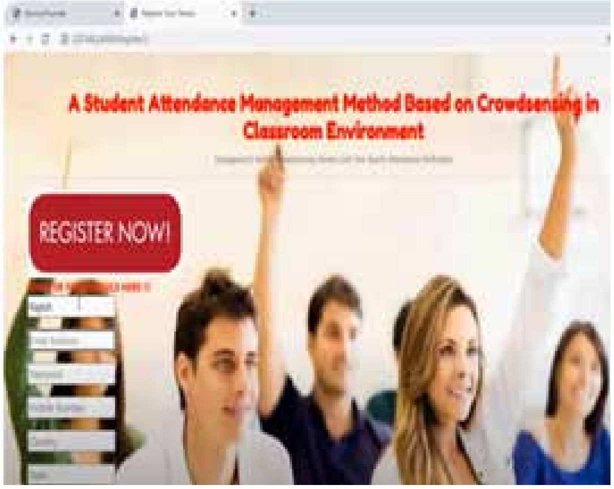
People who use Display and Licence

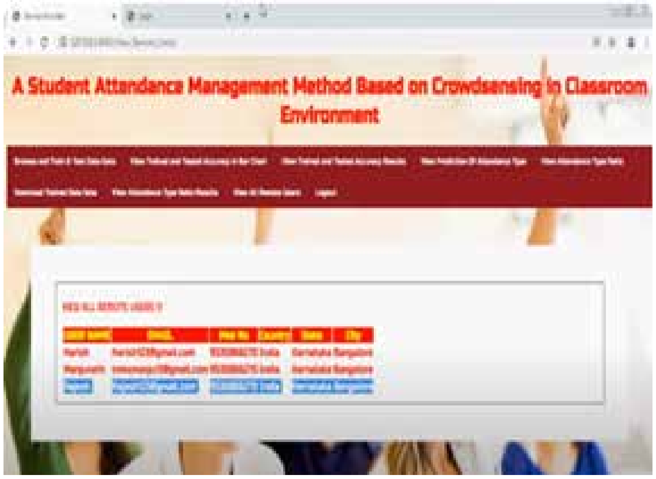
Here the administrator may see a complete roster of registered users. The admin has access to client data like names, email addresses, and addresses, and may also provide access to specific customers.

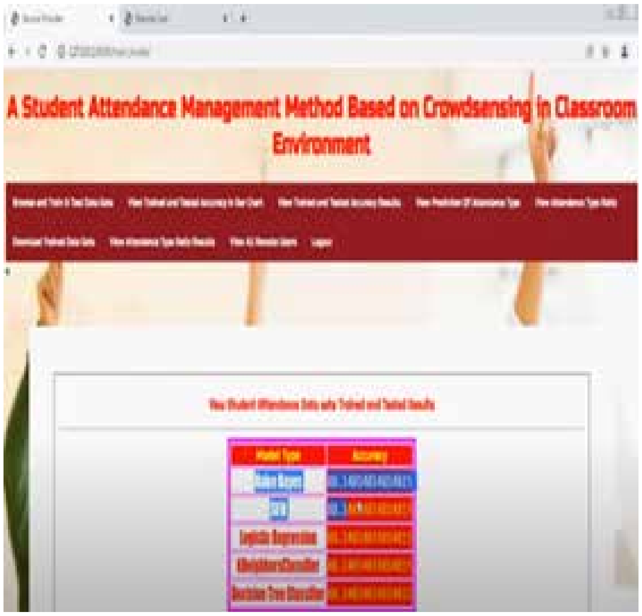
Person Denied Access in a Remote Area

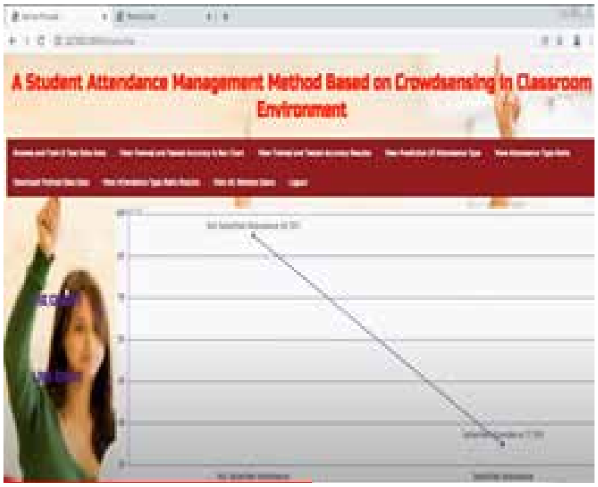
There are a total of n customers in this region. Individuals must register before engaging in any form of activity. Following successful registration, data will be stored in the database. His legal individual name and password will be required for him to visit when he successfully signs up. Features including student attendance form forecast, account viewing, and logout will be accessible to users upon successful login.

A Student Attendance Management Method Based on Crowdsensing in Classroom Environment


A Student Attendance Management Method Based on Crowdsensing in Classroom Environment
REGISTER NOW!


A Student Attendance Management Method Based on Crowdsensing in Classroom Environment


A Student Attendance Management Method Based on Crowdsensing in Classroom Environment
Model Type
Accuracy
SVM
Logistic Regression
Decision Tree Classifier


A Student Attendance Management Method Based on Crowdsensing in Classroom Environment

CONCLUSION

Our proposed intelligent presence administration approach, AMMOC, is detailed in this article. Both the authentication and startup phases are included in AMMOC. During the startup step, all students are required to provide their place details. During authentication, AMMOC optimises group sensing tasks first, and the MCTS algorithm selects a batch of students to carry out trainee confirmation. With the student number of sub regions supplied by the verifies, AMMOC will check the sent locations for veracity. The AMMOC has the benefits of a short presence monitoring time and great accuracy, according to the experimental findings. As a result, AMMOC can successfully provide class participation monitoring.

We want to transition from analogue to digital attendance tracking in the near future so that we may extend the on-site class participation verification to the online discovery environment. In order to be suitable for applications of numerous understanding scenarios, we also want to achieve continuous non-disturbance presence verification.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled “A STUDENT ATTENDANCE MANAGEMENT METHOD BASED ON CROWDSENSING IN CLASSROOM ENVIRONMENT”, which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department Of Computer Science and

Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. A. Lukkarinen, P. Koivukangas, and T. Seppälä, “Relationship between class attendance and student performance,” Procedia-Social and Behavioral Sciences, vol. 228, no. 16, pp. 341-47, Jun. 2016.
2. V. Kassarning, A. Bjerre-Nielsen, E. Mones, S. Lehmann, and D. D. Lassen, “Class attendance, peer similarity, and academic performance in a large field study,” PloS one, vol. 12, no. 11, pp. e0187078, Nov. 2017.
3. N. K. Balcoh, M. H. Yousaf, W. Ahmad, and M. I. Baig, “Algorithm for efficient attendance management: Face recognition based approach,” International Journal of Computer Science Issues, vol. 9, no. 4, pp. 146, Jul. 2012.
4. S. C. Kohalli, R. Kulkarni, M. Salimath, M. Hegde, and R. Hongal, “Smart Wireless Attendance System,” International Journal of Computer Sciences and Engineering, vol. 4, no. 10, pp. 131-137, Sept. 2016.
5. M. M. Islam, M. K. Hasan, M. M. Billah, and M. M. Yddin, “Development of smartphone-based student attendance system,” in proceedings of 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Dhaka, Bangladesh, 2017, pp. 230-233.
6. S. M. Čisar, P. Printer, V. Vojnić, V. Tumbas, and P. Čisar, “Smartphone application for tracking students’ class attendance,” in proceedings of 2016 IEEE 14th international symposium on intelligent systems and informatics (SISY), Subotica, Serbia, 2016, pp. 227-232.
7. R. Cappelli, M. Ferrara, and D. Maltoni, “Minutia cylinder-code: A new representation and matching technique for fingerprint recognition,” IEEE transactions on pattern analysis and machine intelligence, vol. 32, no. 12, pp. 2128-2141, Mar. 2010.
8. I. Naseem, R. Togneri, and M. Bennamoun, “Linear regression for face recognition,” IEEE transactions on pattern analysis andmachine intelligence, vol. 32, no. 11, pp. 2106-2112, Jul. 2010.
9. A. Boles and P. Rad, “Voice biometrics: Deep learning-based voiceprint authentication system,” in proceedings of 2017 12thSystem of Systems Engineering Conference (SoSE), Waikoloa, HI, USA, 2017, pp. 1-6.
10. E. Harinda and E. Ntagwirumugara, “Security & privacy implications in the placement of biometric-based ID card for Rwanda Universities,” Journal of Information Security, vol. 6, no. 02, pp. 93, Jan. 2015.
11. R. K. Ganti, F. Ye, and H. Lei, “Mobile crowdsensing: current state and future challenges,” IEEE communications Magazine, vol. 49, no. 11, pp. 32-39, Nov. 2011.
12. Z. H. Arif, N. S. Ali, N. A. Zakaria, and M. N. Al-Mhiqani, “Attendance Management System for Educational Sector: Critical Review,” International Journal of Computer Science and Mobile Computing, vol. 7, no. 8, pp. 60-66, Aug. 2018.
13. H. D. Rjeib, N. S. Ali, A. A. Farawn, B. Al-Sadawi, and H. Alsharqi, “Attendance and information system using RFID and web-based application for academic sector,” International Journal of Advanced Computer Science and Applications, vol. 9, no. 1, 2018.
14. B. I. Ahmad, “TouchIn: An NFC supported attendance system in a university environment,” International Journal of Information and Education Technology, vol. 4, no. 5, pp. 448, Jan. 2014.

15..J. Jacob, K. Jha, P. Kotak, and S. Puthran, “Mobile attendance using Near Field Communication and One-Time Password,” in proceedings of 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), Noida, India, 2015.

A Deep Attentive Multimodal Learning Approach for Disaster Identification from Social Media Posts

P. Nithya Sree, P. Raju, K. Varun
B.Tech Student
Department of Information Technology
CMR Technical Campus, Medchal
Hyderabad, Telangana

V. Malsoru
Faculty
Department of Information Technology
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ malsoru.it@cmrtc.ac.in

ABSTRACT

Twitter and other micro blogging platforms have become indispensable for disseminating critical information, especially in the aftermath of both natural and man-made disasters. In order to relay critical information like deaths, facility damages, and urgent needs of impacted people, people often upload multimedia components using photographs and/or videos. Humanitarian organisations may greatly benefit from this data in order to plan an adequate and timely response. The need for an automated method to sort through social media for actionable and non-actionable disaster-related material arises from the difficulty of extracting useful information from massive amounts of communications. Previous work mostly examined textual methodsand/or used standard frequent neural networks (RNNs) or convolutional neural networks (CNNs), which might lead to efficiency degradation in the case of lengthy input sequences, although numerous studies have shown the effectiveness of integrating message and picture components for disaster recognition. Using a combination of visual and linguistic information, this article presents a multi-modal catastrophe detection system that can identify tweets by affixing salient word characteristics with aesthetic purposes. A retrained convolutional neural network (e.g., ResNet50) is used for visual attribute extraction, while a bidirectional long-lasting memory (BiLSTM) coupled with an attention device is employed for textual attribute extraction. A function combination technique and the soft max classifier are then used to accumulate visual and textual functions. The results demonstrate that the proposed multi-modal system outperforms the current baselines, which include both multi-modal and uni-modal models, by around 1% and 7% of performance improvement, respectively.

KEYWORDS: *CNN, LSTM, RNN, Twitter, Social media.*

INTRODUCTION

Systems of social media sites may play an important role in disseminating a large amount of vital information during disasters like earthquakes, floods, and storms [1]. When people utilise these social media platforms, they often establish connections across hierarchies, such as those between individuals, between businesses and the federal government, between neighbourhoods, and between the government and its citizens [2]. Tweets from catastrophe victims often detail the events of the disaster, including casualties, infrastructure damage, and the location of affected areas. Additionally, impacted individuals are pleading for quick assistance via the publication of images, tweets, and videos. Charitable organisations may greatly benefit the harmed people if they evaluate these social media posts and derive practical conclusions in real-time [3]. However, manually analysing and extracting actionable insights from a large volume of crisis-related tweets is an incredibly challenging and time-consuming task.

The nonprofit sector of the IT industry has attempted to address the aforementioned challenge by creating automated systems that can sift through social media posts pertaining to crises and extract relevant information. [4] For instance, researchers have developed classifiers to categorise humanitarian features (e.g., kinds of damage), article informativeness, and event categories (e.g., floods, storms) [5, 6]. Current employment opportunities are severely constrained in two ways, notwithstanding these advancements. Research on disaster response via social media has, up until recently, mostly concentrated on textual or image content assessment independently. On the other hand, new studies show that combining textual and visual information frequently yields better insights into an event and leads to more correct reasonings than just reading the text [7]. The second issue is that lengthier phrases may not be well-represented by the CNN or RNN versions used for message attribute portrayal in the few multi-modal feature-based tasks that have been published so far [7], [8].

A dependable computational approach for identifying disaster-related information via the synergistic integration of visual and textual modalities is our goal in this assignment. Our primary focus is on extracting picture functions using a pre-trained visual model, namely ResNet50. To solve the long-range dependency problem with conventional RNN and CNN architecture, we further extract textual characteristics by combining a focus mechanism with the BiLSTM network. After that, we use Deep degree fusion to combine the two sets of features, and then we classify the provided tweet using the soft max layer. [9] In order to determine the types of damage (such as fire, flood, and framework damage) from a group of images and tweets, we conducted extensive experiments using a multi modal damage dataset. Several baselines that do not use attention devices or multi modal functionalities are compared to our designs. The main takeaways from these trials are as follows: (i) using multi modal functions yields much better results than using uni modal features, and (ii) a focus system integrated into an RNN design may significantly outperform a design without such a device. [1]

Our main sources of income from this position are:

- We propose a multi-modal approach that classifies damage-related articles using both visual and textual information, using ResNet50 and BiLSTM recurrent neural networks with an attention mechanism.
- In this study, we evaluate the proposed model in comparison to a set of preexisting multimodal and unimodal (i.e., picture, message) categorization algorithms.
- We conducted an in-depth evaluation of the suggested model using a benchmark dataset and proved that providing emphasis improves system efficiency.
- We use quantitative and qualitative assessment to learn more about the types of mistakes, which will help us improve the model in the future.

RELATED STUDY

Another CNN-based method for categorising tweets about disasters was proposed by Aipe et al. [22], however their focus was on multilabel classification rather than simple binary classification. Similarly, Yu et al. [3] classified the tweets related to several hurricanes into many categories using CNN, logistic regression, and support vector machines. They improved upon SVM and LR with their CNN-based approach. To better capture relationships between word tokens, we examine BiLSTMs with focus systems as an alternative to CNN-based approaches.

Using the Storm Sandy and Boston Marathon fight datasets, Li et al. [4] investigated the possibility of domain name adaptation for assessing catastrophic tweets using the uninformed Bayes classifier. Graf et al. [5] aimed to make the classifier applicable to many types of disasters by focusing on cross-domain categorization. Emotional, nostalgic, and etymological functions were extracted from the damage-related tweets and used by a cross-domain classifier. Message mining and summary approaches have really been the attention of others. One example is the work of Rudra et al. [6], which summarises tweets after classifying them into several scenario courses. In their recommendation of an ESAAWTM system, Cameron et al. [7] sought to alert charitable organisations about disaster situations via the detection of beneficial damage-related Twitter posts. We just concentrate on a multi-class category problem on tweets connected to disasters, in contrast to existing systems that heavily focused on text mining and summarization.

Photos shared on social media platforms may be classified into three types of disasters: severe, medium, and no harm at all, according to a deep convolutional neural network (CNN) architecture developed by Nguyen et al. [9]. A pretrained convolutional neural network (CNN) based structure that can detect catastrophe photographs released on online platforms was also suggested by Alam et al. [10]. In order to identify the fire occurrence, Daly and Thom [31] used pretrained classifiers to filter out flicker photographs. Finally, a method to classify whether the picture shows a fire event or not was devised by Lagerstrom et al. [11]. Chen et al. [12] looked at the photos and texts and used visual attributes and socially relevant contextual attributes (e.g., time of uploading, variety of comments, retweets) to determine catastrophe details, in contrast to these works that developed binary classifiers for categorising catastrophes. Human and environmental harm were the primary foci of the damage discovery investigation by Mouzannar et al. [7]. They used a CNN style for textual characteristics and used the Inception pre-trained design for visual feature extraction.

Similarly, Rizk et al. [35] proposed a multimodal approach to categorise Twitter data according to structural and natural damage types. The tweets were also categorised by Ferda et al. [8] using a multimodal method. The first category was for insightful tasks, such as useful vs. non-informative, and the second was for humanitarian jobs, such as affected persons, rescue volunteering or contribution initiatives, infrastructure and energy damage. To extract the aesthetic and linguistic functions, they used a CNN-based technique. Using the CrisisMMD [14] dataset, Gautam et al. compared unimodal and multimodal methods. Their strategy for integrating the image-tweet sets was the late combination [15] technique. When comparing the works that employ multimodal data to those that use uni-modal data, all of them found that the latter significantly improved performance.

PROPOSED SYSTEM

Our work's main contributions include the following: we provide a multimodal design that manipulates both visual and textual information to detect damage-related postings; this design makes use of ResNet50 and BiLSTM permanent semantic network with interest device. We compare the suggested model's performance to that of many preexisting unimodal (i.e., image, text) and multimodal categorization approaches. We conducted an in-depth study to show how adding focus may improve system efficiency and then tested the proposed model on a benchmark dataset. To get a better grasp of the mistake types that provide guidance for future model improvements, we combine quantitative and qualitative study.

METHODOLOGY

Everything about the Deep Multi-level Attentive network (DMLANet) that has been suggested is laid out here. We provide a high-level description of the proposed network in this section. Our next offering is the visual attention module, which employs both spatial and channel attention to provide noteworthy bi-attentive visual characteristics. Lastly, it delves into a joint attended multimodal learning process that leverages semantic attention to learn a combined representation for textual and visual features. This process involves measuring the semantic closeness of text and visual features, and then using a self-attention mechanism to extract the crucial multimodal features for sentiment classification. To create a bi-attentive visual feature map, the visual attention module uses channel-based attention to improve information-rich channels and spatial or region-based attention to hone in on emotional areas based on attended channels. Semantic attention is used in joint attended multimodal learning to quantify the emotive terms associated with the bi-attentive visual characteristics. After that, we feed the attended word features and the bi-attentive visual features into the self-attention block, and it will automatically choose the most relevant multimodal aspects to highlight.

The majority of the prior work on multimodal sentiment analysis has focused on fusion-based approaches, which integrate data from many sources and input them into a classifier [7]. To get the multimodal sentiment label, some have combined the sentiment predictions from several sources [8]. This process is called late fusion. The fact that these works don't depict the intricate relationship between the modalities is their biggest flaw. The modalities in the network's intermediate layers may be combined using intermediate fusion, which has been used in certain experiments [9]. On the other hand, it calls for meticulous planning and could not work if part of the multimodal material is missing.

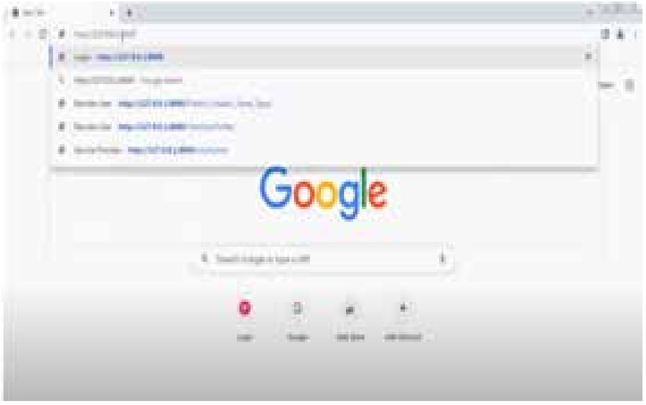


Fig. 1. Web page design

A Deep Attentive Multimodal Learning Approach for Disaster Identification From Social Media Posts
Login Service Provider
admin

Fig.2. Admin page

A Deep Attentive Multimodal Learning Approach for Disaster Identification From Social Media Posts
Login Using Your Account
User Name
Password

Fig.3. Login details.

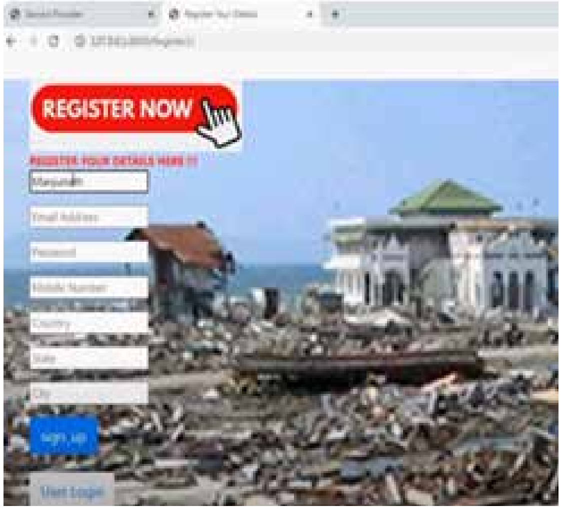


Fig.4. Registration page.

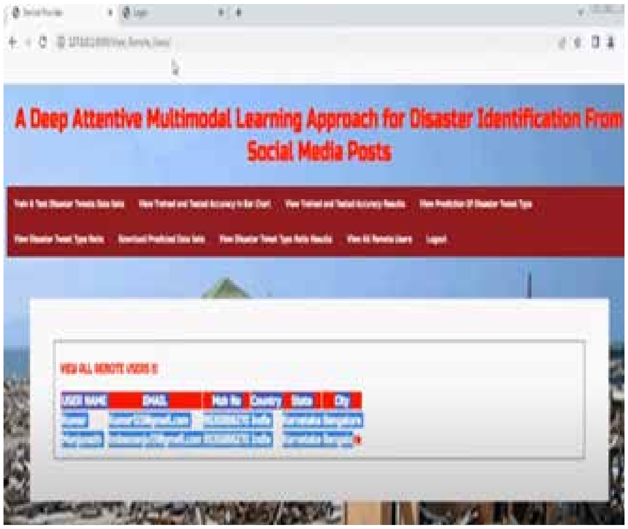

A Deep Attentive Multimodal Learning Approach for Disaster Identification From Social Media Posts
VIEW ALL REMOTE USERS !!!
USER NAME
EMAIL
Mob No
Country
State
City
Logout

Fig.5. users' details

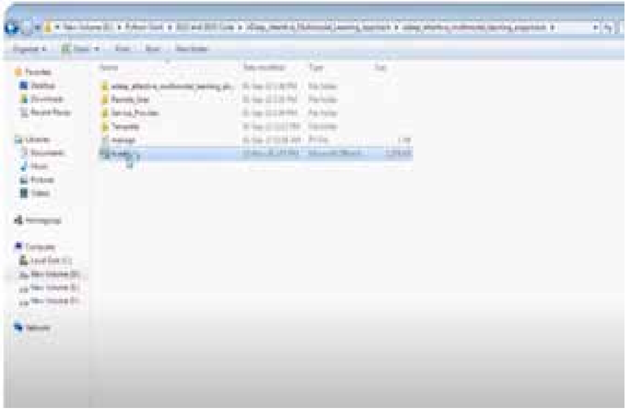
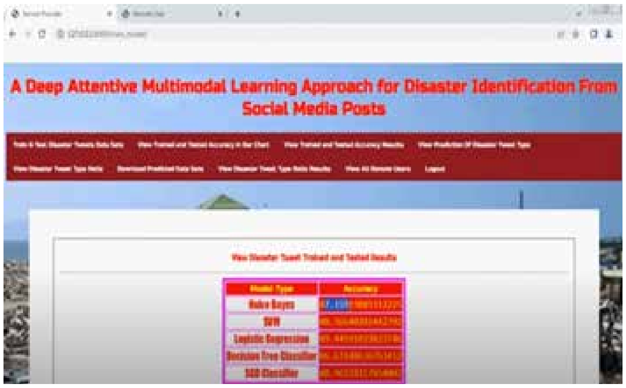


Fig.6. Upload data set.



A Deep Attentive Multimodal Learning Approach for Disaster Identification From Social Media Posts

View Disaster Tweet Trained and Tested Results

Model Type	Accuracy
Naive Bayes	[illegible]
SVM	[illegible]
Logistic Regression	[illegible]
Decision Tree Classifier	[illegible]
SGD Classifier	[illegible]

Fig.7. Accuracy of output.

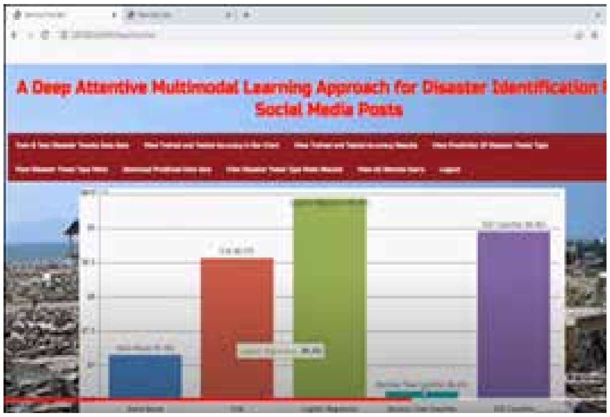


Fig.8. output results

CONCLUSION

In order to classify Twitter posts pertaining to damage, we have presented a multimodal approach that effectively leverages picture and message data. To extract the aesthetic features, we use the pre-trained ResNet version. To extract the tweet functionalities, we use the attention device with a BiLSTM architecture. In order to combine the best parts of both approaches, the early fusion method is used. Also, for the baseline analysis, this study used a plethora of visual and textual approaches, such as VGG19 and Inception, as well as BiLSTM, CNN, BiSTM+CNN, and BiLSTM+ Attention. With a weighted F1-score of 93:21%, the suggested model outperforms both the baseline unimodal (photo/text) and multimodal designs. In addition, the results of the comparison demonstrated that the suggested method outperforms the current state-of-the-art versions by a margin of between one percent and seven percent. It follows that the results validated the efficacy of the proposed method for catastrophe content recognition using multimodal features. Another finding from the error analysis was how difficult it is to distinguish between damaged and non-damaged materials when using a single analysis method. Simultaneously, study of inherent efficiency revealed that adding an interest system improves overall performance.

Despite outperforming unimodal techniques, there remains room for improvement in the suggested method. We want to investigate multitask learning strategies and other multimodal combination techniques to disaster identification in the near future. In addition, we want to use state-of-the-art aesthetic (Vision transformer), textual (BERT, XLM-R), and multimodal (VL-BERT, Visual BERT) transformer designs to better capture the combination of visual and textual features.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled “A DEEP ATTENTIVE MULTIMODAL LEARNING APPROACH FOR DISASTER IDENTIFICATION FROM SOCIAL MEDIA POSTS”, which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department Of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. K. K. Kapoor, K. Tamilmani, N. P. Rana, P. Patil, Y. K. Dwivedi, andS. Nerur, “Advances in social media research: Past, present and future,”Information Systems Frontiers, vol. 20, no. 3, pp. 531–558, 2018.

2. J. Kim and M. Hastak, “Social network analysis: Characteristics of onlinesocial networks after a disaster,” International Journal of Information Management, vol. 38, no. 1, pp. 86–96, 2018.

3. J. Son, H. K. Lee, S. Jin, and J. Lee, “Content features of tweets for effective communication during disasters: A media synchronicity theory perspective,” International Journal of Information Management, vol. 45,pp. 56–68, 2019.

4. A. Elbanna, D. Bunker, L. Levine, and A. Sleigh, “Emergency management in the changing world

of social media: Framing the research agenda with the stakeholders through engaged scholarship," International Journal of Information Management, vol. 47, pp. 112–120, 2019.

5. R. Dubey, A. Gunasekaran, S. J. Childe, D. Roubaud, S. F. Wamba,M. Giannakis, and C. Foropon, "Big data analytics and organizational culture as complements to swift trust and collaborative performance in the humanitarian supply chain," International Journal of Production Economics, vol. 210, pp. 120–136, 2019.

6. S. Akter and S. F. Wamba, "Big data and disaster management: asystematic review and agenda for future research," Annals of Operations Research, vol. 283, no. 1, pp. 939–959, 2019.

7. H. Mouzannar, Y. Rizk, and M. Awad, "Damage identification in social media posts using multimodal deep learning." in ISCRAM, 2018.

8. F. Ofli, F. Alam, and M. Imran, "Analysis of social media data using multimodal deep learning for disaster response," arXiv preprintarXiv:2004.11838, 2020.

9. F. Ofli, P. Meier, M. Imran, C. Castillo, D. Tuia, N. Rey, J. Briant, P. Millet,F. Reinhard, M. Parkan et al., "Combining human computing and machine learning to make sense of big (aerial) data for disaster response," Big data,vol. 4, no. 1, pp. 47–59, 2016.

10. P. Jain, B. Schoen-Phelan, and R. Ross, "Automatic flood detection in sentinei-2 images using deep convolutional neural networks," in Proceedings of the 35th Annual ACM Symposium on Applied Computing,2020, pp. 617–623.

11. A. Kumar, J. P. Singh, Y. K. Dwivedi, and N. P. Rana, "A deep multimodal neural network for informative twitter content classification duringemergencies," Annals of Operations Research, pp. 1–32, 2020.

12. T. Ghosh Mondal, M. R. Jahanshahi, R.-T. Wu, and Z. Y. Wu, "Deep learning-based multi-class damage detection for autonomous post-disaster reconnaissance," Structural Control and Health Monitoring, vol. 27, no. 4,p. e2507, 2020.

13. M. Imran, F. Ofli, D. Caragea, and A. Torralba, "Using ai and social mediamulti modal content for disaster response and management: Opportunities,challenges, and future directions," 2020.

14. L. Dwarakanath, A. Kamsin, R. A. Rasheed, A. Anandhan, and L. Shuib,"Automated machine learning approaches for emergency response and coordination via social media in the aftermath of a disaster: A review,"IEEE Access, vol. 9, pp. 68 917–68 931, 2021.

15. C. Castillo, M. Imran, P. Meier, J. Lucas, J. Srivastava, H. Leson, F. Ofli,and P. Mitra, "Together we stand—supporting decision in crisis response: Artificial intelligence for digital response and micro mappers," by OCHA and partners. Istanbul: Tudor Rose, World Humanitarian Summit, pp. 9395, 2016.

A Novel Time-Aware Food Recommender-System based on Deep Learning and Graph Clustering

Pankitha, Sirisha, Hari Vivek
B.Tech Students
Department of Information Technology
CMR Technical Campus, Medchal
Hyderabad, Telangana

Md. Sajid Pasha
Faculty
Department of Information Technology
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ sajidpasha.it@cmrtc.ac.in

ABSTRACT

People look to food recommender-systems as a trustworthy tool to help them change their eating habits for the better. Using a novel hybrid food recommender-system, this study aims to address the drawbacks of existing systems, such as their inability to take into account food components, time, cool start clients, chilly start food goods, and neighbourhood features. There are two parts to the suggested method: first, recommendations based on food composition, and second, recommendations based on the user. Phase one uses chart clustering, while phase two gathers consumers and food items using a deep-learning based approach. To top it all off, the advice given to the person is of higher quality since a holistic-like approach is used to take into consideration time and user-community related concerns. Using five separate efficiency measures. Accuracy, Remember, F1, AUC, and NDCGwe evaluated our model against a suite of state-of-the-art recommender-systems. Trials using data extracted from "Allrecipes.com" demonstrated that the industrialised food recommendation system performed the best.

KEYWORDS: *F1, NDCG, NDCG, AUC, Recall, Dl, ML*

INTRODUCTION

From leisure activities (e.g., chatting with friends, shopping, finding hotels, and vacation deals) to professional development (e.g., building expert solutions on a web platform), the internet has become an essential component of people's lives and is used for a variety of tasks. [1] Unpredictability and obscurity, caused by the massive amounts of data available to users via their requests, might easily deter them from completing their original requests [2]. Internet search engines have made some efforts to reduce data duplication in recent years, but they still have a long way to go before they can really customise search results and significantly reduce the amount of irrelevant data [3]. Even for individuals with very different interests and talents, several of these systems provide the same outcomes. As one of the most effective online punishment tools, recommender-systems have recently attracted more attention from researchers [4]. Its many potential applications include guiding users to the most suitable service, alleviating information overload, guiding them towards personalised behaviour, and locating their favoured items amid massive amounts of data, among many more. A typical recommender system finds people's interest rates and then suggests products and services based on those rates. Food recommendation is an integral part of many lifestyle apps and services, serving as a tool to encourage users to make changes and embrace a more balanced and healthy lifestyle [5]. A typical goal of food referral is to provide the client with a tailored meal recommendation in terms of recipes, amount of change, and time needed to achieve certain goals, which may be associated with dietary needs or other lifestyle demands [16]. Cultural boundaries and the difficulties in predicting what individuals may want to eat may explain why research in food recommendation has historically received less attention than referral in other leisure and pleasure domains (e.g., music, publications, retail suggestion systems). However, about 60% of all fatalities are attributable to lifestyle-and diet-related health issues, including diabetes and obesity [7]

A lot of people consider the process of making a meal suggestion to be a machine learning problem [8]. To effectively provide a meal suggestion, it is crucial to have a thorough understanding of the customer's culinary preferences. Customers are more likely to follow a proposal if it aligns with their taste preferences, and this is true even when it comes to creating health-oriented food services [9].

There have been a number of recommender-systems developed in the last few decades that attempt to anticipate people's preferences and/or summarise their choices in light of certain goals. Previous food recommender-systems have been rather effective in discovering people's preferences by mapping their interactions with food and meals in the past, but these systems still have some drawbacks.

1) Ingredients in food: Many earlier food recommender-systems neglected food components in favour of relying on consumer evaluations from the past to generate meal suggestions using a collaborative filtering method. This follows from the finding that a private person is more likely to like a certain food item if it contains active elements that they like. The suggestion may be missing some key points because of this. A person may have a preference for hen wings and other similar dishes, yet have a strong aversion to certain flavours that are used in cooking. Because of this, it's possible that collective filtering recommender-systems won't be enough to handle these customers' specific needs.

2) The temporal dimension: The standard model for recommender systems assumes that users would continue to have similar tastes if they have previously made similar purchases. Appropriately, these recommender-systems rely on static data and fail to account for the possibility that people's eating habits, diet plans, or overall way of life may change over time.

3) Cold start products and consumers: Standard food recommender systems based on communal filtering struggle to identify active user next-door neighbours or comparable foods as users usually only rate a small number of items. In light of this, communal filtering-based food referral systems can only provide recommendations to users who have given each dish a sufficient rating. People who are just starting out and have rated a small number of food items are thus ignored. Similarly, a collaborative filtering-based technique will also neglect new food items (food cool starting) that haven't received enough consumer ratings yet.

Community of individuals: Existing recommender-systems also fail to take into account the individual's community or local aspect. Community feature makes it easy to predict the score of hidden food products and the likelihood of a diet's success by extrapolating from the actions of active users in your region. As a general rule, clustering-based architectures can handle community elements. The optimal number of collections and the performance of the similarity measures used are two examples of the many additional issues that have been shown to arise with this approach, all of which are related to the clustering algorithms that are used.

SURVEY OF RESEARCH

As stated in the article "A Hybrid Technique to Suggesting Healthy Food" by Charu Aggarwal, Hui Wang, and Jianfeng Xu (2018), This research introduces a hybrid approach to personalised meal recommendation systems that combines content-based filtering with joint filtering.

Article titled "Suggesting Healthy And Balanced and Personalised Dishes Using Recurring Neural Networks" was written by Shuo Yang, Yanqiao Zhu, and Shijie Sun in 2018. Based on consumers' dietary preferences and health and wellness objectives, this article proposes a recommendation system that use frequent neural networks (RNNs) to provide individualised meal plans. A food recommendation system based upon food picture acknowledgment and nutritional worths is presented in the work of Ahmet Cakir, Ali Selman Aydin, and Mustafa Yildirim (2020) [3]. This research introduces a meal recommender system that can identify foods from photos and provide people with personalised recommendations based on their nutritional content.

Peng Li, Chen Liang, and Zhiyong Cheng (2017) - "Food Recommender Solutions: From Fundamental to Hybrid Approaches" In this study, we survey several methods for food recommender systems, including content-based filtering, collaborative filtering, and hybrid methods.

“A Context-Aware Recommender System for Food and Drink Pairing” (2019) by Seunghwan Kim, Hyeonjin Kim, and Minsoo Lee is referenced in [5]. In order to better combine foods and drinks, this article proposes a context aware recommender system that takes into account the customer’s mood, the event, and their culinary preferences, among other contextual factors.

III EXISTING SYSTEM

1) Nutritional components: Traditional food recommender systems have ignored nutritional components in favour of using users’ past scores to inform meal recommendations made via a community filtering method. This is a direct outcome of the observation that a certain food item is often liked by an individual since it has ingredients that those individuals could enjoy eating. There is a chance that this will ignore some important parts of the suggestion. To illustrate the point, even when someone loves chicken wings, there’s a chance that they’re allergic to certain flavours used in cooking. Consequently, such user’s preferences and limitations may be too much for collaborative filtering recommender-systems to handle.

2) The temporal dimension: The foundation of traditional recommender-systems [19], [26] _ [28] is the assumption that users who have made similar decisions in the past would continue to do so in the future. To be fair, many recommender-systems rely on static data and fail to account for the fact that people’s eating habits, diet plans, and overall way of life might change over time.

3) People with a cold start and foods with a cold start: Traditional food recommender systems based on collaborative filtering struggle to identify active individual neighbours or comparable foods since customers usually only evaluate a small number of items. Customers who have rated items adequately are the only ones that collective filtering-based food suggestions can properly propose. This results in cold start consumers, who have rated a small number of food products, being disregarded. Additionally, this collaborative filtering-based approach does not include brand-new food items (food cold starting) that have not received sufficient user ratings.

Individuals’ neighbourhood: The user’s neighbourhood or community aspect is another worry that current recommender-systems once again disregard. By extrapolating from the neighbourhood activities of active persons, community aspect may be used to estimate the rating of concealed food products and the success probability of a provided diet regimen with no effort at all. In most cases, clustering-based solutions can handle the neighbourhood issue. Nonetheless, evidence suggests that this strategy also has a plethora of other issues that are intrinsic to the clustering techniques used (e.g., optimal number of clusters, efficacy of similar operations).

PROPOSED SYSTEM

1) A food recommender system that takes ingredients into account: Our design differs from conventional collaborative-based food recommender systems in that it combines a user-based stage based on collective filtering with a food-based stage based on content. As a result, it is recommended that they eat a variety of meals that cater to their tastes while also taking into account their past evaluations.

2. A system that recommends meals based on the current time: In this study, we provide a new time-aware similarity metric that accounts for gradual changes in food choices or diet. For scenarios where clients’ ratings or preferences change over time, this makes the proposition a good fit.

The third option is a trust-aware food recommender-system, which is designed to address the issues with cold start individuals and cold start foods that have long plagued conventional collaborative filtering-based feeds. In our proposed revision, we build a customer trust fund network using rely on (fan following) assertions to accurately predict individual ratings.

A crucial role in solving the neighbour selection problem is the reliance on network creation. One way to predict the score of hidden items is to utilise depend on declarations.

In food recommender systems, since the trust of persons is highly correlated with the similarity stage based on user evaluations. This research study handles the data sparsity issue by using expertise stored outside of the customer's geographical neighbourhood. It incorporates the individual's trust fund network and customer ratings-based similarity.

4) A community-aware food recommender-system: Unlike prior work that ignored people's neighbourhoods throughout the food suggestion process, our model explicitly includes these details so that the optimal number of people's collections is quickly determined. Using a visual representation is also covered.

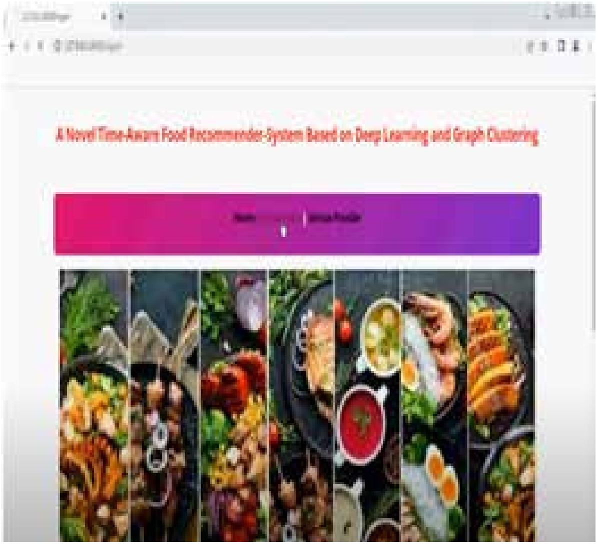
With sparse datasets, the suggested approach uses network-dependent side weights derived according to individual ratings-based similarity.

WORKING METHODOLOGY

Below, we will go over the many aspects of aesthetic food recognition and how they may be used to the process of identifying and evaluating ingredients. In the field of nutrition and the food industry, knowing these characteristics is crucial for creating effective management strategies and preventing negative outcomes. Recognizing the features of visual food identification may greatly aid in comprehending the sources of data on food components and their possible implications. Artificial intelligence methods, such support vector machines and convolutional neural networks (CNNs), are crucial to the visual food recognition data gathering and assessment processes. By eliminating unnecessary functions and constructing classifiers from input attributes, these methods improve ingredient classification. Additionally, using more detailed input characteristics may help reduce processing time, which in turn increases recognition performance. The visual identification category of food goods relies heavily on feature extraction and the development of strong classifiers. We might learn more about this crucial topic with the use of classification models and function extraction. By employing artificial intelligence algorithms, we can fully use visual food identification to enhance our knowledge of food ingredients and how to make it healthier.

Fig. 1. Web page details

As part of our research, we reviewed a number of studies that dealt with picture recognition and ingredient identification in food. We found a few of well-known studies that added a lot of fresh information to the current body of research while we were doing our investigation. Researchers in these works have used a wide range of deep learning models, datasets, and evaluation criteria to determine the efficacy of their methods. We shall summarise the searches of these studies in the following sections, focusing on the methodologies, datasets, and performance indicators that were used. We hope that by doing these searches, we will be able to shed light on the pros and cons of this emerging field by providing a comprehensive evaluation of current approaches for food picture identification and element detection.



A Novel Time-Aware Food Recommender-System Based on Deep Learning and Graph Clustering

Fig. 2. Home page

A Novel Time-Aware Food Recommender-System Based on Deep Learning and Graph Clustering
Login

Fig. 3. Admin login page

Recommender system, food recommendation, healthcare, deep learning, graph clustering
Login Using Your Account
User Name
Password

Fig. 4. User registration page

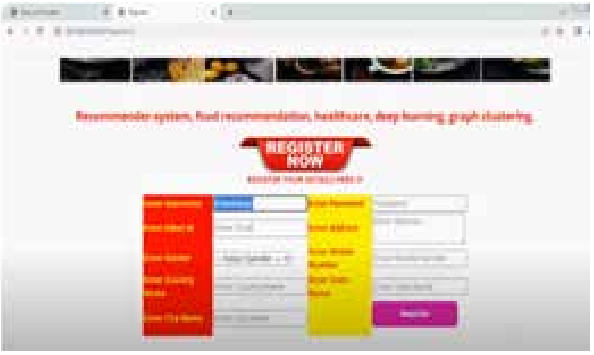

Recommender system, food recommendation, healthcare, deep learning, graph clustering
REGISTER NOW

Fig. 5. User details page

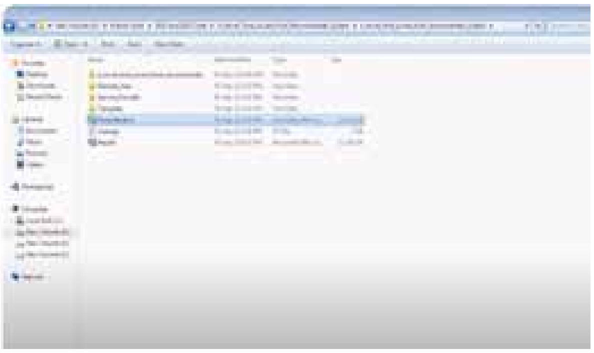


Fig. 6. Upload dataset

A Novel Time-Aware Food Recommender-System Based on Deep Learning and Graph Clustering

Fig. 7. Accuracy details

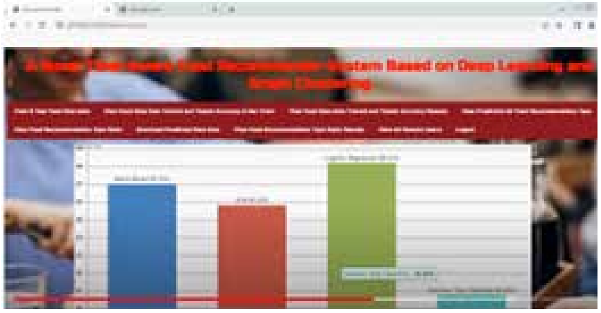

A Novel Time-Aware Food Recommender-System Based on Deep Learning and Graph Clustering

Fig. 8. Output results

CONCLUSION

Recommender systems that choose products that are fairly suitable to people's wants are becoming more common as the Net grows in popularity and the number of internet users increases. Food recommender systems are essential to many lifestyle services and rely on them for a variety of lifestyle applications. In order to address the drawbacks of earlier food recommender-systems—such as ignoring time stamps, cold start customers, cool start items, and particular communities—this study establishes a novel hybrid food recommender-system. The proposed method incorporates user-and content-based designs, time data, count on network, and consumer regions to simultaneously handle all four concerns and enhance the recommender-system's final accuracy. Suggestion based on food composition and recommendation based on user are the two parts of the suggested method. Phase one employs graph clustering, while phase two employs a deep-learning based technique to cluster consumers and food items. Precision, Remember, F1, AUC, and NDCG are five metrics that have been used to compare the design to the latest proposed food recommender-system, which consists of LDA, HAFR, and FGCN approaches.

According to the findings of the experiments, the produced food recommender-system outperformed the sophisticated food recommender-systems by a clear margin and attained the most effective efficiency. Our long-term goal is to enhance the meal suggestion's final performance by including people' demographic data (such as gender, age, weight, height, location, and culture) into the suggestion's structure. Furthermore, a healthy diet might lessen the severity of symptoms associated with non-infectious diseases. We want to include the nutritional value of each item as an additional metric in our future works so that we may tailor meal recommendations to each individual's health status and medical conditions.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "A NOVEL TIME-AWARE FOOD RECOMMENDER-SYSTEM BASED ON DEEP LEARNING AND GRAPH CLUSTERING", which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. S. Wang, L. Cao, Y. Wang, Q. Z. Sheng, M. A. Orgun, and D. Lian, ``A survey on session-based recommender systems,'' ACM Comput. Surv., vol. 54, no. 7, pp. 138, Sep. 2022.

2. P. Wang, Y. Wang, L. Y. Zhang, and H. Zhu, ``An effective and efcient fuzzy approach for managing natural noise in recommender systems,'' Inf. Sci., vol. 570, pp. 623637, Sep. 2021.

3. A. D. Viniski, J. P. Barddal, A. D. S. Britto, Jr., F. Enembreck, and H. V. A. D. Campos, ``Acase study of batch and incremental recommender systems in supermarket data under concept drifts and cold start,'' Expert Syst. Appl., vol. 176, Aug. 2021, Art. no. 114890.

4. X. Yu, Y. Chu, F. Jiang, Y. Guo, and D. Gong, ``SVMs classication based two-side cross domain collaborative ltering by inferring intrinsic user and item features,'' Knowl.-Based Syst., vol. 141, pp. 8091, Feb. 2018.

5. N. Hazrati and F. Ricci, ``Recommender systems effect on the evolution of users' choices distribution,'' Inf. Process. Manage., vol. 59, no. 1, Jan. 2022, Art. no. 102766. [6] L. Xie, Z. Hu, X. Cai, W. Zhang, and J. Chen, ``Explainable recommendation based on knowledge graph and multi-objective optimization,'' Complex Intell. Syst., vol. 7, no. 3, pp. 12411252, Jun. 2021.

7. M. Wasid and R. Ali, ``A frequency count approach to multi-criteria recommender system based on criteria weighting using particle swarm optimization,'' Appl. Soft Comput., vol. 112, Nov. 2021, Art. no. 107782.

8. S. Forouzandeh, M. Rostami, and K. Berahmand, ``A hybrid method for recommendation systems based on tourism with an evolutionary algorithm and topsis model,'' Fuzzy Inf. Eng., vol. 14, no. 1, pp. 2650, 2022.

9. S. Forouzandeh, M. Rostami, and K. Berahmand, ``Presentation a trust Walker for rating prediction in recommender system with biased random walk: Effects of H-index centrality, similarity in items and friends,'' Eng. Appl. Artif. Intell., vol. 104, Sep. 2021, Art. no. 104325.

10. T. N. T. Tran, A. Felfernig, and N. Tintarev, ``Humanized recommender systems: State-of-the-art and research issues,'' ACM Trans. Interact. Intell. Syst., vol. 11, no. 2, pp. 141, Jul. 2021.

11. M. Slokom, A. Hanjalic, and M. Larson, ``Towards user-oriented privacy for recommender system data: A personalization-based approach to gender obfuscation for user proles,'' Inf. Process. Manage., vol. 58, no. 6, Nov. 2021, Art. no. 102722.

12. X.Yu, F. Jiang, J. Du, and D. Gong, ``Across-domain collaborative ltering algorithm with expanding user and item features via the latent factor space of auxiliary domains,'' Pattern Recognit., vol. 94, pp. 96109, Oct. 2019.

13. M. Ge, F. Ricci, and D. Massimo, ``Health-aware food recommender system,'' in Proc. 9th ACMConf. Recommender Syst., Sep. 2015, pp. 333334.

14. D. Bianchini, V. De Antonellis, N. De Franceschi, and M. Melchiori, ``PREFer: A prescription-based food recommender system,'' Comput. Standards Interfaces, vol. 54, pp. 6475, Nov. 2017.

15. M. B. Vivek, N. Manju, and M. N. Vijay, ``Machine learning based food recipe recommendation system,'' in Proc. Int. Conf. Cogn. Recognit. Singapore: Springer, 2018, pp. 1119.

Prediction of used Car Prices using Artificial Neural Networks and Machine Learning

Ch. Vishnu Vardhan, M. Shreya, Md. Rizwan
B.Tech Students
Department of Information Technology
CMR Technical Campus, Medchal
Hyderabad, Telangana

L. V. Ramesh
Faculty
Department of Information Technology
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ rameshlv.it@cmrtc.ac.in

ABSTRACT

Over the last decade, the number of cars driving on Mauritius's roads has actually been steadily increasing, rising by 5% annually. The National Transport Authority received 173,954 vehicle registrations in 2014. As a result, 1 in 6 Mauritian adults own a vehicle, with many of those vehicles being pre-owned or replaced vehicles. In this study, we want to find out whether it's possible to utilise artificial semantic networks to forecast the value of used vehicles. As a result, four separate maker learning models were given data pertaining to 200 automobiles culled from diverse sources. Results were somewhat better using support vector equipment regression compared to semantic networks and straight regression, according to our findings. However, for more expensive vehicles in particular, several of the predicted values are far off from the actual pricing. Consequently, more investigations using a bigger dataset are necessary, as is a great deal of trial and error with other types of networks and frameworks, to get much improved predictions.

KEYWORDS: *ML, DL, SVM, Car price, Linear regression.*

INTRODUCTION

Obtained from the National Transportation Authority (2014), the number of cars increased by 254% from 2003 (68,524) to 2014 (173, 954), as shown in Number 1. Given that new cars and trucks make up such a small percentage of the overall number of vehicles sold each year, it is reasonable to assume that the sale of used imported (refurbished) automobiles and pre-owned previously owned vehicles has eventually increased. When purchasing a new automobile in Mauritius, most buyers are also interested in finding out how much their vehicle will be worth when they're ready to sell it on the used car market. A number of factors influence the prediction of used car prices. Year of production, manufacturer, model, gas economy, horsepower, and nation of origin are the most crucial ones. Additional considerations include the following: fuel type and quantity per use, braking system type, acceleration, interior style, vehicle condition, number of cylinders (in cubic centimetres), vehicle size, number of doors, vehicle weight, customer reviews, paint colour and type, gearbox type, sports car status, audio system, planetary wheels, power steering, air conditioning, GPS navigation, safety and security index, and so on. Whether the vehicle has been in any serious accidents and the identities of its former owners are two of the distinctive factors often considered in a Mauritius setting. Predicting the rate of used cars is, therefore, a quite commendable company. This research will investigate the feasibility of using semantic networks for used car rate prediction. Additional approaches, such as support vector regression and straight regression, will also be used to compare the results. In accordance with, this paper continues. Several service semantic networks and cost forecasts have been consolidated in this system. In this system, the procedure and the data gathering are detailed. The technology calculates what used car prices are likely to be. A conclusion and some suggestions for further research round out the report.

SURVEY OF RESEARCH

[1] Predicting the value of pre-owned vehicles has been the subject of several studies. It is common practice for researchers to use historical data to estimate product prices. In order to forecast the pricing of used automobiles in Mauritius, Pudaruth used a number of methods, including decision trees, k-nearest neighbours, Ignorant Bayes, and direct regression. The cars in question were not brand new. When several methods' predictions were compared, it became clear that their prices are rather close to one another. Decision trees and the Naive Bayes method were both shown to fail miserably when it came to predicting numerical values. The small sample size does not provide excellent prediction accuracy, according to Pudaruth's study.

[2] in A multivariate regression model that helps with numerical value classification and forecasting was presented by Kuiper, S. (2008). In it, the author shows how to predict the GM truck rate in 2005 using this multivariate regression model. Automobile rate forecasting does not need specialised knowledge. So, it is possible to predict prices using the data that is readily accessible online. In the same vein as the article's car rate forecast, the author offered variable selection techniques to help identify which variables would work best in the model.

As a method for predicting the prices of secondhand vehicles using Random Forest, Pal et al. [3] found in 2019. With an accuracy of 83.62% for test data and 95% for train-data, the Kaggle information established was utilised to forecast used-car prices in this research. After eliminating outliers and uninteresting characteristics from the dataset, the most relevant functions used for this prediction were price, km, brand name, and truck type. As a novel implementation, Random Woodland outperformed earlier work with similar datasets in terms of accuracy.

[4] The need to create a model to predict the cost of used cars in Bosnia and Herzegovina is shown by Gegic, E. et al. (2019). They used AI methods including made-up semantic networks, assistance vector devices, and made-up woods. Still, all of those methods were used in tandem. To get the data for the projection, we used a web scraper that was made in PHP to collect information from the autopijaca.ba website. In order to determine which method was the best match for the given data, we compared the individual algorithms' results. The final prediction model was a Java programme. The design was further evaluated using examination data, which confirmed its accuracy of 87.38%. In 2019, Dholiya et al. showcased a method for reselling cars that relies on equipment learning.

Giving the person a realistic idea of how much the vehicle may cost is the main objective of the method developed by Dholiya, M., et al. The system, which is an online application, may also provide the user with a list of options for different types of cars based on the details of the vehicle they are attempting to discover. By doing so, it helps provide the buyer or seller with useful information upon which to make a decision. The system trains utilising past data acquired over a long period of time, and it uses the multiple direct regression technique to create projections. At first, the KDD (Knowledge Exploration in Databases) method was used to gather the raw data. Afterwards, it cleaned and preprocessed the data in an effort to find meaningful patterns, which it subsequently used to draw conclusions.

PROPOSED SYSTEM

Because, like the prices of other items, the prices of autos also change with time, data for this research has been sourced from a variety of automotive websites as well as the small ads sections seen in regular newspapers. One hundred and twenty records were gathered. Old vehicles' production year (YEAR), make (MAKE), engine capacity (ENGINE) measured in cubic centimetres, paint type (typical or metallic), gearbox type (guidebook or automatic), mileage (GAS MILEAGE) and price (RATE) in Mauritian rupees are all part of the detailed information. A large number of tests were conducted to determine the optimal semantic network criteria and network topology. Out of all the semantic network frameworks that were tested, we found that a network with just one hidden layer and two nodes made the most insignificant error. However, out of these four approaches, k-Nearest Neighbour had the lowest accuracy, while Support Vector Regression and a multi layer understanding with back-propagation made somewhat better predictions than linear regression. A cross validation value of ten folds was used to conduct all trials.

WORKING METHODOLOGY

This part contains the research study methodology. In order to compile the vehicle dataset for this analysis, olx.com was used. Every vehicle was filmed along with its make, version, vendor, mileage, year of production, fuel type, and price. Table 1 displays a sample of the collected data. There will probably be a lot of used car data in these databases, so they will need some engineering and tweaking. Excluding duplicate data is a good first step since they could affect the model's output. Rate predictions are greatly aided by power and engine. Additionally, new_price is a very good cost forecaster. In predictive statistics and AI, top traits with a high connection coefficient have a stronger impact on the prediction variable. However, this is not always the case. The correlation coefficient, as its name implies, is a statistical measure that describes the relationship between two or more variables. Always between 1 and -1 (positive to negative), with 0 indicating complete disorganisation, is the range of values for the variation of the correlation coefficient between 2 criteria.

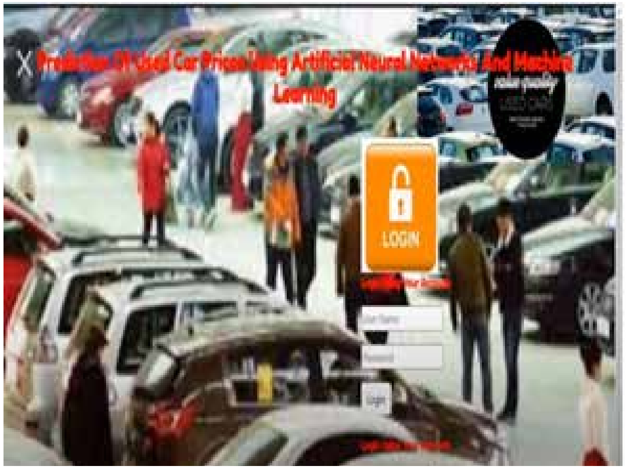

X
Prediction Of Used Car Prices Using Artificial Neural Networks And Machine Learning
LOGIN
User Name
Password
Login

Fig. 1. Home page

This task makes use of the Scikit-learn machine learning package to build a number of AI formulae. We use the same training data to train each design, and we utilise the same screening data to assess it. The findings are defined and compared in the section that follows. For predictable variables, the regression-based method in monitored machine learning is reliable. When X is the independent variable and Y is the dependent variable, it is clear that a single straight regression model can predict Y. In order to predict Y's future value, the design will use the Y-intercept, slope, and noise of the regression line.

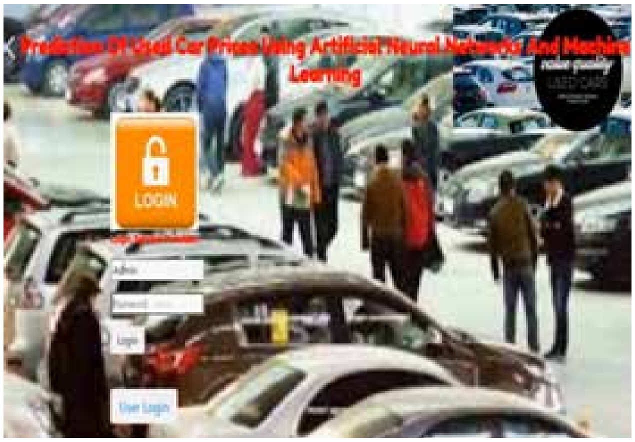

Prediction Of Used Car Prices Using Artificial Neural Networks And Machine Learning
LOGIN
Admin
Password
Login
User Login

Fig.2. Admin login page.

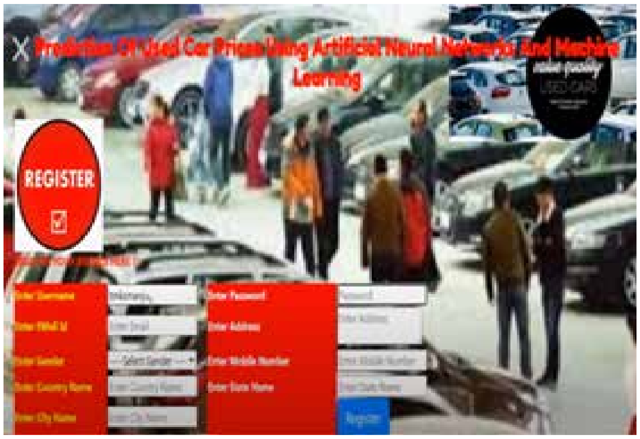


Fig.3. user registration page.

Fig.4. Login page sign in page.

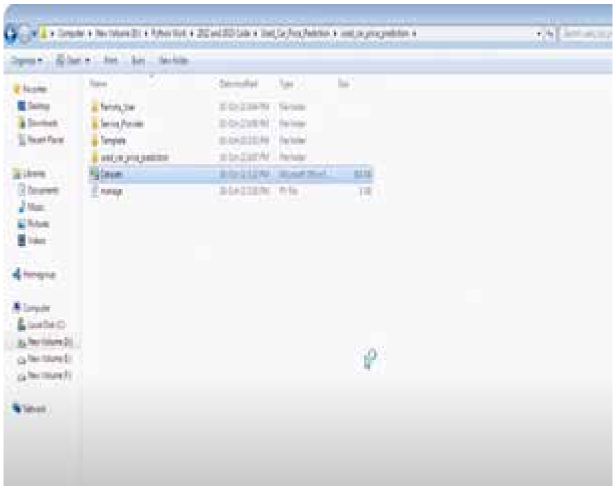
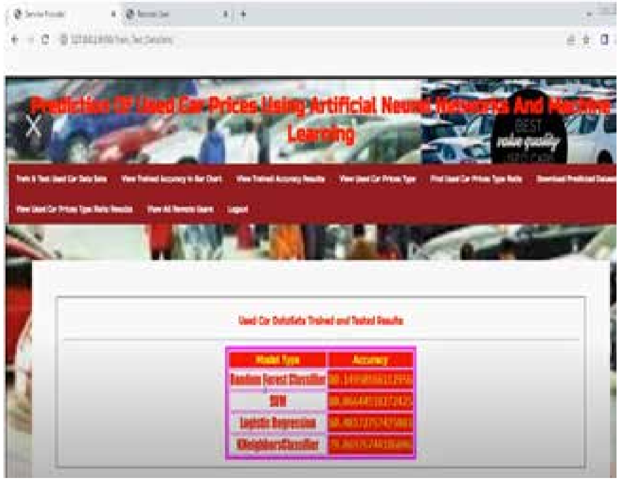


Fig.5. Upload dataset page.

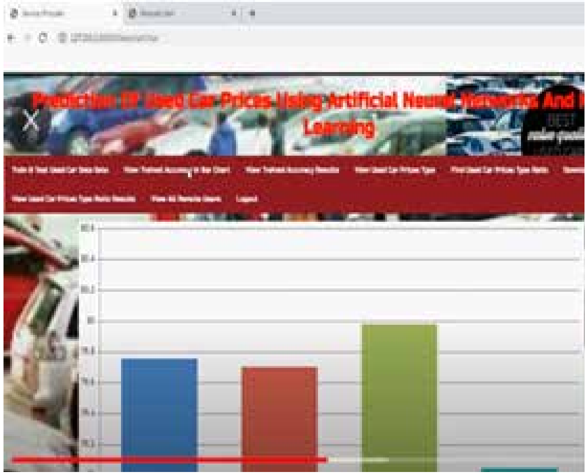


Prediction Of Used Car Prices Using Artificial Neural Networks And Machine Learning

Used Car DataSets Trained and Tested Results

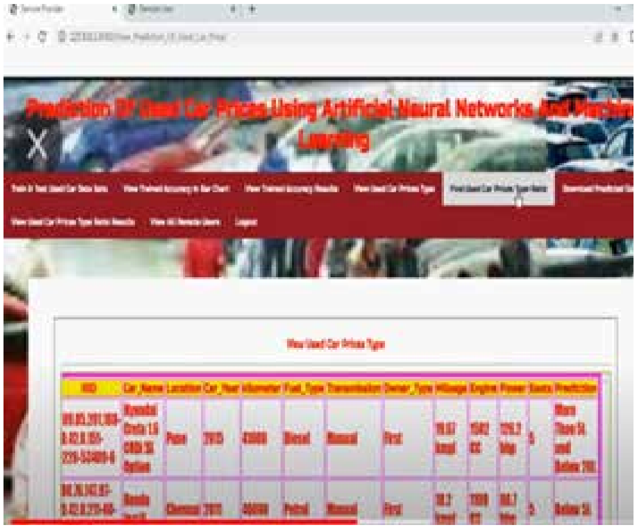
Model Type	Accuracy
Random Forest Classifier	[illegible]
SVM	[illegible]
Logistic Regression	[illegible]
KNeighborsClassifier	[illegible]

Fig.6. Accuracy details.



Prediction Of Used Car Prices Using Artificial Neural Networks And Machine Learning

Fig.7 . Graph results.



Prediction Of Used Car Prices Using Artificial Neural Networks And Machine Learning

View Used Car Prices Type

RID	Car_Name	Location	Car_Year	Kilometer	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	Prediction
[illegible]	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67 kmpl	1582 CC	126.2 bhp	5	More Than 5L and Below 10L
[illegible]	Honda [illegible]	Chennai	2011	46000	Petrol	Manual	First	18.2 kmpl	1199 CC	88.7 bhp	5	Below 5L

Fig.8. Output results.

CONCLUSION

Predicting the rate of used, refurbished, and second-hand automobile sales in Mauritius was the primary objective of this work. The fact that the car market has been steadily increasing by around 5% over the last decade is evidence of the significant demand for cars among Mauritian citizens. There are a plethora of automotive websites on Mauritius, but none of them provide a tool to estimate the value of pre-owned vehicles according to their specifications. We employed the cross-validation method with a ten-fold increase on our dataset of 200 documents. Using four separate machine learning methods, we can predict the price of used cars based on their make, year, colour, gearbox type, engine capacity, and mileage. With each of the four approaches, we saw a considerable drop in the ordinary residual value. Consequently, we draw the conclusion that predicting the rate of used vehicles is a highly risky but potentially lucrative endeavour. Dealers and owners of automobiles looking to get an idea of their worth will find this approach incredibly useful. We want to use a wider variety of machine learning formulae for future predictions, as well as to gather additional data and characteristics.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "PREDICTION OF USED CAR PRICES USING ARTIFICIAL NEURAL NETWORKS

AND MACHINE LEARNING", which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department Of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. NATIONAL TRANSPORT AUTHORITY. 2015. Available at: http://nta.govmu.org/English/Statistics/Pages/Arch ives.aspx. [Accessed 24 April 2015].
2. Bharambe, M. M. P., and Dharmadhikari, S. C. (2015) "Stock Market Analysis Based on Artificial Neural Network with Big data". Fourth Post Graduate Conference, 24-25th March 2015, Pune, India.
3. Pudaruth, S. (2014) "Predicting the Price of Used Cars using Machine Learning Techniques". International Journal of Information & Computation Technology, Vol. 4, No. 7, pp.753- 764.
4. Jassibi, J., Alborzi, M. and Ghoreshi, F. (2011) "Car Paint Thickness Control using Artificial Neural Network and Regression Method". Journal of Industrial Engineering International, Vol. 7, No. 14, pp. 1-6, November 2010
5. Ahangar, R. G., Mahmood and Y., Hassen P.M. (2010) "The Comparison of Methods, Artificial Neural Network with Linear Regression using Specific Variables for Prediction Stock Prices in Tehran Stock Exchange". International Journal of Computer Science and Information Security, Vol.7, No. 2, pp. 38-46.
6. Listiani, M. (2009) "Support Vector Regression Analysis for Price Prediction in a Car Leasing Application". Thesis (MSc). Hamburg University of Technology.
7. Iseri, A. and Karlik, B. (2009) "An Artificial Neural Network Approach on Automobile Pricing". Expert Systems with Application: ScienceDirect Journal of Informatics, Vol. 36, pp. 155-2160, March 2009.
8. Yeo, C. A. (2009) "Neural Networks for Automobile Insurance Pricing". Encyclopedia of Information Science and Technology, 2nd Edition, pp. 2794-2800, Australia.
9. Doganis, P., Alexandridis, A., Patrinos, P. and Sarimveis, H. (2006) "Time Series Sales Forecasting for Short Shelf-life Food Products Based on Artificial Neural Networks and Evolutionary Computing". Journal of Food Engineering, Vol. 75, pp. 196–204.
10. Rose, D. (2003) "Predicting Car Production using a Neural Network Technical Paper- Vetronics (Inhouse)". Thesis, U.S. Army Tank Automotive Research, Development and Engineering Center (TARDEC).
11. LEXPRESS.MU ONLINE. 2014. [Online] Available at: http://www.lexpress.mu/ [Accessed 23 September 2014].
12. LE DEFI MEDIA GROUP. 2014. [Online] Available at: http://www.defimedia.info/ [Accessed 23 September 2014].
13. He, Q. (1999) "Neural Network and its Application in IR". Thesis (BSc). University of Illinois.
14. Cheng, B. and Titterington, D. M. (1994). "Neural Networks: A Review from a Statistical Perspective". Statistical Science, Vol. 9, pp. 2-54.
15. Anyaeche, C. O. (2013). "Predicting Performance Measures using Linear Regression and Neural Network: A Comparison". African Journal of Engineering Research, Vol. 1, No. 3, pp. 84-89.

Phishing URL Detection a Real-Case Scenario through Login URLS

G. Meghana, G. Sai Prakash, K. Pravalika
B.Tech Students
Department of Information Technology
CMR Technical Campus, Medchal
Hyderabad, Telangana

Ravi Regulagadda
Faculty
Department of Information Technology
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ ravi.it@cmrtc.ac.in

ABSTRACT

In phishing, a social engineering cyberattack, fraudsters use a login form to deceive users into giving up sensitive information, which is then sent to a malicious web server. In this research, we compare and contrast deep learning and machine learning approaches to provide a solution that can evaluate URLs and identify phishing sites. The real course in many current state-of-the-art phishing detection solutions consists of homepages without login fields. Instead, we use login page links in both classes because we think it's more representative of a real-world situation and since we demonstrate that current tactics have a high false-positive rate when tested with links from genuine login sites. Furthermore, by training a base design using historical information and then testing it with up-to-date URLs, we can show how designs lose accuracy with time. Furthermore, in order to discover the different strategies used by phishers in their projects, we conduct a frequency assessment across active phishing domains. In order to back up these claims, we have created a new dataset called Phishing Index Login URL (PILU-90K). This dataset contains 30,000 phishing links and 60,000 legitimate links, including login and index websites. Lastly, we showcase a Logistic Regression model that achieves 96:50% accuracy on the given login link dataset when combined with Term Frequency - Inverse Record Regularity (TF-IDF) attribute extraction.

KEYWORDS: *URL, SVM, Light GBM, Cyber security, Phishing website.*

INTRODUCTION

The web's process has been greatly enhanced in recent years, simplifying and improving our lives. Communication, learning and education, conditioning organisations, and business are all heavily reliant on it. For individual, group, business, and societal progress, the internet is a treasure trove of useful information, data, and facts. Thanks to the internet, we can access dynamic information at any time, from any location in the globe, and it's simple to provide a wide variety of solutions online. To commit phishing, one must use the same email address, domain name, or malicious website to trick the victim into divulging sensitive information such as social security numbers, bank account details, birthdays, credit card numbers, or personal identification numbers (PINs). Internet addicts all across the globe are susceptible to phishing attacks. Companies and individuals have lost a tonne of money and sensitive information due to phishing attacks. It turns out that finding the phishing attack is no easy feat. This kind of attack might use creative methods to trick even the most vigilant users, including swapping out a few link characters with similar Unicode characters. One potential drawback is that it may be provided in sloppy forms, such as using an IP address instead of the domain. With the application of artificial intelligence and data mining techniques, some literary works [5–9] successfully overcame the phishing assault detection challenge, reaching a satisfying acknowledgment rate of 99.62%. However, due to their complex computers and high battery consumption, those systems aren't ideal for smart devices and other embedded devices. This is because they rely on photo processing to accomplish recognition, which requires HTML web pages or at least HTML links, tags, and website JavaScript components. In contrast to competing systems, our recognition algorithm only needs six characteristics that have been removed from the link as input, which means it uses less CPU and memory. This study will first provide a brief overview of the relevant field studies, and then describe in detail the link properties that our system uses for the acknowledgment. Then, in the practical section, we will assess the suggested system while providing the findings we acquired after defining our recognition system. Finally, we will clarify the benefits and consequences of our solution compared to the phishing assault.

Project Objective

Phishers aim to steal sensitive information such as login credentials and bank account data. Now more than ever, cyber security professionals are searching for reliable and ongoing discovery ways to identify phishing websites. This study discusses device detecting technology that can identify phishing links and remove them, as well as evaluate various characteristics of legitimate and malicious links. Phishing sites are identified using algorithms from Support Vector Machines, Random Forest, and Choice Tree. By comparing the accuracy rate, mistake positive rate, and mistake negative price of each algorithm, this research aims to discover phishing URLs and the limit to the optimum maker figuring out formula.

LITERATURE SURVEY

The process of the internet has been much improved in the last several years, which has made our lives easier and better. A lot of things rely on it, including communication, education, conditioning organisations, and business. The internet is a treasure trove of useful information for personal, group, service, and society development. Because of the internet, we have constant, global access to dynamic information, and it's simple to provide a plethora of choices to our customers. In order to commit phishing, one must trick the victim into giving sensitive information like SSNs, bank account details, birthdays, credit card numbers, or personal identification numbers (PINs) by using the same email address, domain, or malicious website. Phishing attempts might target internet junkies worldwide. A lot of money and personal information has been stolen due to phishing attacks. Discovering the phishing attack is clearly not a walk in the park. A few web link personalities might be swapped out for identical Unicode characters as part of this sort of attack's unique approaches to fool even the most vigilant persons. A possible downside is that it might be provided in a careless manner, such using an IP address instead of the domain. Some literature [5–9] successfully overcame the phishing attack detection problem by using expert system and data mining techniques, and they achieved a satisfying recommendation rate of 99.62%. However, such systems aren't ideal for smart devices and other embedded devices because to their facility computing systems and high battery consumption. This is due to the fact that in order for them to accomplish recognition, they depend on image processing, which necessitates HTML web pages or at least certain HTML web connections, tags, and JavaScript components for websites. Our recognition technique utilises far less CPU and memory than competing systems since it just requires 6 attributes that have been removed from the link as input. First, this study will provide a brief overview of the relevant field investigations, and then it will describe in detail the link structures that our system uses for the recognition. After we've defined our acknowledgment system, we'll look at the proposed system in the functional area and provide the results of our searches. Finally, in contrast to the phishing attack, we will undoubtedly outline the advantages and disadvantages of our approach.

Job Objective

Passwords and bank account details are among the sensitive information that phishers try to steal. More than ever before, professionals in the field of cyber safety and security are on the lookout for reliable and ongoing discovery techniques to identify phishing websites. This research study delves into the latest advancements in device detection technology that can identify and eliminate phishing connections. It also examines the different traits that distinguish between trustworthy and hazardous online links. Use of Support Vector Machine, Random Forest, and Selection Tree formulae allows for the recognition of phishing sites. Aiming to identify phishing Links and the ideal maker figuring out formula limit, this study compares each formula's accuracy rate, error positive rate, and mistake negative cost.

EXISTING SYSTEM

In phishing, an online scammer poses as a trustworthy entity in order to trick unsuspecting victims into divulging sensitive information. Your personal information will be stolen or your machine will be infected with a virus the moment you click on a link or open a file in the email. Typically, massive spam campaigns that arbitrarily targeted large groups of individuals were the means by which phishing attempts were carried out. Purging as many individuals as possible into opening infected papers or clicking on infected links was the objective. This sort of attack may be detected using a variety of methods. Artificial intelligence is one of the methods. By entering the URLs that the user has obtained into the equipment finding design, the formula will refine the input and present the result indicating whether the URL is legitimate or not. Multiple machine learning algorithms may be used to detect these connections, including support vector machines, neural networks, random forests, choice trees, XG increase, and many more. Care for Random Woodland and Choice Tree classifiers are handled by the suggested approach. The proposed method achieved an 87.0% success rate for identifying phishing URLs and an 82.4% success rate for identifying legitimate URLs using Random Woodland and decision tree classifiers, respectively.

PROPOSED SYSTEM

The sophistication and volume of phishing attacks have both increased in recent times. This has led to corresponding advancements in the techniques used to evade phishing attempts, which provide formidable challenges to the privacy and security of smart device users. This paper proposes a machine learning-based approach to detect phishing websites and maintain the security of smart devices by using LightGBM and domain functions. A large number of domain-name-generation methods maintain consistency in domain name properties, sometimes called proportion. Attributes of the domain of the provided website, including character-level functionalities and information on the domain name, are first removed using the proposed discovery design. In order to make the version more accurate, the functions are filtered before being utilised for categorization. By combining two types of features for training, the suggested detection model outperforms the one that uses just one kind of feature, according to experimental comparison results. Also, the proposed technique is well-suited to the real-time detection of many phishing websites and has higher detection accuracy than competing methods.

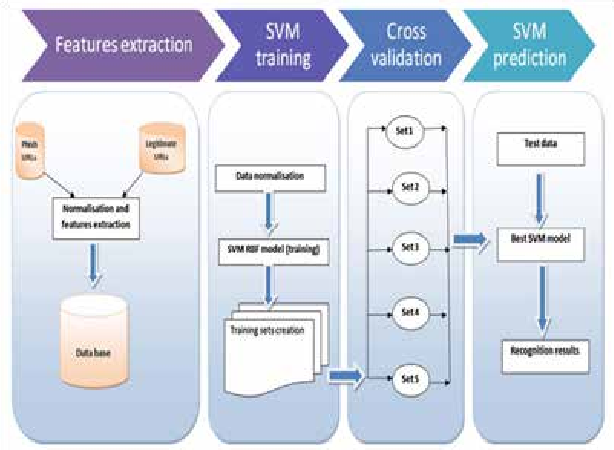


Fig. 1. Phishing website process

METHODOLOGY

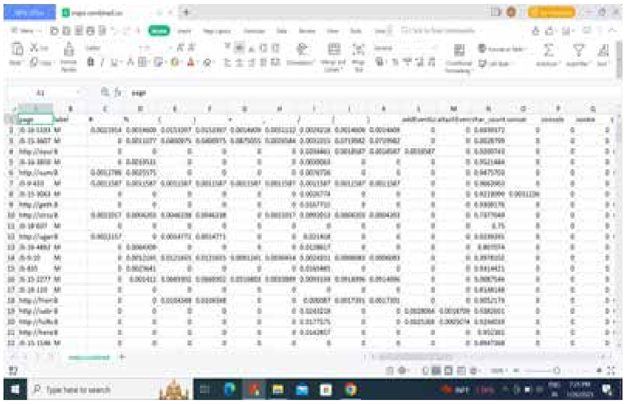
Here we will learn about the classifiers that machine learning makes use of to conceptualise phishing. In this section, we will outline the method we think is best for identifying phishing websites. One part of this is for classifiers and the other is to explain our proposed system.

Viewing the phishing website using machine learning classifiers There is a deep and energetic problem with distinguishing and recognising phishing websites. In fact, AI has been extensively applied in many places to provide automated outcomes. There are several forms that phishing attacks may take, such as send-off, website, virus, and voice. In this research, we employ the Hybrid Algorithm Technique to detect phishing attempts on websites (LINK). The system's accuracy and pricing estimates are enhanced by a combination of many classifiers. We may use any classification technique, depending on the application and the kind of the dataset. We can't tell which formulae transcend or not since there are so many uses.

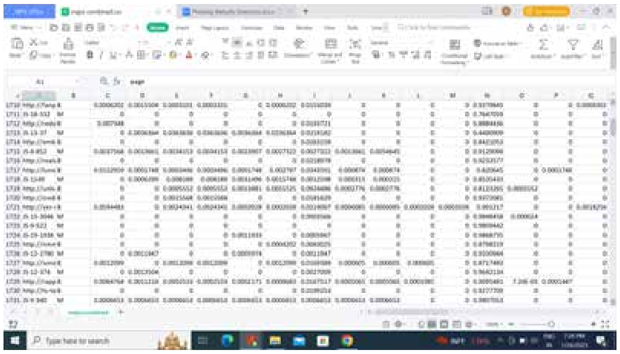
Vector Support System (SVM): This is also one of the easy-to-use classification methods that is supervised. Classification and regression are two possible uses for it, although the former is where it really shines. In contrast to previous classification techniques, support vector machines (SVMs) use the distance between neighbouring data points of all courses to determine the decision boundary. As a result of using SVMs to draw their decision boundaries, we get the best margin classifier, also known as the maximum margin active plane. Different aeroplanes' information set factors—differences in the courses—form the basis of the category.

Information Collection

Still today, phishing is one of the best ways for hackers to trick us into giving up our personal and financial information. Fraudsters now have the ideal environment to execute targeted phishing attacks, thanks to our increasing reliance on the internet to carry out much of our everyday business. The sophistication and invisibility of modern phishing assaults is growing. The majority of security professionals (97%) are unable to distinguish between legitimate and phishing emails, according to study by Intel.



There are 11,430 URLs and 87 extracted functions in the provided dataset. The dataset is specifically created to be used as a criterion for phishing detection systems that rely on machine learning. There are three sources for the attributes: 56 culled from the phrases and structure of Links, 24 from the web content of their reporting pages, and 7 from other services that were queried. There are exactly half legitimate URLs and half phishing ones in the collection.



IMPLEMENTATION

2. This section will serve as a review of the steps taken to carry out the experiment. The method for analysing the data and making the phishing prediction will be described in detail. We have made use of jumbled data that is just url-based. The number of urls retrieved from the internet is 11064. The majority of the links obtained are phishing, while it does include legitimate URLs as well.

1. The first piece of information is the Phish storage tank website's unstructured data, which consists of URLs.
2. Eight functions are generated from disorganised data during feature creation in pre-processing. This includes characteristics such as link size, http tokens, suspicious characters, prefix/suffix, slash variation, phishing phrase, subdomain length, and IP address.
3. Then, a structured dataset is created and distributed to all the classifiers; this dataset has binary values (0,1) for each attribute.
4. The fourth step is to train the two different classifiers, SVM and light gbm, and then compare and contrast how well they perform in terms of accuracy.
5. A classifier uses the training data to determine if the provided URL is legitimate or phishing. If it is legitimate, the classifier accesses the page in the browser; otherwise, it displays an error.
6. After comparing several classifiers' accuracy, we found that light gbm provides the greatest precision.
7. You can see screenshots of the implementation process down below.

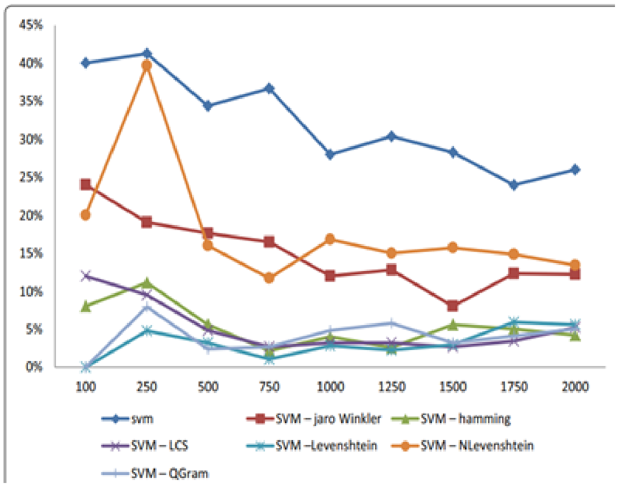


Fig.2. Graphical representation

Our experiments validate the hypothesis that using the similarity distance might improve the detection of phishing websites. The inclusion of the range of similarity has significantly improved our discovery system's identification rate in three out of four tests conducted on 4. Also in this regard, the test using Probabilistic Semantic network that records the worst detection rate of all our tests is the only one where the similarity did not have a positive effect on the phishing website identification rate. Since our system's recognition rate was enhanced by 21.8% when the Hamming range was used as an input feature, this impact is most apparent in tests conducted using the SVM approach.

CONCLUSION

Using state-of-the-art maker learning technology, this article seeks to enhance detection methods for phishing sites. Using a random forest approach with the lowest possible false positive rate, we achieved a detection accuracy of 97.14 percent. Furthermore, the results show that classifiers perform much better when we utilise additional data for training. To improve the accuracy of phishing site detection in the future, a hybrid approach combining the random forest algorithm of machine learning with a blacklist technique will be used.

Analysis of Attributes

Without collecting customer privacy-related data, including network traffic, the characteristics of the domain used here can only be retrieved by utilising known strings of domain. Attributes of the persons used in the domain and characteristics of information on the domain are the two primary types of domain features that may be classified according to the acquisition approach. While the matching website or other search engines may be utilised to get the domain's data, a local feature-extraction algorithm can be used to obtain the domain's personalities' traits even if the user doesn't visit the website.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "PHISHING URL DETECTION A REAL-CASE SCENARIO THROUGH LOGIN URLS", which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. Ms. Sophiya Shikalgar, Mrs. Swati Narwane (2019), Detecting of URL based Phishing Attack using Machine Learning. (vol. 8 Issue 11, November – 2019)
2. Rashmi Karnik, Dr. Gayathri M Bhandari, Support Vector Machine Based Malware and Phishing Website Detection.
3. Arun Kulkarni, Leonard L. Brown, III2 , Phishing Websites Detection using Machine Learning (vol. 10, No. 7,2019)
4. R. Kiruthiga, D. Akila, Phishing Websites Detection using Machine Learning.
5. Ademola Philip Abidoye, Boniface Kabaso, Hybrid Machine Learning: A Tool to detect Phishing Attacks in Communication Networks. (vol. 11 No. 6,2020)
6. Andrei Butnaru, Alexios Mylonas and Nikolaos Pitropakis, Article Towards Lightweight URL-Based Phishing Detection.13 June 2021
7. Ashit Kumar Dutta (2021), Detecting phishing websites using machine learning technique. Oct 11 2021
8. Nguyet Quang Do, Ali Selamat, Ondrej Krejcar, Takeru Yokoi and Hamido Fujita (2021) Phishing Webpage Classification via Deep Learning-Based Algorithms: An Empirical study.

9. Ammara Zamir, Hikmat Ullah Khan and Tassawar Iqbal, Phishing website detection using diverse machine learning algorithms.

10. Vahid Shahrivari, Mohammad Mahdi Darabi and Mohammad Izadi (2020), Phishing Detection Using Machine Learning Techniques.

11. A. A. Orunsolu, A. S. Sodiya and A.T. Akinwale (2019), A predictive model for phishing detection.

12. Wong, R. K. K. (2019). An Empirical Study on Performance Server Analysis and URL Phishing Prevention to Improve System Management Through Machine Learning. In Economics of Grids, Clouds, Systems, and Services: 15th International Conference, GECON 2018, Pisa, Italy, September 18-20, 2018, Proceedings (Vol. 11113, p. 199). Springer.

13. Desai, A., Jatakia, J., Naik, R., & Raul, N. (2017, May). Malicious web content detection using machine leaning. In 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 1432-1436). IEEE.

Block Hunter Federated Learning for Cyber Threat Hunting in Blockchain-Based IIOT Networks

Akash, Manoj, Haseeb
B.Tech Student
Department of Information Technology
CMR Technical Campus, Medchal
Hyderabad, Telangana

Y. Satyam
Assistant Professor
Department of Information Technology
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ satyam.it@cmrtc.ac.in

ABSTRACT

In order to improve data safety and security, several sectors are now developing blockchain-based contemporary solutions. A blockchain-based network is among the most notable uses of blockchain technology within the framework of the IIoT. Industrial Internet of Things (IIoT) devices are becoming more common in our digital environment, particularly for the purpose of building smart factories. Despite its usefulness, blockchain technology is vulnerable to cyber attacks. In order to protect networks and systems from unforeseen attacks, it is necessary to detect abnormalities in smart manufacturing facilities' blockchain-based IIoT networks. In this article, we build a threat hunting framework named Block Hunter using Federated Understanding (FL) to instantly search for attacks on IIoT networks that are built on blockchain. In a federated setting, Block Hunter employs a cluster-based approach for anomaly identification that incorporates many machine learning versions. Our research indicates that Block Seeker is the first federated risk hunting version for IIoT networks to identify suspicious patterns while protecting user privacy. Our results show that the Block Seeker is effective in detecting suspicious activity with a high degree of precision and a low amount of transmission capacity required.

KEYWORDS: *FL, Block hunter, IIOT, High security data, IoT.*

INTRODUCTION

Blockchain technology is becoming a useful tool in many fields, including healthcare, the military, finance, and networking, thanks to its immutable and tamper-proof data security features. Factories, in particular, are becoming more intelligent and efficient as a result of technological advancements, and this trend is driven by the ever-increasing use of Industrial Internet of Things (IIOT) solutions. [1] One subset of the Internet of Things (IoT) is the Industrial Internet of Things (IIOT). Nevertheless, when it comes to the requirements for safety and security, IIOT and IOT are diverse. While the IIOT improves the quality of life for consumers, its primary goal is to strengthen production security, efficiency, and safety. When it comes to business-to-business (B2B) environments, IIOT tools are more often employed, but IOT devices are more commonly considered in business-to-customer (B2C) settings. Because of this, the risk profile for IIOT networks would be different from that of their IOT equivalents, where device-to-device transactions are very valuable.

IIoT networks allow us to meet the demands of our clients and support a wide range of applications, especially in industrial settings like smart factories. [1] Smart factories, smart homes/buildings, smart farms, smart cities, connected drones, and medical care systems are just a few examples of the IIOT-based networks that have embraced blockchain technology due to its many benefits [1, 2]. This research primarily focuses on smart manufacturing facility block chain-based IIoT network security [3, 4], however the suggested architecture might be applied to other IIoT environments as well.

Modern smart factories use Internet-enabled lighting, temperature monitoring systems, IP electronic cameras, and IP phones to power a plethora of operations that rely on these technologies. These devices are storing confidential information and could provide answers that are vital to public safety. in [3], (1) The primary issue will undoubtedly be the secure storage, accumulation, and exchange of data as the number of IIOT devices in smart factories increases. Consequently, in this kind of situation, industrial, critical, and personal data are all at risk. With blockchain technology, data integrity, strong authentication, and a reliable timetable for communication foundations can be guaranteed both inside and outside of smart manufacturing plants. However, there are still significant obstacles to privacy and security in IIOT [3, 4]. An key challenge with blockchain-based networks is the potential of misleading tasks occurring in them [2, 4]. Blockchain technology is a powerful tool, but it is not immune to cyberattacks. Case in point: Ethereum Standard was hit by a 51% cyber attack [2] and three consecutive strikes in August 2020 [5], leading to the loss of more than $5 million worth of cryptocurrency. These incidents have shown the vulnerabilities of this blockchain network.

During transmission, utilisation, and storage, smart factories must protect the privacy of consumers' information. [4] Scammers may access, alter, or utilise the stored data for malicious purposes, making it susceptible to interference. When looking at the data, these assaults stand out as unusual occurrences that don't follow the norm. [2, 6] For threat hunting programmes and for safeguarding systems against unauthorised access, the ability to detect and filter out-of-the-ordinary activities is essential. the references [6], [7]

The primary objective of this article is to identify dubious clients and transactions inside an IIOT network that is built on blockchain technology, with a focus on smart manufacturing facilities. In this case, out-of-character actions stand in for dubious routines. [4] Machine learning (ML) techniques may be used to detect strikes and abnormalities on the blockchain by finding trends and outliers. Deep neural networks are a promising alternative for anomaly detection since they autonomously learn representations from training data. [4, 7] However, problems might arise with anomaly finding systems that rely on machine learning or deep learning. Concerns about privacy and a lack of training data are addressed by these methods. [7]

It is difficult to detect anomalies on the blockchain. [8] Not only does sending each block to a main server increase training time, but the version also needs fresh block data during testing [8]. Furthermore, malevolent adversaries might use causal/data poisoning attacks to intentionally damage the ML architecture when ML models are routinely updated to respond to new dangers and identify anomalies. To avoid detection of anomalies, attackers may deliberately send out designed payloads.

Using Federated Discovering (FL) architectures to detect anomalies while safeguarding personal information and monitoring data quality is a novel and practical technique. references [7], [9] With FL, edge devices may work together during training while keeping all data locally. Instead of transmitting the data to another place, we may train the model locally on the device, and then communicate just the most recent modifications with the rest of the network.

One recent trend in machine learning is FL, which allows for smart edge devices to make mutual predictions with one another [7], [10]. Also, FL handles crucial issues with data sharing, information security, and digital civil liberties management, and it ensures that several stars build long-lasting machine finding designs without exchanging data. This research adopts an anomaly-detection structure called Block Seeker that is based on FL and can identify attack hauls in IIOT networks that are built on the blockchain, according to these features.

This study primarily contributes to the following areas:

First, create an anomaly detection problem for smart factories that use blockchain technology by using a cluster-based architecture. When it comes to reducing transmission capacity and increasing throughput in IIoT networks, the cluster-based approach improves hunting effectiveness.

2) Identify suspicious activity in IIoT devices linked to smart factories that use blockchain technology by implementing a federated design version. In a federated setting, this provides a privacy-preserving function for machine learning versions.

3.) Applying several methods for finding abnormalities, such as clustering-based, analytical, subspace-based, classifier-based, and tree-based, to efficiently identify abnormalities in smart factories.

4) The Block Hunter structure is examined in relation to block production, block size, and miners. True Positive Rate (TPR) anomaly detection, F1-score, Precision, and Recall are some of the performance metrics that are examined.

SURVEY OF RESEARCH

Digital bitcoin transactions may be analysed with the use of an algorithm that was suggested by Sayadi et al. [5]. In order to classify outliers that were similar in kind and statistical importance, they looked at the K-means and One-Class Support Vector Machines (OCSVM) algorithms. After reviewing their work via the creation of discovery results, they discovered that we can get excellent results in terms of accuracy.

Anomaly semantics in blockchain-based IoT networks was the basis for the authors' proposed solution in [6]. An approach was already in place to detect suspicious activity in blockchains by gathering metadata in forks and using it to determine common informational identification of unusual tasks. They developed a device that improves the security of blockchains and interconnected devices. Similarly, in order to discover blockchain security, has really presented encoder-decoder deep learning regression in [7]. An anomaly finding framework based on collected data from bitcoin blockchain monitoring was constructed in this study. Their testing has shown that their system can detect publicly known attacks by mining the Ethereum network's previous records.

The authors Chai et al. [2] suggested using FL and a hierarchical blockchain structure to discover and exchange environmental data. For large-scale vehicle networks, this design is practical and dependable. The distributed pattern and personal privacy demands of the Net of Autos are met by FL-based discovery. Knowledge sharing is encouraged via the use of a model that mimics a multi-leader, multi-player trading market mechanism. The results of the substitution show that an ordered structure-based approach may improve the sharing, discovery, and handling of specific harmful attacks. Furthermore, the authors of [3] provide an exhaustive analysis of how FL might offer enhanced cybersecurity and evade many attackers simultaneously. This study identifies key challenges and opportunities for further research on FL's implementation in real-world settings.

PROPOSED SYSTEM

Automatically protecting a system from unforeseen attacks relies heavily on detecting suspicious behaviours. In order to detect blockchain anomalies, every time a block is upgraded, the data from that block must be sent to a central server. In addition to being ineffective, this raises issues of individual privacy. When it comes to addressing this issue, FL options appear promising. To discover anomalies, we utilise FL to get an international version of the model and to update it periodically. Once we have gathered information on each smart factory's data, devices, and business, we will send the version's specs to the parameter server so that we can aggregate them and improve our core design. With collection-based architecture, the blockchain can function in any smart factory with much more dependable usage sources and throughput. Clustering simplifies the hierarchical hidden network construction process in terms of computational complexity.

WORKING METHODOLOGY

Fig.1. Home page

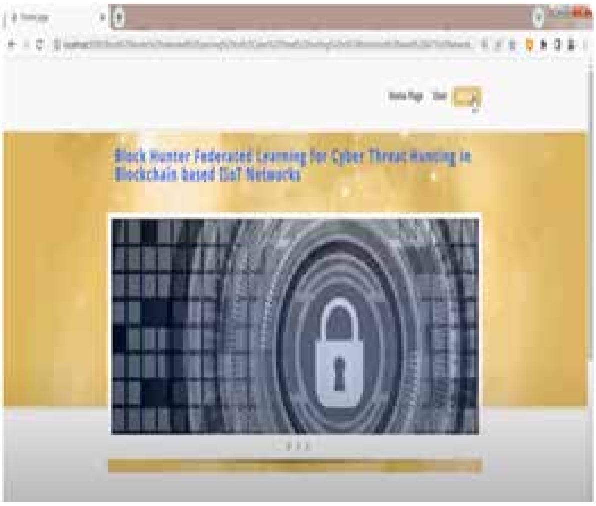

Home Page
User
Block Hunter Federated Learning for Cyber Threat Hunting in Blockchain based IIoT Networks

Fig.2. Home page

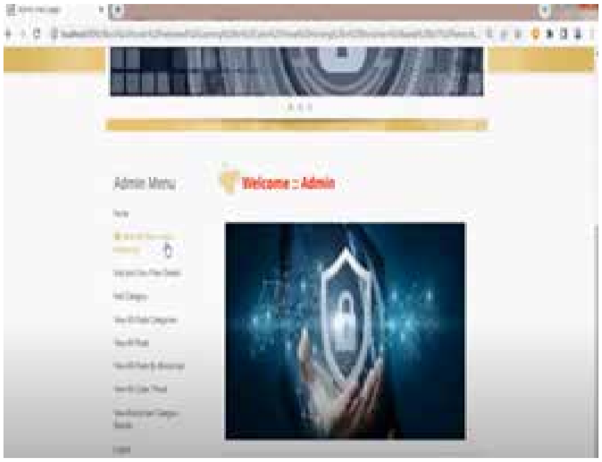


Fig.3. Admin login page

Fig.4. All details of users

Fig.5. Cyber thefts details

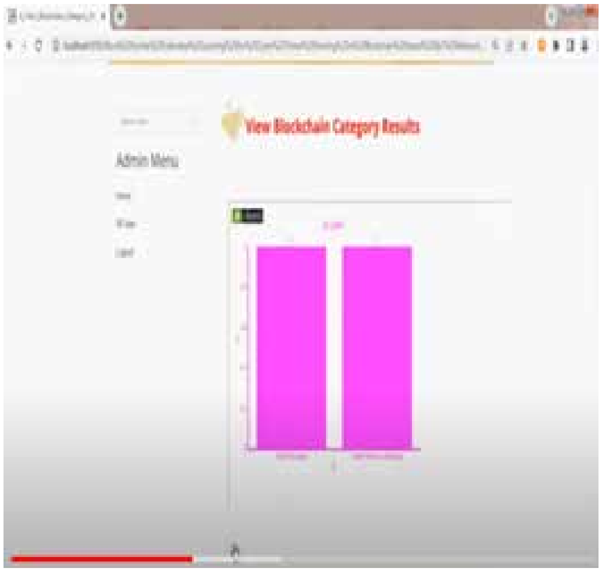


Fig.6. Output results

In the intelligent sectors, IIoT tools are often used. Although blockchain is secure, it may be hacked. In order to protect themselves against assaults, smart manufacturing facilities that use blockchain-based IIoT networks need to detect issues. Block seeker is a threat-hunting framework that hunts for dangers in blockchain-based IIoT networks automatically. It is constructed using federated learning. When looking for anomalies, block hunter employs federated maker learning designs with a cluster-based architecture. A privacy-preserving, federated danger-hunting method called "block hunter" for IIoT networks. When using the FedAvg approach with a discovery rate of 95%, the block hunter is able to accurately identify suspicious behaviours even when bandwidth is limited.

CONCLUSION

Using a federated understanding approach, we built the Block Hunter framework in this article to seek for anomalies in IIOT smart factories that are based on the blockchain. To improve the search performance of IIOT networks that utilise blockchain technology and reduce their associated sources, Block Hunter employs a cluster-based design. We used several AI techniques (NED, IF, CBLOF, K-means, PCA) to look for anomalies in the Block Seeker framework. The effects of block size, block creation interval, and the number of miners on Block Seeker performance were also investigated. An intriguing area for future research may be the use of generative adversarial networks (GAN) to build and run a framework similar to block hunters. It would also be worthwhile to investigate, down the road, the possibility of developing and implementing IIOT-related blockchain connections with other agreement formulae.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled “BLOCK HUNTER FEDERATED LEARNING FOR CYBER THREAT HUNTING IN BLOCKCHAIN-BASED IIOT NETWORKS”, which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. J. Wan, J. Li, M. Imran, D. Li, and F. e Amin, “A blockchain-based solution for enhancing security and privacy in smart factory,” IEEE Transactions on Industrial Informatics, vol. 15, no. 6, pp. 3652–3660, 2019.
2. F. Scicchitano, A. Liguori, M. Guarascio, E. Ritacco, and G. Manco, “Blockchain attack discovery via anomaly detection,” Consiglio Nazionale delle Ricerche, Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR), 2019, 2019.
3. Q. Xu, Z. He, Z. Li, M. Xiao, R. S. M. Goh, and Y. Li, “An effective blockchain-based, decentralized application for smart building system management,” in Real-Time Data Analytics for Large Scale Sensor Data. Elsevier, 2020, pp. 157–181.
4. B. Podgorelec, M. Turkanovi´c, and S. Karakati˘c, “A machine learningbased method for automated blockchain transaction signing including personalized anomaly detection,” Sensors, vol. 20, no. 1, p. 147, 2020.
5. A. Quintal, “Veriblock foundation discloses mess vulnerability in ethereum classic blockchain,” VeriBlock Foundation. [Online]. Available: https://www. prnewswire. com/news-releases/veriblock-foundation-discloses-mess-vulnern ability-in-ethereum-classic-blockchain-301327998.html
6. M. Saad, J. Spaulding, L. Njilla, C. Kamhoua, S. Shetty, D. Nyang, and D. Mohaisen, “Exploring the attack surface of blockchain: A comprehensive survey,” IEEE Communications Surveys & Tutorials, vol. 22, no. 3, pp. 1977–2008, 2020.
7. R. A. Sater and A. B. Hamza, “A federated learning approach to anomaly detection in smart buildings,” arXiv preprint arXiv:2010.10293, 2020.
8. O. Shafiq, “Anomaly detection in blockchain,” Master’s thesis, Tampere University, 2019.
9. A. Yazdinejadna, R. M. Parizi, A. Dehghantanha, and H. Karimipour, “Federated learning for drone authentication,” Ad Hoc Networks, p. 102574, 2021.
10. D. Preuveneers, V. Rimmer, I. Tsingenopoulos, J. Spooren, W. Joosen, and E. Ilie-Zudor, “Chained anomaly detection models for federated learning: An intrusion detection case study,” Applied Sciences, vol. 8, no. 12, p. 2663, 2018.
11. L. Tan, H. Xiao, K. Yu, M. Aloqaily, and Y. Jararweh, “A blockchainempowered crowdsourcing system for 5g-enabled smart cities,” Computer Standards & Interfaces, vol. 76, p. 103517, 2021.
12. L. Tseng, X. Yao, S. Otoum, M. Aloqaily, and Y. Jararweh, “Blockchainbased database in an iot environment: challenges, opportunities, and analysis,”

Cluster Computing, vol. 23, no. 3, pp. 2151–2165, 2020.

13. M. Signorini, M. Pontecorvi, W. Kanoun, and R. Di Pietro, "Bad: a blockchain anomaly detection solution," IEEE Access, vol. 8, pp. 173 481–173 490, 2020.

[14] S. Iyer, S. Thakur, M. Dixit, R. Katkam, A. Agrawal, and F. Kazi, "Blockchain and anomaly detection based monitoring system for enforcing wastewater reuse," in 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2019, pp. 1–7.

[15] S. Sayadi, S. B. Rejeb, and Z. Choukair, "Anomaly detection model over blockchain electronic transactions," in 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC). IEEE, 2019, pp. 895–900.

Waternet a Network for Monitoring and Assessing Water Quality for Drinking and Irrigation Purposes

D Samuel, P Bhanuteja, B Vinay
B.Tech Students
Department of Information Technology
CMR Technical Campus, Medchal
Hyderabad, Telangana

Ch. Ramesh
Faculty
Department of Information Technology
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ ramesh.it@cmrtc.ac.in

ABSTRACT

Building Long-Term, Resilient Smart As cities expand at an unprecedented rate, water supply systems are encountering serious problems on a global scale. All municipal monitoring is focused on water quality, which affects our lives everywhere. Testing for physical, chemical, and organic indicators of water quality on a regular basis was the main focus of traditional methods for controlling urban water quality. Still, many major cities have seen an increase in health risks due to accidents like massive infections brought on by the inevitability of waiting for biological signs. We begin by analysing the issue, the technological challenges, and the research study questions in this article. Then, we provide a framework for evaluating risks to the metropolitan water system, which might be a service in and of itself. In order to see changes to water top quality and further for danger identification, it uses indicator data that we gathered from industrial activities. We provide a Flexible Regularity Evaluation (AdpFA) method to resolve the data employing signs' regularity domain name information for their private predictions and internal partnerships in order to provide findings that can be explained. We also investigate the method's scalability features from indication, location, and time domain names. To power the app, we mined four separate metropolitan water systems in Norway—Oslo, Bergen, Strommen, and Aalesund—for high-quality industrial data. In order to compare the suggested technique to conventional Artificial Semantic network and Random Forest methods, we analyse the spectrogram, prediction precision, and time consumption. The results demonstrate that our approach performs much better across the board. Commercial water systems may continue to provide high-quality water with early indications of potential problems and greater choice support.

KEYWORDS: *ANN, FA, Risk, Quality.*

INTRODUCTION

Building Long-Term, Resilient Smart As cities expand at an unprecedented rate, water supply systems are encountering serious problems on a global scale. All municipal monitoring is focused on water quality, which affects our lives everywhere. Testing for physical, chemical, and organic indicators of water quality on a regular basis was the main focus of traditional methods for controlling urban water quality. Still, many major cities have seen an increase in health risks due to accidents like massive infections brought on by the inevitability of waiting for biological signs. We begin by analyzing the issue, the technological challenges, and the research study questions in this article. Then, we provide a framework for evaluating risks to the metropolitan water system, which might be a service in and of itself. In order to see changes to water top quality and further for danger identification, it uses indicator data that we gathered from industrial activities. We provide a Flexible Regularity Evaluation (AdpFA) method to resolve the data employing signs' regularity domain name information for their private predictions and internal partnerships in order to provide findings that can be explained. We also investigate the method's scalability features from indication, location, and time domain names. To power the app, we mined four separate metropolitan water systems in Norway—Oslo, Bergen, Strommen, and Aalesund—for high-quality industrial data. In order to compare the suggested technique to conventional Artificial Semantic network and Random Forest methods, we analyse the spectrogram, prediction precision, and time consumption. The results demonstrate that our approach performs much better across the board. Commercial water systems may continue to provide high-quality water with early indications of potential problems and greater choice support.

LITERATURE SURVEY

In contrast to the suggested flexible regularity analysis method, existing ANN and arbitrary forests will almost likely lack control over dataset processing operations, resulting in a high error rate. - A Although the author of the suggested work utilised a dataset from the Norwegian national water supply, he did not make it publicly available, so we do not have access to it. However, we were able to locate a high-quality dataset from the Indian state water supply.

DAMAGE FEATURES

Several obstacles stand in the way of our ability to assess the dangers of water top quality modification and investigate the mechanism behind the data sources: 1. Information Sparsity: Usually, there is a big pool of accessible information. Examples of water high quality indicators often show very little or nonexistent overlaps between two situations (such as the exact same time and location). Two main considerations have led to this conclusion. The first issue is that the drivers who collect the samples don't adhere to protocol, which leads to data loss and inadequate indication collecting. Second, there has been a shift in the information standard over the last several years, with certain signs being added and others removed. The data becomes quite thin as a result of these.

2. Data Synchronisation: New, cutting-edge technology can support the continuous gathering of data on various physical and chemical indicators of water quality in real-time. In contrast, tests for organic indicators—the most important aspects of health—tend to take far longer, anything from several hours to days. This complicates the process of integrating the collected data.

The third and last goal of controlling water quality for drinking is to improve health and wellbeing; this goal is known as risk modelling. The presence of some organic markers, such as microbes like E. coli, may lead to severe outbreaks of sickness. Immediate and catastrophic damage may result from their transmission via the water distribution system for alcohol use. A revised version of the cooperation between these biological markers and the water danger associated with alcohol intake is required.

METHODOLOGY

In contrast to the suggested flexible regularity analysis method, existing ANN and arbitrary forests will almost likely lack control over dataset processing operations, resulting in a high error rate. - A Although the author of the suggested work utilized a dataset from the Norwegian national water supply, he did not make it publicly available, so we do not have access to it. However, we were able to locate a high-quality dataset from the Indian state water supply.

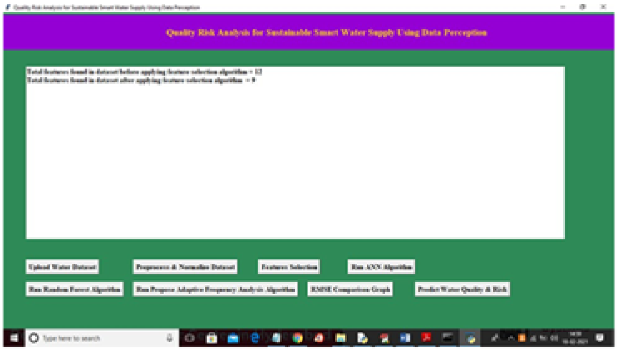
DAMAGE FEATURES

Several obstacles stand in the way of our ability to assess the dangers of water top quality modification and investigate the mechanism behind the data sources: 1. Information Sparsity: Usually, there is a big pool of accessible information. Examples of water high quality indicators often show very little or nonexistent overlaps between two situations (such as the exact same time and location). Two main considerations have led to this conclusion. The first issue is that the drivers who collect the samples don't adhere to protocol, which leads to data loss and inadequate indication collecting. Second, there has been a shift in the information standards over the last several years, with certain signs being added and others removed. The data becomes quite thin as a result of this.

2. Data Synchronisation: New, cutting-edge technology can support the continuous gathering of data on various physical and chemical indicators of water quality in real-time. In contrast, tests for organic indicators—the most important aspects of health—tend to take far longer,

anything from several hours to days. This complicates the process of integrating the collected data.

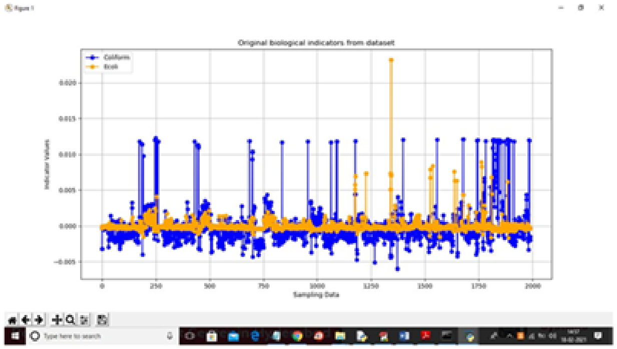
The third and last goal of controlling water quality for drinking is to improve health and well being; this goal is known as risk modelling. The presence of some organic markers, such as microbes like E. coli, may lead to severe outbreaks of sickness. Immediate and catastrophic damage may result from their transmission via the water distribution system for alcohol use. A revised version of the cooperation between these biological markers and the water danger associated with alcohol intake is required.


Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception
Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception
Upload Water Dataset
Preprocess & Normalize Dataset
Features Selection
Run ANN Algorithm
Run Random Forest Algorithm
Run Propose Adaptive Frequency Analysis Algorithm
RMSE Comparison Graph
Predict Water Quality & Risk
Type here to search

```
Total features found in dataset before applying feature selection algorithm : 12
Total features found in dataset after applying feature selection algorithm : 9
```

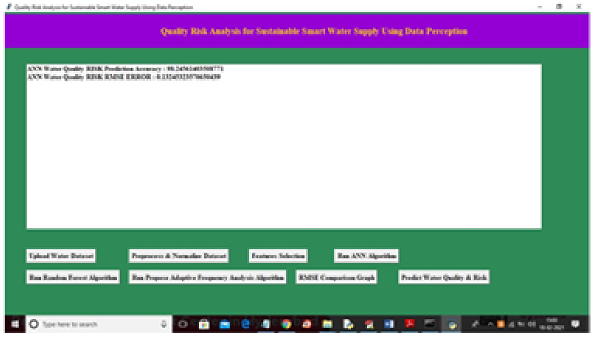
Hospital Database

The characteristic selection procedure reduces the dataset's characteristics from 12 within the preceding screen to 9 within the one under, and the resulting graph is proven beneath.



AI Search String

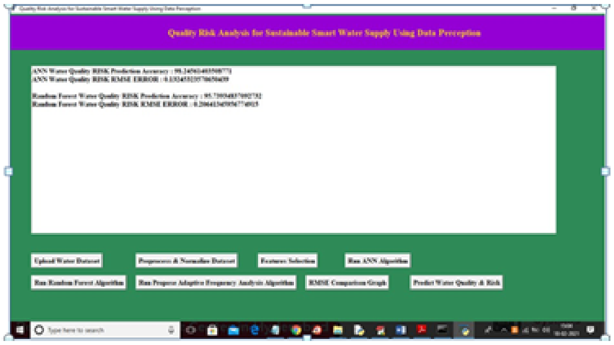
Blue shows the presence of COLI bacteria inside the dataset, even as orange indicates the presence of ECOLI micro organism. Both styles of micro organism are visible within the water dataset proven in the above graph. Once the dataset is ready, you may teach an ANN on it after which determine its root-imply-squared errors (RMSE) by using clicking the "Run ANN Algorithm" button.


Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception
Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception
Upload Water Dataset
Preprocess & Normalize Dataset
Features Selection
Run ANN Algorithm
Run Random Forest Algorithm
Run Propose Adaptive Frequency Analysis Algorithm
RMSE Comparison Graph
Predict Water Quality & Risk
Type here to search

```
ANN Water Quality RISK Prediction Accuracy : 98.24561403508771
ANN Water Quality RISK RMSE ERROR : 0.13245323570650439
```

AI Search Result

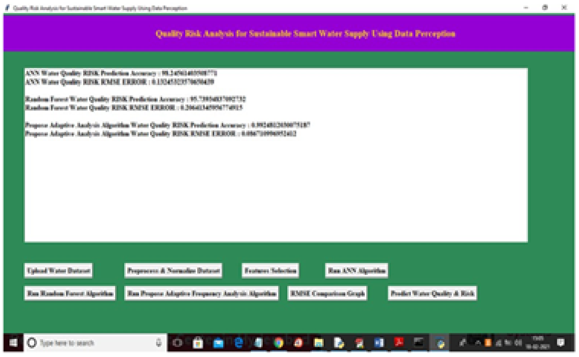
To train the dataset using random wooded area, click the "Run Random Forest Algorithm" button. The results could be proven underneath. The ANN accuracy is 98% and its RMSE is 0.Thirteen%.


Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception
Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception
Upload Water Dataset
Preprocess & Normalize Dataset
Features Selection
Run ANN Algorithm
Run Random Forest Algorithm
Run Propose Adaptive Frequency Analysis Algorithm
RMSE Comparison Graph
Predict Water Quality & Risk
Type here to search

```
ANN Water Quality RISK Prediction Accuracy : 98.24561403508771
ANN Water Quality RISK RMSE ERROR : 0.13245323570650439

Random Forest Water Quality RISK Prediction Accuracy : 95.73934837092732
Random Forest Water Quality RISK RMSE ERROR : 0.20641545956774915
```

Patient Login

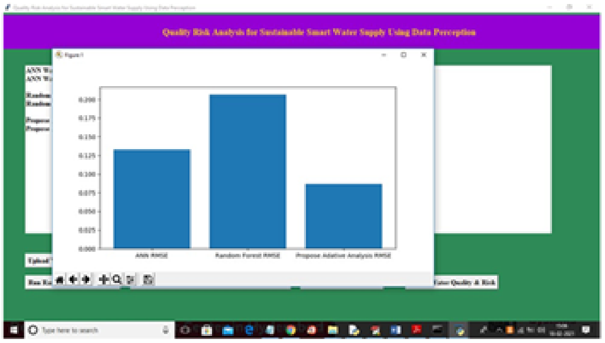
To train the algorithm using the dataset and get the results below, click on the "Run Propose Adaptive Frequency Analysis Algorithm" button. The accuracy of Random Forest is 95.73% and its error rate is 0.20%.


Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception
Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception
Upload Water Dataset
Preprocess & Normalize Dataset
Features Selection
Run ANN Algorithm
Run Random Forest Algorithm
Run Propose Adaptive Frequency Analysis Algorithm
RMSE Comparison Graph
Predict Water Quality & Risk
Type here to search

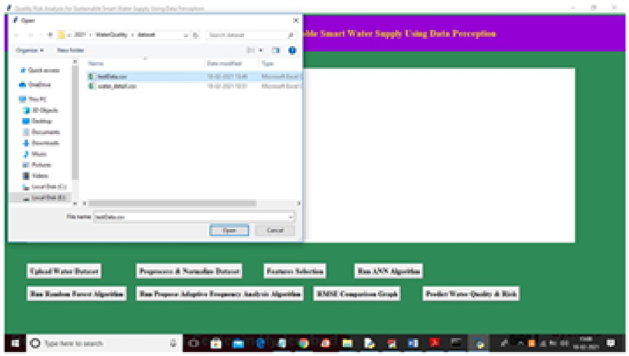
```
ANN Water Quality RISK Prediction Accuracy : 98.24561403508771
ANN Water Quality RISK RMSE ERROR : 0.13245323570650439

Random Forest Water Quality RISK Prediction Accuracy : 95.73934837092732
Random Forest Water Quality RISK RMSE ERROR : 0.20641545956774915

Propose Adaptive Analysis Algorithm Water Quality RISK Prediction Accuracy : 0.9924812030075187
Propose Adaptive Analysis Algorithm Water Quality RISK RMSE ERROR : 0.0867109969052412
```

Patient Details

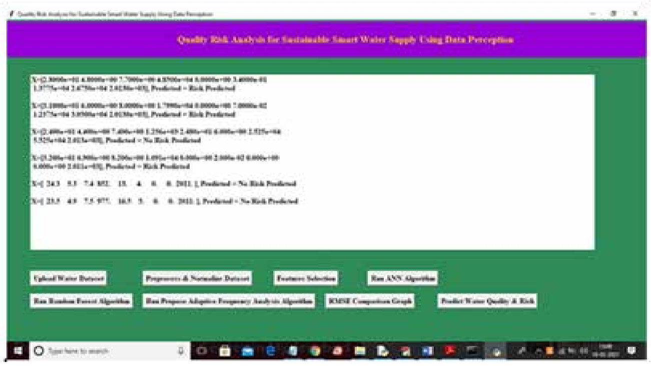
You can see that the proposed method outperforms the other two by clicking the "RMSE Comparison Graph" button; its accuracy is 99% (0.99 * 100= 99%) and its error rate is 0.086%.


Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception
Figure 1
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
ANN RMSE
Random Forest RMSE
Propose Adative Analysis RMSE
Type here to search

From the three techniques shown, the adaptive approach had the best overall performance, with the lowest root-mean-square error rate (RMSE) shown on the y-axis. To submit test water data, click the "Predict Water Quality & Risk" option. The programme will then determine whether the data is RISKY.


Open
Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception
testData.csv
water_data.csv
File name: testData.csv
Open
Cancel
Upload Water Dataset
Preprocess & Normalize Dataset
Features Selection
Run ANN Algorithm
Run Random Forest Algorithm
Run Propose Adaptive Frequency Analysis Algorithm
RMSE Comparison Graph
Predict Water Quality & Risk
Type here to search

Click the "Open" button on the previous page after choosing and uploading the "testData.csv" file to get the following prediction result.


Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception

```
X=[2.3000e+01 4.8000e+00 7.7000e+00 4.8700e+04 5.0000e+00 3.4000e+01
 1.3775e+04 2.6750e+04 2.0150e+03], Predicted = Risk Predicted

X=[3.1000e+01 4.0000e+00 5.0000e+00 1.7990e+04 8.0000e+00 7.0000e+02
 1.2175e+04 3.0500e+04 2.0150e+03], Predicted = Risk Predicted

X=[2.400e+01 4.400e+00 7.400e+00 1.256e+03 2.400e+01 6.000e+00 2.525e+04
 5.525e+04 2.015e+03], Predicted = No Risk Predicted

X=[3.200e+01 4.900e+00 5.200e+00 1.695e+04 5.000e+00 2.000e+02 8.000e+00
 4.000e+00 2.015e+03], Predicted = Risk Predicted

X=[ 24.3   5.3   7.4 852.   15.    4.    6.    0. 2011. ], Predicted = No Risk Predicted

X=[ 23.5   4.9   7.5 977.   16.5   5.    6.    0. 2011. ], Predicted = No Risk Predicted
```

Upload Water Dataset
Preprocess & Normalize Dataset
Features Selection
Run ANN Algorithm
Run Random Forest Algorithm
Run Propose Adaptive Frequency Analysis Algorithm
RMSE Comparison Graph
Predict Water Quality & Risk
Type here to search

CONCLUSION

In today's modern urban living, water quality is of the utmost importance, particularly for the expansion of Smart Water Supply systems. It is difficult to identify transmitted germs on time and provide trustworthy decision support using conventional monitoring and risk regulation methods. We provide a data-perception-based method for early warning of water top quality dangers in this article. We have validated the practicality, precision, and efficacy of our method via its implementation in four different Norwegian cities. Domain specialists have analysed the early findings, and they are quite promising. The three main areas in which this work excels are:

In the water resource regions, it uses cost-free data analysis methodologies to give an early warning system. This keeps more options open for the latter steps of water delivery and lengthens the response time of preventative measures.

Sign, location, and time are all brought together in this method. This provides a fresh perspective on regularity domain name analysis that may be used to find the link between various indications and their predictions. Concurrently, it is open to the idea of scalability for those three domain names.

This project is based on actual industrial water supply systems in four different Norwegian towns.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "WATERNET A NETWORK FOR MONITORING AND ASSESSING WATER QUALITY FOR DRINKING AND IRRIGATION PURPOSES", which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department Of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. S. Franco, V. Gaetano, and T. Gianni, "Urbanization and climate change impacts on surface water quality: Enhancing the resilience by reducing impervious

surfaces," Water Research, vol. 144, pp. 491–502, 2018.

2. T. Hak, S. Janou ´ skov ˇ a, and B. Moldan, "Sustainable development ´ goals: A need for relevant indicators," Ecological Indicators, vol. 60, pp. 565–573, 2016.

3. World Health Organization (WHO), Guidelines for drinking-water quality: recommendations. World Health Organization, 2004.

4. E. Weinthal, Y. Parag, A. Vengosh, A. Muti, and W. Kloppmann, "The eu drinking water directive: the boron standard and scientific uncertainty," European Environment, vol. 15, no. 1, pp. 1–12, 2005.

5. R. W. Adler, J. C. Landman, and D. M. Cameron, The clean water act 20 years later. Island Press, 1993. [6] D. Berge, "Overvakingavfarrisvannet med tilløpfra 1958-2010," ˚ 2011.

7. I. W. Andersen, "EUs rammedirektiv for vann–miljøkvalitetsnormer for vannmiljøetimøte med norskrett," Kart og Plan, vol. 73, no. 5, pp. 355–366, 2013.

8. V. Novotny, Water quality: prevention, identification and management of diffuse pollution. Van Nostrand-Reinhold Publishers, 1994.

9. A. Hounslow, Water quality data: analysis and interpretation. CRC press, 2018.

10. S. Yagur-Kroll, E. Schreuder, C. J. Ingham, R. Heideman, R. Rosen, and S. Belkin, "A miniature porous aluminum oxide-based flowcell for online water quality monitoring using bacterial sensor cells," Biosensors and Bioelectronics, vol. 64, pp. 625–632, 2015.

11, H. R. Maier and G. C. Dandy, "The use of artificial neural networks for the prediction of water quality parameters," Water Resources Research, vol. 32, no. 4, pp. 1013–1022, 1996.

12. H. Orouji, O. Bozorg Haddad, E. Fallah-Mehdipour, and M. Marino, "Modeling of water quality parameters using data- ˜ driven models," Journal of Environmental Engineering, vol. 139, no. 7, pp. 947–957, 2013.

Crop Recommender System using Machine Learning Approach

Namratha, Shareef, Rakesh
B.Tech Students
Department of Information Technology
CMR Technical Campus, Medchal
Hyderabad, Telangana

M. Siva Jyothi
Faculty
Department of Information Technology
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ sivajyothi.it@cmrtc.ac.in

ABSTRACT

In India, the agricultural industry and its related industries unquestionably constitute the most important sources of income. Additionally, the agricultural sector contributes significantly to the GDP of the nation. The agricultural market is a boon to the nation due to its massive size. However, disappointingly, crop returns per hectare fall short of global standards. This is one of the possible explanations for the higher suicide rate among India's marginal farmers. A simple and reasonable method for farmers to anticipate their returns is suggested in this research. Using a smartphone app, the proposed solution connects farmers to the internet. Locating a certain person is much easier with the use of GPS. The location and soil type are provided by the person. Selecting the most productive plant checklist or predicting the harvest from a user-selected plant are both made possible by machine learning algorithms. Machine learning methods including Support Vector Machine (SVM), Artificial Neural Network (ANN), Random Woodland (RF), Multivariate Linear Regression (MLR), and K-Nearest Neighbour (KNN) are used to predict when the plants would return. With a 95% confidence level, the Random Woodland model outperformed the others. The algorithm also suggests when it's ideal to apply the plant foods in order to maximise production.

KEYWORDS: *Crop, Agriculture, Farmer, Wrong crop.*

INTRODUCTION

Over half of our countrymen rely on agriculture as a primary source of income. In 17 states, the average monthly income of a farmer is only Rs.1700/-, which is so low that it causes farmer suicides and the conversion of farmland to non-farm uses, according to the 2016–17 economic survey. Furthermore, almost half of farmers do not want their children or grandchildren to continue the family business but rather to live in urban regions. The main cause of this is because farmers often choose crops that won't yield much for their soil type or that aren't suitable for their growing season, among other common mistakes [9]. Due to the farmer's lack of expertise, the choice may have been made after acquiring the property from someone else. Picking the wrong plants will always result in lower yields. It becomes even more difficult to bear if the family's livelihood is totally reliant on this income. Prospective scientists are unable to assist with creating nation studies due to the lack of both accessibility and suitable, up-to-date information. A system has been proposed to address this issue using available resources. It provides predictions regarding plant sustainability and suggestions based on AI designs that take significant ecological and economical factors into account. When recommending a crop to an individual, the proposed system takes into account both the soil characteristics (such as type, pH value, and nutrient concentration) and the climatic factors (such as rainfall, temperature, and geographical position in terms of the state). In addition, the farmer will also get information on the expected harvest if they choose the right plant. The objective is to construct a long-term model that accurately predicts how long plants will survive in a given environment, taking into account the specific soil type and weather conditions. To avoid financial loss, recommend the best plants for the area so that the farmer may reap the benefits.3. Use data from the previous year to assess the profitability of various plants. One use of expert systems is artificial intelligence, which the proposed system makes use of. This technology allows systems to learn and progress instantaneously, even when not explicitly instructed to do so by a developer. When it is complete, the programme will be more accurate even when no human intervention is involved. In order to help farmers make the best decision possible when selecting a crop, researchers are looking at several aspects such as physical, environmental, and economic considerations. Prior to farming, the crops were graded using techniques such as Choice Tree Learning-ID3 (Iterative Dichotomiser 3) and K Nearest Neighbours Regression. A Synthetic Neural Network is then used to choose the plant with the highest return price [1]. [9] Using BigML and the random woodland method, we looked for plant characteristics. [10] To help plants cope with water stress, AI algorithms have developed decision-making principles that are included into plant health status forecasts. To forecast the price of crops, artificial intelligence approaches were used, and smart technologies were employed to provide ideas in real-time. As part of this assignment, we investigated several machine learning formula applications in agricultural production systems. Recommendations on plant management were supplied by other AI-enabled systems. Improved crop yields in plant cultivation are possible with the use of deep learning algorithms. An effective return projecting mechanism is designed in this study taking into consideration real-time monthly weather conditions. The aforementioned forecasting system was being implemented using a non-parametric statistical version in conjunction with non parametric regression methodologies.

LITERATURE SURVEY

Choice of Plants Method for optimising plant return rate using machine learning When it comes to agro-based economies and food security, farming preparation is king. One major consideration in agricultural planning is the selection of plants. A lot of factors, including production cost, market price, and government initiatives, play into determining this. Using data methods or machine learning techniques, several researchers have investigated the prediction of crop yield pricing, weather prediction, soil categorization, and plant category for agricultural preparation. Choosing a plant becomes more of a challenge when there is more than one way to cultivate it consistently while using limited land resources. This research proposed the Crop choice Technique (CSM) as a means to resolve the crop choice problem, maximise the plant's web return rate over time, and, finally, achieve the country's economic development objectives. Plant net yield price could go up if the proposed approach is implemented.

Mining Data for More Accurate Plant Yield and Pesticide Forecasts to Help the Agricultural Economy Risk is inherent in farming. Many factors, including weather, geography, biology, politics, and the economy, influence crop production. Some risks arise from these elements, and they may be assessed using suitable statistical or mathematical methods. Accurate information about the characteristics of plant returns in the past is crucial for modelling purposes; this information aids farmers and government agencies in making decisions and creating effective strategies. There is now access to vast amounts of data because to advancements in computing and data storage. Information mining is one innovative approach that has emerged as a result of the need to extract useful insights from large datasets in order to improve agricultural production evaluations. In order to determine whether meaningful links may be found, this research set out to evaluate these novel data mining techniques by applying them to the various database variables.

Thirdly, assessing soil data with the use of classification algorithms and the prediction of dirt attributes. Scientific research in agriculture has benefited from technological developments like data mining and automation. Data mining in agricultural datasets is an emerging area of research, despite data mining's widespread usage and the availability of both generic and domain-specific information mining tools. The massive volumes of data that are now often gathered alongside crops should be thoroughly investigated and used to their full potential. The goal of this study is to use data mining techniques to examine a soil dataset. Various easily accessible formulae are used to concentrate on certain types of dirt. The prediction of unknown characteristics via the use of regression analysis and the implementation of automated soil sample categorization is another critical aim.

Intelligent Farming Implementing Machine Learning In India's current economic climate, farming is vital. The agricultural sector in India is, nevertheless, now facing a precarious architectural adjustment. The only way out of this jam is to encourage farmers to keep up the plant manufacturing operations by making farming a lucrative enterprise. In an attempt to go in that direction, this term paper will discuss how farmers may use machine learning to make better agricultural choices. In this study, we utilise supervised equipment finding algorithms to predict the best plant to employ in a given situation by looking at past crop yields and potential weather conditions. Along with it, an online app has been created.

Existing System

About 58% of our countrymen rely on farming as one of their primary sources of income. A survey conducted in 2016–17 found that farmers in 17 states receive an average of Rs.1700/–per month. This low income causes many to give up farming and divert their land from agriculture to other uses. On top of that, over half of all farmers would prefer that their children and grandchildren stay in the city rather than continue the family business. This is due to the fact that farmers often make poor decisions when it comes to crop options, such as choosing a plant that won't do well in their soil, planting at the wrong season, etc. It is possible that the farmer acquired the property from other parties, allowing them to make the choice even without prior knowledge. Choosing the wrong plants will always result in lower yields. It becomes quite difficult to subsist if the family is totally reliant on these incomes. A random woodland formula was stored in the previous system. still haven't figured out which plant is best recommend.

PROPOSED SYSTEM

Here, we implement a system that uses machine learning to help farmers improve their return rate by recommending the right plants depending on factors like soil type, sowing season, weather, physical, environmental, and financial variables, and which plants are still in need. Systems may now learn and adapt on their own, thanks to today's technology, even when no explicit configuration from a designer is present. Eliminating human intervention will unquestionably improve the program's accuracy. In order to aid farmers in making informed crop selections that take into account many factors, including physical, ecological, and economic ones, a number of academics are delving into this topic. mechanism. In order to implement the aforementioned system of projections, a non-parametric analytical design and non-parametric regression techniques were used. This challenge involves feeding the algorithm through a variety of datasets obtained from Kaggle and the government website. Following the pre-processing step, several versions of the equipment are trained to achieve maximum accuracy using the stored dataset.

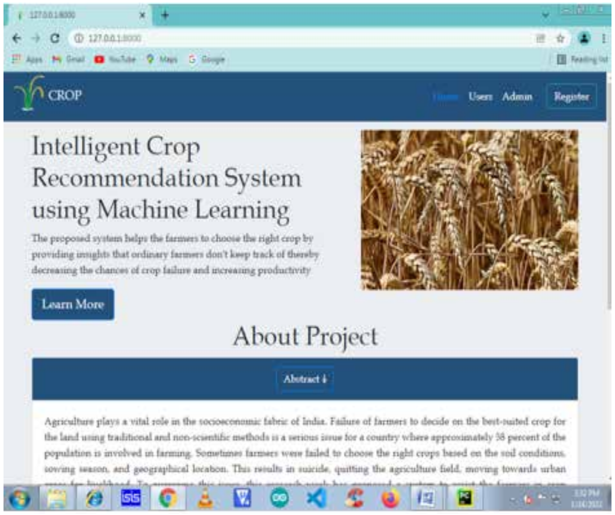
MODULES DESCRIPTION

Everyone has the ability to initiate the registration process. He had to provide a real email and phone number for the registration to go forward. The administrator may activate the user once they successfully register. We guarantee that our system will let the client log in whenever the management prompts them to do so. Clients may provide datasets that are a good fit for our dataset columns. The algorithm can only process data that is in float or integer format. Here, a ph was conducted. In addition, a dataset concerning climate change issues for evaluation purposes. Customers may also add new data to existing datasets using our Django application. Anyone may begin the data cleansing process by going to the website and selecting the "Information Preparations" tab. You may be sure that a chart including the cleaned data will be shown.

With his credentials, the manager may be able to access the system. The administrator has control over each activation. Once turned on, only that person will have access to our system. The data aggregated may be seen on the administrator's web browser. Algorithm accuracy, complexity matrices, and ROC contours are all under his purview. You may also find a bench graph that shows the relative accuracy here. Once all formula execution is complete, the administrator will have a clear view of the websites' overall accuracy.

Details for the Preparation: Dataset components are often referred to by several names, including records, points, vectors, patterns, events, instances, monitoring, and entities. A number of characteristics that define an entity capture its most salient attributes, including its mass or the moment an event occurred. Variables, attributes, measures, regions, and functions go by many different names. Methods used in the data preparation for this forecast include, but are not limited to, the following: noise removal, missing data removal, default value changes when needed, and the incorporation of features for prediction at multiple levels.

Working with the cleaned-up dataset, we use five AI classifiers: Logistic Regression (LR) with pipe, Assistance Vector Maker (SVM), Decision Tree (DT), and Random Woodland (RF). According to the split requirements, training data must make up 60% and testing data must account for 40%. The accuracy of the classifiers was evaluated using the confusion matrix. It is possible to identify the top classifiers by comparing their accuracy levels.



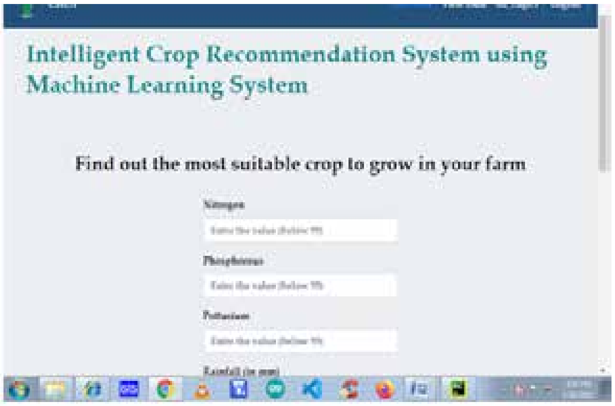
Intelligent Crop Recommendation System using Machine Learning

The proposed system helps the farmers to choose the right crop by providing insights that ordinary farmers don't keep track of thereby decreasing the chances of crop failure and increasing productivity

About Project

Agriculture plays a vital role in the socioeconomic fabric of India. Failure of farmers to decide on the best-suited crop for the land using traditional and non-scientific methods is a serious issue for a country where approximately 58 percent of the population is involved in farming. Sometimes farmers were failed to choose the right crops based on the soil conditions, sowing season, and geographical location. This results in suicide, quitting the agriculture field, moving towards urban

Fig.1. Home page



Intelligent Crop Recommendation System using Machine Learning System

Find out the most suitable crop to grow in your farm

Nitrogen

Enter the value (Below 99)

Phosphorous

Enter the value (Below 99)

Potassium

Enter the value (Below 99)

Rainfall (in mm)

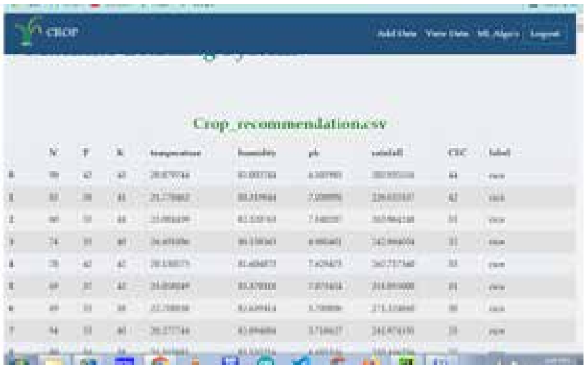
Fig.2. Crop grow materials

Intelligent Crop Recommendation System using Machine Learning System

Datasets Here are:

1. Crop_Recommendation.csv
2. Crop_Sustainability.csv

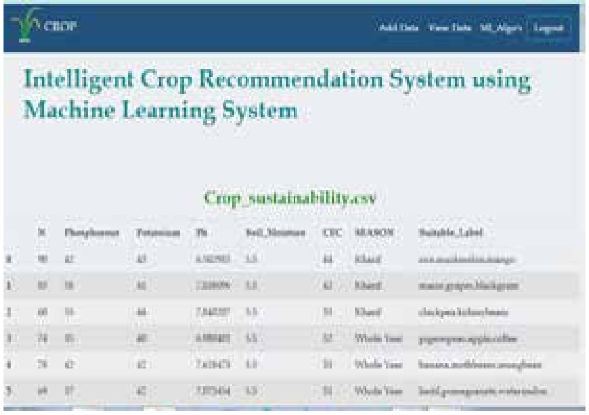
Fig.3. Dataset



Crop_recommendation.csv

	N	P	K	temperature	humidity	ph	rainfall	CEC	label
0	90	42	43	20.879744	82.002744	6.502985	202.935536	44	rice
1	85	58	41	21.770462	80.319644	7.038096	226.655537	42	rice
2	60	55	44	23.004459	82.320763	7.840207	263.964248	33	rice
3	74	35	40	26.491096	80.158363	6.980401	242.864034	33	rice
4	78	42	42	20.130175	81.604873	7.628473	262.717340	33	rice
5	69	37	42	23.058049	83.370118	7.073454	251.055000	33	rice
6	69	55	38	22.708838	82.639414	5.700806	271.324860	38	rice
7	94	53	40	20.277744	82.894086	5.718627	241.974195	35	rice

Fig.4. Crop recommendation

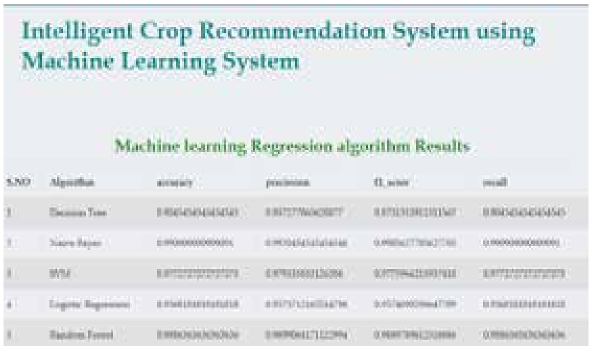


Intelligent Crop Recommendation System using Machine Learning System

Crop_sustainability.csv

	N	Phosphorous	Potassium	Ph	Soil_Moisture	CEC	SEASON	Suitable_Label
0	90	42	43	6.502985	5.5	44	Kharif	rice,muskmelon,mango
1	85	58	41	7.038096	5.5	42	Kharif	maize,grapes,blackgram
2	60	55	44	7.840207	5.5	33	Kharif	chickpea,kidneybeans
3	74	35	40	6.980401	5.5	32	Whole Year	pigeonpeas,apple,coffee
4	78	42	42	7.628473	5.5	33	Whole Year	banana,mothbeans,mungbean
5	69	37	42	7.073454	5.5	33	Whole Year	lentil,pomegranate,watermelon

Fig.5. Crop sustainability



Intelligent Crop Recommendation System using Machine Learning System

Machine learning Regression algorithm Results

S.NO	Algorithm	accuracy	precision	f1_score	recall
1	Decision Tree	[illegible]	[illegible]	[illegible]	[illegible]
2	Naive Bayes	[illegible]	[illegible]	[illegible]	[illegible]
3	SVM	[illegible]	[illegible]	[illegible]	[illegible]
4	Logistic Regression	[illegible]	[illegible]	[illegible]	[illegible]
5	Random Forest	[illegible]	[illegible]	[illegible]	[illegible]

Fig.6. ML based algorithm results

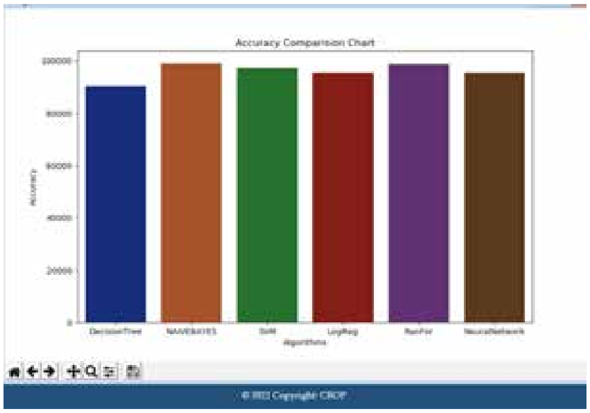


Fig.7. Output results

CONCLUSION

In order to help farmers choose the best plants, the suggested system provides insights that regular farmers overlook, which in turn decreases the likelihood of plant failure and increases performance. They also avoid losing money because of it. An online user interface and a mobile app will soon be integrated to provide farmers with plant growing advice that can be accessible by many farmers around the nation.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "CROP RECOMMENDER SYSTEM USING MACHINE LEARNING APPROACH", which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. R. Kumar, M. P. Singh, P. Kumar, and J. P. Singh, "Crop Select ion Method to maximize crop yield rate using machine learning technique", 2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy, and Materials (ICSTM), Chennai, 2015, pp. 138-145, DOI: 10.1109/ICSTM.2015.7225403.

2. H. Lee and A. Moon, "Development of yield prediction system based on real-time agricultural meteorological information", 16t h International Conference on Advanced Communication Technology, Pyeongchang, 2014, pp. 1292- 1295, DOI: 10.1109/ICACT.2014.6779168.

3. T.R. Lekhaa, "Efficient Crop Yield and Pesticide Prediction for Improving Agricultural Economy using Data Mining Techniques", International Journal of Modern Trends in Engineering and Science (IJMTES), Volume 03, Issue 10, 2016.

4. Jay Gholap, Anurag Ingole, JayeshGohil, Shailesh Gargade and Vahida At t ar, " SoilData Analysis Using Classification Techniques and Soil Attribute Prediction", International Journal of Computer Science Issues, Volume 9, Issue 3,2012.

5. S. R. Rajeswari, Parth Khunteta, Subham Kumar, Amrit Raj Singh, Vaibhav Pandey, "Smart Farming Prediction using 76 Machine Learning", International Journal of Innovative Technology and Exploring Engineering, 2019, Volume-08, Issue07.

6. Z. Doshi, S. Nadkarni, R. Agrawal, and N. Shah, "Agro Consultant : Intelligent Crop Recommendation System Using Machine Learning Algorithms", Fourth International Conference on Computing Control and Automation (ICCUBEA), Pune,India, 2018, pp. 1-6, DOI: 10.1109/ICCUBEA.2018.8697349.

7. S. Pudumalar, E. Ramanujam, R. H. Rajashree, C. Kavya, T. Kiruthika and J. Nisha, "Crop recommendation system for precision agriculture," Eighth International Conference on Advanced Computing (ICoAC), 2017, pp. 32-36, DOI: 10.1109/ICoAC.2017.7951740.

8. Konstantinos G. Liakos, Patrizia Busato, Dimitrios Moshou, Simon P earson and Dionysus Bocht is, "Machine Learning in Agriculture: A Review", Article on sensors, 2018, pp 1 -29, doi:10.3390/s18082674.

9. M. T. Shakoor, K. Rahman, S. N. Rayta and A. Chakrabarty, "Agricultural production output prediction using Supervised Machine Learning techniques", 1st International Conference on Next Generation Computing ions (NextComp), Mauritius, 2017, pp. 182-187, DOI: 10.1109/NEXTCOMP.2017.8016196.

Artificial Intelligence Crime an Overview of Malicious use and Abuse of AI

Ganji Shreeya, Chintala Swapnil Gupta,
K Kasi Viswanath Shastri
B.Tech Students
Department of CSE(AI&ML)
CMR Technical Campus, Medchal
Hyderabad, Telangana

Shaik Sharif
Faculty
Department of CSE(AI&ML)
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ sksharif.cse@cmrtc.ac.in

ABSTRACT

Expert system (AI) capabilities are expanding rapidly and impacting almost every cultural industry. There has been a noticeable uptick in the use of AI in illegal and dangerous activities, which has increased current vulnerabilities and introduced new ones. In order to provide a taxonomy of the detrimental use and abuse of AI-capable systems, this article examines the pertinent literary works, reports, and descriptive events. Sorting out the many jobs and matching dangers is the primary objective. Recognising the weaknesses of AI versions and outlining how malicious actors might exploit them is our starting point. Finally, we analyse attacks that are facilitated by or improved by AI. We do not aim for a final and exhaustive categorization, but we do provide a thorough overview. Instead, we hope that our synopsis of the risks associated with increased AI applications will contribute to the existing literature on the subject. We highlight four specific forms of problematic AI use: social engineering, misinformation/false information, hacking, and autonomous tool systems, as well as four forms of problematic AI abuse: stability assaults, unanticipated AI consequences, mathematics trading, and membership inference attacks. By creating a visual picture of these dangers, we can better understand where we are in terms of governance and what we can do to mitigate or eliminate them. To better prepare for and fight against the harmful use of AI, there has to be more cooperation between federal governments, markets, and civil society groups.

KEYWORDS: *AI, Attacks, Hacking, Fake news, Weapon system.*

INTRODUCTION

As the study of data science grows, machine learning will play an increasingly important role. Statistical methods are used to train different types of algorithms to classify data, generate predictions, and uncover important insights in this field. As a result, these insights should influence critical growth KPIs by driving application and service decision making.

Algorithms used in artificial intelligence create a model using this task data, called training information, so they can make decisions or predictions without being told to. Many datasets make use of machine learning formulae, since traditional formula development for these activities is either too difficult or too costly.

A great deal of academic literature, political discourse, and reports from civic cultural organisations have all pointed to the positive effects of systems that incorporate expert systems (AI) [1, 3, 4]. [5] In fact, AI development has been a source of acclaim due to its exceptional technological capabilities, such as increased potential for automated picture identification (e.g., cancer detection in the medical area) [6, 7]. The uncertain effects of automation on the job market (e.g., issues of mass unemployment [8, pp. 26_27]) are one source of criticism and worry. Cybersecurity and cybercrime provide extra background for determining the pros and cons of contemporary technology.

While governments use AI to bolster their capabilities, the same technology may be used to launch attacks against them [9].

The commercial sector and customer-oriented applications have driven the present spike in AI progress, but sectors like protection might benefit from similar capabilities in their operations [10]. But it's becoming more and harder to separate the actions of states from those of non-state actors. A recent wave of ransomware attacks targeting public infrastructure in several countries, including the Colonial Pipe in the USA in Might 2021 [11, pp. 127_128], has shown this. Even software with good intentions may be run or altered to do harm, even if the original aim was not malevolent. When discussing cybercrime1 and (cyber-) safety, the issue of innovation's dual-use is not new. However, there are distinct vulnerabilities associated with the potential for AI to be used in harmful ways. In order to build and modify governance tools, launch aggressive operations, and enhance (cyber-)strength, it is essential to conduct long-term analyses of the threat environment. This article explores the main categories of AI misuse and usage in criminal contexts in order to build upon prior work [14] _ [16] and get a better grasp of how AI expands the possibility for harmful jobs on the internet. To illustrate the difficulties, we provide a number of notable examples. In light of these instances, we provide a taxonomy that outlines the most harmful AI-based pursuits. Cybersecurity organisations and government agencies may better prepare for potential disasters and assaults by gaining knowledge and understanding about the harmful uses and misuses of AI. Furthermore, a typology is quite useful for organising research projects and pinpointing knowledge gaps that require further investigation.

EXISTING SYSTEM

Machine learning will become more and more important as the field of information science continues to expand. Algorithms of various sorts are trained using statistical methods to classify data, provide predictions, and unearth critical insights in this domain. Consequently, these findings need to influence critical growth KPIs by influencing application and service decision-making. In order for algorithms used in expert systems to generate choices or predictions without human intervention, they first need task information, also known as training information. Due to the difficulty or expense of developing conventional formulas for these tasks, many datasets rely on machine learning formulae.

The positive outcomes of systems that include AI have been highlighted in a great deal of academic writing, political discourse, and reports from civic cultural groups. Indeed, advancements in AI have garnered praise for their remarkable technical capabilities, such as the increased likelihood of automatic picture identification (for instance, cancer diagnosis in a clinical setting). six, seven Concerns about widespread unemployment [8, pp. 26_27] and other unpredictabilities in the labour market are examples of sources of opposition to and anxiety about automation. To better understand the pros and cons of contemporary technology, it is helpful to have some prior knowledge on cybersecurity and cybercrime.

There is a risk that governments would use AI to enhance their capabilities, but the technology may also be used to launch attacks against them. Although the current surge in AI development has been fueled by industrial applications and customer-oriented sectors, other businesses, such as defence, might benefit from similar capabilities in their operations. Determining the behaviour of states apart from non-state stars is becoming more difficult. This was made public in May 2021[11, pp. 127_128] by a ransomware epidemic that hit public frameworks in several nations, including the United States' Colonial Pipeline. Additionally, even software developed with good intentions has the potential to do damage, even if the initial goal was not malicious. Cybercrime1 and (cyber-)safety and security reviews have long addressed the problem of development's dual-use. Still, some risks associated with AI's potentially negative use stand out. It is important to do long-term assessments of the risk environment in order to construct and change administration devices, launch aggressive processes, and increase (cyber-)stability. Building on previous work, this article examines the main types of AI misuse and criminal use to get a better grasp of how AI increases the likelihood of harmful online activities. Several noteworthy examples are provided to emphasise the issues. Given these facts, we provide a taxonomy that details one of the riskiest AI-based endeavours. Cybersecurity companies and government agencies may benefit from learning more about the risks associated with AI in order to better prepare for potential attacks and catastrophes. Typologies also help in organising research tasks and identifying knowledge gaps that need more investigation.

Proposed System

We want to make the following contributions using the typology given in this paper:

a. Contribute to the growing data set that catalogues various forms of abusive AI system use. In order to develop active reactions to these attacks and much-needed preventative measures, it is essential to comprehend the main principles, risk circumstances, and possibilities.

b. Contribute to the development of a common vocabulary across and within fields, particularly in the STEM fields and among legitimate practitioners and legislators. In order to bridge existing gaps and reduce complexity caused by too technical or monodisciplinary terminology, interdisciplinary research on the issue is recommended.

b. Suggest strategies for reduction and show that the market, academics, and government must work together.

This method is based on research into cybercrime and the potential exploitation of artificial intelligence systems. Using the following databases, this investigation and searches are informed by a literature testimonial: Scientific Research Direct, Google Scholar, Wiley Online Library, and IEEE Xplore. We used titles, keywords, and abstracts that had been pre-screened. (Artificial Intelligence, AI, ML, and damaging, criminal, harmful, or cyber attack) are the search phrases that were used. We also looked at news sources that explained previous AI incidents, as well as lists of references from evaluated papers and publications. Only articles, reports, and websites with English and Portuguese versions were considered. We were able to distinguish between several forms of harmful AI use and abuse after reviewing these sources.

The use of machine learning (ML) has grown substantially in recent years. Because of this, ML versions are vulnerable to stability attacks, which may occur when adversaries decide to alter designs (such as the software programme itself) or the underlying data. Hackers attempt to undermine a system's trustworthiness by injecting inaccurate information into it in an integrity attack.

Benefits

The system's goal is to categorise harmful AI uses based on current discussions and actual evidence, specifically looking at how AI systems are used to put data availability, secrecy, and stability at risk.

The goals are limited to identifying key components of AI misuse and damaging applications and gathering evidence of their actual usage. With this data in hand, we can examine the potential ways in which criminals might corrupt AI systems.

IMPLEMENTATION

The Provider is required to provide their authentic credentials in order to access this module. Once he has successfully logged in, he will be able to do activities such as browsing datasets and examination information sets, seeing trained and checked accuracy results, seeing the kind of criminal activity predictions, and seeing educated and checked precision in bar charts. Ratio of Offence Types for Criminal Offences See All Remote Users, Licence Individuals, Download and Install Predicted Data Sets, and View Outcomes by Crime Type Ratio.

The administrator may see the whole roster of registered users in this section. Customers' names, email addresses, and physical addresses are viewable to the admin, who may also provide access to certain users.

Solo Traveller

There are n number of users in this part. Before any operations can be performed, the customer must register. The data source will be updated with the individual's information the moment they join up. Once his registration is complete, he will need to log in using the credentials he was given. Upon successful login, customers will be able to do actions such as registering and logging in, predicting the kind of criminal offence, and seeing their accounts.

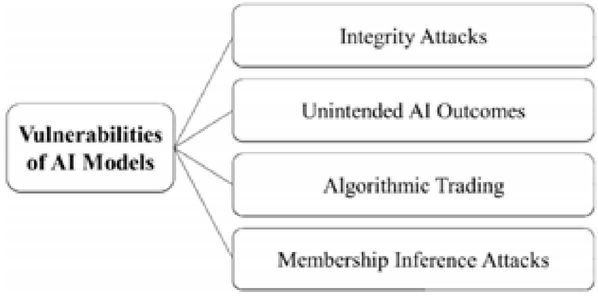


Fig.1. Malicious Abuse of AI

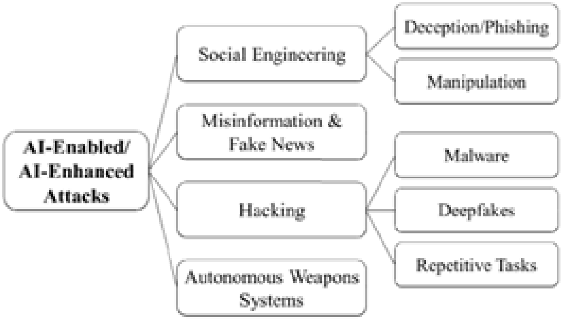


Fig.2. . Malicious Use of AI

CONCLUSION

To develop safeguards to protect cultural and essential frameworks against attacks, it is necessary to have a thorough understanding of the risks posed by AI use and abuse. Using the existing literature, data, and case studies, we set out to categorise the ways in which harmful stars might exploit or misuse AI technologies. Any and all forms of harm, including mental, emotional, political, and financial ones, fall under this category. We looked at AI model weaknesses, such as unforeseen outcomes, and AI-enabled and -enhanced assaults, like imitating. Past events, such as the 2010 _ash collapse and the Cambridge Analytics detraction, are also explained in this article, making the current problems more tangible.Additionally, we described attacks that have only been shown via "proof of principle" so far, such IBM's DeepLocker, as far as we are aware. We have also found several practical ways to lessen the impact of the dangers highlighted in this research. Everyone, from businesses and governments to nonprofits and people, must pitch in to become experts in the field, educate the public, and provide practical solutions to the problems at hand.

This form of classification has certain benefits, but it also has some drawbacks. The established categories were unable to accommodate all attacks that made use of AI or were boosted by AI. To further assess the unreliability and defectiveness of the provided categorization system, further study might use empirical methodologies. Approaches like analytical analysis might be useful to acquire a more complete picture of the danger situation when enough data is provided. It is possible to better prepare for and react to attacks by continuously mapping the threats connected with AI abuse and malicious usage.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "ARTIFICIAL INTELLIGENCE CRIME AN OVERVIEW OF MALICIOUS USE AND ABUSE OF AI", which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department of CSE(AI&ML), Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. K. Crawford, Atlas of AI: Power, Politics, and the Planetary Costs of Arti_cial Intelligence. London, U.K.: Yale Univ. Press, 2021.

2. D. Garcia, ``Lethal arti_cial intelligence and change: The future of international peace and security,'' Int. Stud. Rev., vol. 20, no. 2, pp. 334_341, Jun. 2018, doi: 10.1093/isr/viy029.

3. T. Yigitcanlar, K. Desouza, L. Butler, and F. Roozkhosh, "Contributions and risks of arti_cial intelligence (AI) in building smarter cities: Insights from a systematic review of the literature," Energies, vol. 13, no. 6, p. 1473, Mar. 2020, doi: 10.3390/en13061473.

4. I. van Engelshoven. (Oct. 18, 2019). Speech by Minister Van Engelshoven on Arti_cial Intelligence at UNESCO, on October the 18th in Paris. Government of The Netherlands. Accessed: Apr. 15, 2021. [Online]. Available: https://www.government.nl/documents/speeches/2019/10/18/speech-by-minister-van-engelshoven-on-arti_cial-intelligence-atunesco

5. O. OsobaandW.Welser IV, The Risks of Arti_cial Intelligence to Security and the Future of Work. Santa Monica, CA, USA: RAND Corporation, 2017, doi: 10.7249/PE237.

6. D. Patel, Y. Shah, N. Thakkar, K. Shah, and M. Shah, “Implementation of arti_cial intelligence techniques for cancer detection,” Augmented Hum. Res., vol. 5, no. 1, Dec. 2020, doi: 10.1007/s41133-019-0024-3.

7. A. Rodríguez-Ruiz, E. Krupinski, J.-J. Mordang, K. Schilling, S. H. Heywang-Köbrunner, I. Sechopoulos, and R. M. Mann, ``Detection of breast cancer with mammography: Effect of an arti_cial intelligence support system,'' Radiology, vol. 290, no. 2, pp. 305_314, Feb. 2019, doi: 10.1148/radiol.2018181371.

8. J. Furman and R. Seamans, ``AI and the economy,'' Nat. Bur. Econ. Res., NBER, Cambridge, MA, USA,Work. Paper, 2018, doi: 10.3386/w24689.

9. D. R. Coats, Worldwide Threat Assessment of the U.S. Intelligence Community. New York, NY, USA, 2017, p. 32.

10. L. Floridi, ``Soft ethics: Its application to the general data protection regulation and its dual advantage,'' Philosophy Technol., vol. 31, no. 2, pp. 163_167, Jun. 2018, doi: 10.1007/s13347-018-0315-5.

SCA Sybil-based Collusion Attacks of IIoT Data Poisoning in Federated Learning

Draksharapu Sai Praharshitha,
Gugulothu Praveen, K Joy Wycliff
B.Tech Students
Department of CSE(AI&ML)
CMR Technical Campus, Medchal
Hyderabad, Telangana

Badam Prashanth
Faculty
Department of CSE(AI&ML)
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ prashanth.aiml@cmrtc.ac.in

ABSTRACT

When it comes to the massive amounts of data produced by IIoT instruments at any one time, federated understanding (FL) allows these dispersed, fascinating gadgets to work together to construct a machine learning model while protecting the privacy of the data. Unfortunately, bad actors continue to launch damaging attacks on design gathering's safety and security holes. First of its kind, this article suggests using sybil-based collusion strikes (SCA) on the IIoT-FL system to address the vulnerabilities listed above. In order to finish their local poisoning training, the malevolent participants use label flipping attacks. Simultaneously, they may virtualize many sybil nodes to provide the neighbourhood poisoning schemes the greatest chance of accumulation. While other non-attack classes kept the primary task accuracy similar to the non-poisoned condition, they focused on having the joint design misclassify the selected attack course instances during the screening phase. The results of our SCA's extensive experimental investigation demonstrate that it outperforms the advanced on several metrics.

KEYWORDS: *SCA, FL, Sybil, IIOT, IIOT-FL, Attacks.*

INTRODUCTION

Industrial Internet of Things (IIOT) applications and the rapid development of market 4.0 have led to the proliferation of data produced by industrial devices and the success of smart transport and intelligent healthcare. In order to create a safe combined version for identifying road issues, innovations like autonomous driving [1] must educate all data produced by sensor and electronic camera devices. Distributed IIOT devices may also generate massive amounts of data rapidly [2]. To consider the efficacy of handling massive amounts of data while protecting the confidentiality of customers. A new approach based on distributed training to minimise the efficiency bottleneck and personal privacy danger caused by central processing was proposed, a device learning standard called federated understanding (FL) [3]. Incorporating a large number of smartphones or massive amounts of data into traditional equipment discovery methods [4] often involves centralising the storage and operation of this data, which results in high computational and interaction costs. Due to the need of real-time data transfer, this renders it unsuitable for sensitive IIOT applications, such as autonomous driving, intelligent robotics, and smart medical devices [5]. Furthermore, there is a substantial risk of privacy leaking while depending on central storage space. [6] Customers' private information is usually kept on a local level when FL conducts a collaborative training process including many remote workers (e.g., IIOT tools) [7]. In order to do collaborative calculations in a joint environment with dangerous persons, FL has shown to be very efficient throughout the distributed implementation process, protecting participants' anonymity via autonomous neighbourhood training and version updates. This further increases FL's profile in a number of domains, such as intelligent health care [8] [9], intelligent function prediction [10], and the Internet of Things among smart homes [11] [12].

Each device in the IIOT network may take part in training and upgrades just like any other FL member, making it a dispersed network of intelligent and highly linked industrial instruments [13]. FL improves the design's performance for IIOT applications by continuous iterative training, and attains a stable international design during convergence. The training procedure, however, exposes FL's weak spots to dangerous foes to a large extent [14]. While doing their best to avoid anomaly detection, malevolent opponents may get access to the global design knowledge in each round and either upload harmful criteria or perform a small portion of the beneficial payout for collective training. For example, malevolent foes use tainted data for local training or alter and remove neighbourhood designs for poisoning collection. Previous research [12] has shown that the global version is more vulnerable to attackers that control more malevolent IIOT devices or use more direct poisoning attacks when FL is being executed. In a diverse federated context, many IIOT technologies are at risk of going down due to issues with communication, power, networks, and more. In this unstable communication network, a malicious actor would undoubtedly virtualize several destructive nodes. The current technique of combination sybil-based attacks is often used to address this oriental fault resistance problem, which causes more significant damage to the common international version's construction. Also, when poisoning attacks are being carried out, the bad actors usually train their models using mislabeled samples or upload the infected models to the primary web server to aggregate. Multiple malevolent actors collaborating on an assault has a greater success rate and greater ability to conceal their actions than a single malevolent actor acting alone. However, the central web server cannot verify the neighbourhood data of all participants due to information personal privacy protection characteristics, and all participants' specification transmission procedures are anonymous, which gives harmful participants more opportunities to launch destructive assaults.

Consequently, we provide an effective sybil-based collusion strikes (SCA) mechanism in this work so that the IIOT-FL system may focus on poisoning attack execution. We represent the harmful IIOT device as an adversarial element inside our system. To be more specific, initially, in the FL computing environment that we set up, no one can view anybody else's data; everyone can only manage their own neighbourhood data. Without drawing attention to themselves, they are able to exert more influence on local data used for poisoning training. Two of the most common methods of data poisoning are tag flipping attacks and backdoor poisoning attacks. With the goal of fooling the global design into classifying the selected strike class samples, this study employs tag turning attacks to conduct poisoned training on the massive amounts of data generated by IIOT technologies. However, such an assault lacks the desired striking effect. Second, we take use of sybil's cloning capabilities to ensure that all sybil nodes vitalized by malicious players would carry out identical detrimental actions throughout training and exert equivalent attack impact. To increase the likelihood of the dangerous version being gathered during FL collection, we are considering this. Lastly, we try to replace the global model with the poisoned version by conspiring with all damaging parties to launch the collusion strikes. Simultaneously, the assault behaviour of such coordinated attacks might be much more concealed. To represent the data collected by IIOT devices and conduct experiments, we utilise the Fashion MNIST and CIFAR-10 datasets. To sum up, our main contributions to this task may be categorised into four parts, as shown below.

- In this IIOT-FL system, we find sybil-based collusion attacks of IIOT data poisoning, conduct poisoning training, and construct collusion assaults.

- A We include the tag-flipping poisoning attacks and make minimal destructive assumptions about destructive opponents such that the global model misclassified the selected strike course instances while keeping the main job accuracy of other non-attack classes.

We should provide a sybil-based collusion assaults (SCA) method that effectively covers their striking patterns and makes poisoned collusion models more likely to develop over aggregation.

The data generated by IIOT devices is represented by us using the F-MNIST and CIFAR-10 datasets. Extensive experimental assessment shows that our SCA outperforms the state-of-the-art in terms of efficiency.

EXISTING SYSTEM

Photo images in a distributed heterogeneous dataset were influenced by Xie et al. [1] into making an inaccurate prediction by inserting adversarial triggers into a subset of the training data. In order to degrade the global version's performance on the target work, Sunlight et al. [2] included backdoor tasks into a portion of the photos. Injecting backdoor triggers into huge training instances may be costly, despite its high assault success rate. Furthermore, we want to misclassify the selected assault course samples as part of our attack. For this reason, we use the tactic of poisoning assaults in our work. Even without interacting with criteria, altering the FL architecture, or pre-training, malicious adversaries may launch label-turning attacks. They train in your region using dirty data that has the inappropriate tag. There is an overt and covert strategy to this assault.

According to Jiang et al. [3], a technique based on Sybil was suggested. Customers of Sybil put the infected device at risk in order to directly upgrade the poisoning design. While simultaneously reducing the convergence of the global version, they demonstrated their efficacy on a variety of sophisticated defensive tactics. Additionally, Fung et al. [4] created a new sybil-based attack technique, which has proven successful on several distributed machine learning error tolerance protocols recently. Additionally, the sybil assaults exposed a striking impact on Internet of Things applications. Despite their shown reliability in the striking effect, their regional poisoning design's drift slope is very easy to identify and eliminate. This study introduces the sybil-based collusion attacks innovation, which improves the neighbourhood poisoning model's aggregation probability and makes it easier for malevolent actors to understand attack patterns.

For IIoT applications, Taheri et al. [5] tested two live poisoning strike methods that included participants' use of Generative Adversarial Networks (GANs) and Federated Adversarial Networks (FedGANs). Attacks in which dishonest users work together with the server were studied by Lim et al. [6]. During the gathering phase, the malicious participant uploads the poisoning model, and the server also gives the evil person access to other people's parameters. In order to prevent anomaly detection during the poisoning operation, they want to accomplish the function of reducing the performance of the worldwide version while evaluating the regional versions of other people.

Negative Aspects

To ensure that all Sybil nodes vitalized by destructive persons carry out identical malicious processes throughout training and have equal strike effect, the technique is not used for cloning residential homes. Using SCA on IIoT-FL designs is not done by the system.

PROPOSED SYSTEM

In this IIoT-FL system, the suggested method investigates sybil-based collusion strikes of IIoT information poisoning, models such strikes, and conducts poisoning training. A The suggested approach incorporates label-flipping poisoning strikes and makes few destructive assumptions about destructive enemies; this causes the global design to misclassify samples from some assault classes while maintaining the accuracy of non-attack classes' key tasks. A In addition, the suggested system suggests a sybil-based collusion strikes (SCA) method that effectively obfuscates their striking behaviours and increases the likelihood that poisonous collusion designs would be gathered throughout collection. A The data generated by IIoT tools is represented in the proposed system by the F-MNIST and CIFAR-10 databases. Our SCA outperforms the state-of-the-art in terms of efficiency, according to our extensive experimental investigation.

Benefits

An very safe and secure method, SCA-based tag-flipping poisoning assaults are used by the system. In the proposed system, it is put into action in the event that the malevolent adversary employs the label-turning tactic to teach locals about poisoning and collaborates with other poisoning schemes.

EXECUTION SERVICE OPERATOR

To access this section, the business must provide the correct client ID and password. After he successfully logs in, he will be able to perform things like Peruse Datasets for Networks and Training and Testing, Check out the Accuracy Results for Trained and Tested Network Datasets, See the Precision of Examined and Educated Network Datasets in a Bar Chart, Sight Sybil-based Collusion Attack Condition Ratio, Sight Sybil-based Collusion Assault Status Prediction,

Get your hands on Predicted Data Sets, View Sybil-based Collaborative Strike Results, View All Remote Users, and View Authorised Clients.

This section allows the administrator to see the whole roster of registered users. Here, the administrator may see the data of each user, including their name, email, and address, and can also grant them licences.

Solo Traveller

There are n different types of people in this section. Before any operations can be done, the individual must register. Data will be entered into the database as soon as a consumer signs up. He will need to log in using the authorised username and password after enrollment is completed. Upon successful login, the user will be able to do actions such as registering and logging in, predicting a Sybyl-based conspiracy assault condition, and seeing their profile.

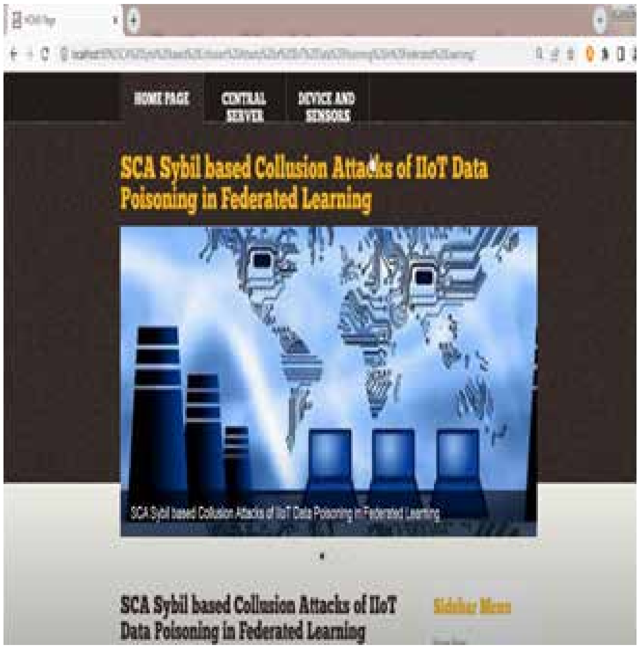

HOME PAGE
CENTRAL SERVER
DEVICE AND SENSORS
SCA Sybil based Collusion Attacks of IIoT Data Poisoning in Federated Learning
SCA Sybil based Collusion Attacks of IIoT Data Poisoning in Federated Learning
SCA Sybil based Collusion Attacks of IIoT Data Poisoning in Federated Learning
Sidebar Menu

Fig.1. Home page

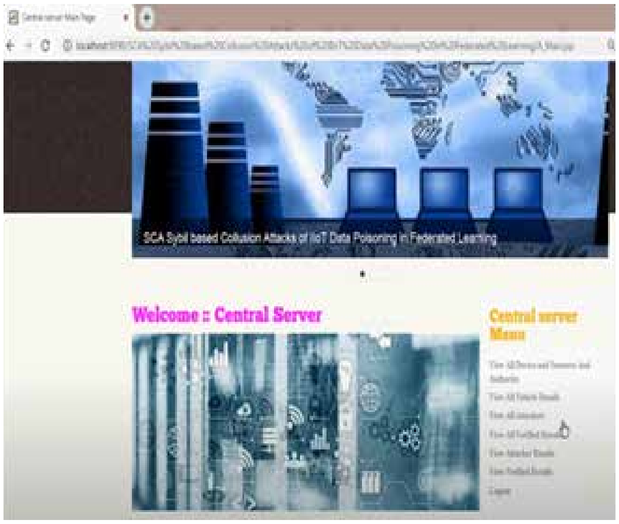

SCA Sybil based Collusion Attacks of IIoT Data Poisoning in Federated Learning
Welcome :: Central Server
Central server Menu

Fig.2. Server details

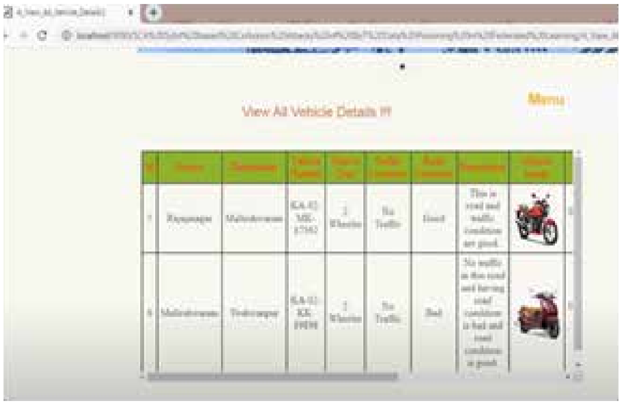


Fig.3. Vehicle details

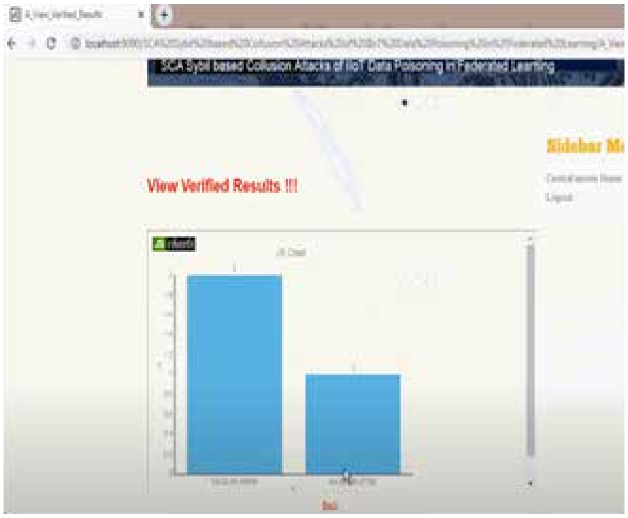
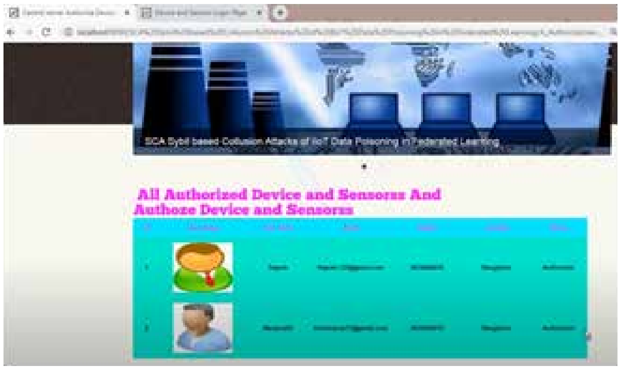

SCA Sybil based Collusion Attacks of IIoT Data Poisoning in Federated Learning
View Verified Results !!!
Sidebar Me
Logout
Bad

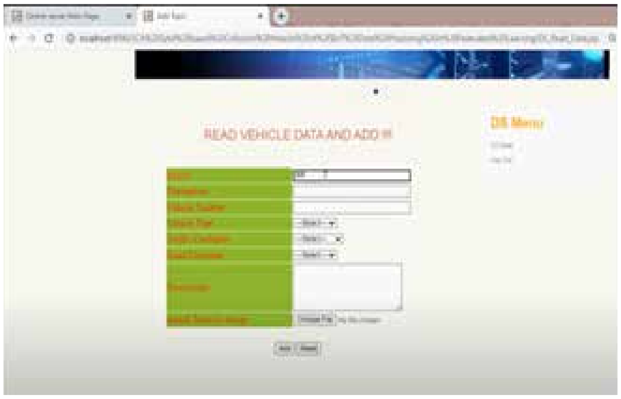
Fig.4. Output results



SCA Sybil based Collusion Attacks of IIoT Data Poisoning in Federated Learning

All Authorized Device and Sensorss And Authoze Device and Sensorss

Fig.5. User details



READ VEHICLE DATA AND ADD !!!

DS Menu

Fig.6. Vehicle data set

CONCLUSION

This study examined the IIOT-FL system's joint training security weaknesses and offered a sybil-based collusion attacks (SCA) approach to exploit them. Simultaneously, we provided further information on the application of relevant formulae, design architecture, and analysis of the experiment's efficiency. Malicious actors in our federated system may vitalize several Sybil nodes and launch coordinated attacks in this role. The goal is to increase the likelihood of the regional poisoning model being collected. While other non-attack courses maintain similar accuracy as before, their goal is to misclassified the instances of the chosen attack course. With fewer harmful actors engaging in cooperation and the ability to properly conceal their attack behaviours, our SCA outperforms the state-of-the-art and achieves an additional large attack outcome. Extensive testing findings demonstrate that our SCA outperforms the competition on several analytical measures when it comes to long-term assault performance.

REFERENCES

1. D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, D. Niyato and H. V. Poor, "Federated learning for industrial internet of things in future industries," IEEE Wireless communications magazine, 2021.
2. P. Zhang, C. Wang, C. Jiang, and Z. Han. "Deep reinforcement learning assisted federated learning algorithm for data management of IIoT," IEEE Transactions on Industrial Informatics (TII), 2021.
3. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), 2017, pp. 1273-1282.
4. Y. Roh, G. Heo, and S. E. Whang, "A survey on data collection for machine learning: A big data - AI integration perspective," IEEE Transactions on Knowledge and Data Engineering (TKDE), vol. 33, no. 4, pp. 1328-1347, 2021.
5. B. Jia, X. Zhang, J. Liu, Y. Zhang, K. Huang, and Y. Liang, "Blockchainenabled federated learning data protection aggregation scheme with differential privacy and homomorphic encryption in IIoT," IEEE Transactions on Industrial Informatics (TII), 2021.
6. V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, "A survey on security and privacy of federated learning," ELSEVIER Future Generation Computer Systems (FGCS), vol. 115, pp. 619-640, 2021.
7. T. Li, A. K. Sahu, A. Talwalker, and V. Smith, "Federated learning: Challenges, methods, and future directions," IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 50-60, 2020.
8. W. S. Zhang, T. Zhou, Q. H. Lu, X. Wang, C. S. Zhu, H. Y. Sun, Z. P. Wang, S. K. Lo, and F. Y. Wang, "Dynamic fusion-based federated learning for COVID-19 detection," IEEE Internet of Things Journal (IoTJ), 2021.
9. M. Parimala, M. S. Swarna, P. V. Quoc, D. Kapal, M. Praveen, T. Gadekallu, and T. H. Thien, "Fusion of federated learning and industrial internet of things: A survey," arXiv preprint arXiv:2101.00798, 2021.
10. M. X. Duan, K. L. Li, A. J. Ouyang, K. N. Win, K. Q. Li and Q. Tian, "EGroupNet: A feature-enhanced network for age estimation with novel age group schemes," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP), vol. 16, no. 2, 2020.

A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos

Nampally Anusha, Guddeti Naresh,
Devaraju Karthik Vaishnav Raju
B.Tech Students
Department of CSE(AI&ML)
CMR Technical Campus, Medchal
Hyderabad, Telangana

Gumpula Aravind
Faculty
Department of CSE(AI&ML)
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ aravind.aiml@cmrtc.ac.in

ABSTRACT

There are now billions of watchers, with the majority being young people, thanks to the meteoric rise of video snippets on YouTube. Even malicious users use this platform as a way to disseminate disturbing images; for example, they may exploit cartoon films to distribute explicit content to youngsters. Consequently, it is highly recommended that social media platforms contain an automated method for screening video clip content in real-time. This study proposes a new approach for detecting and categorising objectionable online material in video clips using deep learning. This is achieved by feeding video clip descriptors into a bidirectional long short-term memory (BiLSTM) network, which in turn finds trustworthy video representations and executes multiclass video category. The design used in this framework is an EfficientNet-B7 pre-trained convolutional neural network (CNN) from ImageNet. Following BiLSTM, a focus device is also added to the network to take use of the interest probability distribution. A dataset of 111,156 animation clips collected from YouTube videos and annotated by hand is used to test these algorithms. Preliminary findings showed that the EfficientNet-BiLSTM framework outperformed the attention device-based EfficientNet-BiLSTM framework (precision D 95.30%) in terms of accuracy (D 95.66%). Second, compared to deep understanding classifiers, typical machine learning classifiers do rather poorly. The combination of Reliable Internet and BiLSTM, which has 128 hidden units, resulted in state-of-the-art performance, as measured by a f1 rating of D 0.9267.

KEYWORDS: *BiLSTM, Deep learning, SVM.*

INTRODUCTION

In actuality, the production and consumption of videos on social media platforms has skyrocketed in recent years. YouTube is the clear leader in social networking platforms if you're seeking for a way to share videos from all across the world. Over 2 billion people have registered on YouTube, and over 500 hours of video footage are posted every minute, according to statistics [1]. As a result, there is a wealth of generic and personalised content accessible to consumers of all ages, with billions of hours of video clips available [2]. Keeping track of and managing the contributed material in accordance with platform requirements is quite tough when dealing with such a large crowd-sourced database. Because of this, dishonest consumers have a better possibility of engaging in spamming operations, which include misleading target audiences with fraudulently marketed material (music, text, etc.). One of the most upsetting things that bad people do is expose young people to unpleasant stuff online, even when it's labelled as safe for them. Now more than ever, kids spend a lot of time online, and YouTube has become a clear alternative to traditional screen media like television for kids [3, 4]. According to the YouTube press release [5], the site's tremendous appeal among younger viewers is due in large part to the lack of restrictions that are in place for older age groups. [6]

Due to the lack of regulations, children may be exposed to any kind of information on the Internet, unlike television. Internet safety concerns include, but are not limited to, the following: cyberbullying, cyberkillers, hate speech, and the exposure of minors to inappropriate material. [7] Constant exposure to upsetting video clip content may have an effect on children's behaviours, emotions, and thoughts, according to Bushman and Huesmann [8]. The tendency of disseminating improper content in children's video clips has been identified in many papers [9] _ [12]. When the mainstream media covered the Elsagate controversy, people became aware of the disturbing scenes on YouTube featuring popular cartoon characters from their childhood—superheroes, Disney personalities, and the like—in which they engaged in light violence, theft, drinking, and nudity or sexual activities [13], [14].

Regulations like the Children's Online Privacy Protection Act (COPPA) mandate that websites have security measures in place to keep minors under the age of thirteen safe while they use the internet. To filter hazardous content, YouTube has also added a "security setting" option. On top of that, YouTube created the YouTube Children app so parents can manage which videos their children may see based on their ratings [15]. Problems in identifying such material mean that disturbing videos continue to appear on YouTube, even in YouTube Kids, despite YouTube's best attempts to curb the phenomenon of harmful web content [16] _ [19]. One possible reason is that YouTube is prone to inappropriate content due to the high volume of videos posted every minute. Additionally, a lot of the video metadata—such as the title, description, view matter, score, tags, comments, and community —is used by YouTube's algorithms to make decisions. To protect children's safety, it is therefore insufficient to simply alter films based on metadata and community _tagging. [21] There are several examples of harmful content on YouTube with seemingly harmless names and images, intended to trick both children and their parents. Destructive uploaders often use the thin insertion of juvenile inappropriate content in videos as a tactic. Figures 1(a), 1(b), and 1(c) show that while the film itself contains improper content, the title and brief video clips are suitable for younger audiences. These videos are concerning since they have millions of views, have been available for years, and have received more likes than dislikes. That goes for a lot of comparable examples. There have also been other cases found (see Fig. 1(d)) where videos or the YouTube network are not well-known but contain content that is harmful to children, particularly cartoons. seems that this issue remains regardless of the channel or video clip appeal based on occurrences. In addition, YouTube has removed the dislike button from videos, so users can't provide indirect comment on video content via statistics. Video clip characteristics, rather than metadata functions associated with videos, should be used for finding of inappropriate content on YouTube since the information may be easily changed [22].

In the past, methods that used common manual operations on frame-level data were able to overcome the difficulty of extracting violent or pornographic video clips from the internet [23] _ [28]. Recent advances in deep learning have led to its application in image and video processing by researchers. One of the most popular applications for categorizing pictures and videos utilizes convolutional semantic networks [29] _ [31]. Furthermore, a particular type of recurrent neural network (RNN) architecture called long-short term memory (LSTM) has been demonstrated through time-series data analysis to be an efficient deep learning version. [32] In order to address the YouTube multiclass video clip category issue, this work uses CNN (Ef_cientNet-B7) and LSTM to discover video clip effective depictions for detection and categorization of inappropriate content. We concentrated on two types of violent and sexually suggestive content that kids see online: one that shows nude individuals and another that contains violent stuff.

There are three primary points that this study brings to the table:

1. For the purpose of finding and classifying inappropriate video material on the web, we propose a new convolutional neural network (CNN) architecture based on biLSTM and eff_cientNet-B7.
2. We provide a ground reality video collection that has been manually annotated. It contains 1860 minutes (111,561 seconds) of animated films designed for children under the age of 13. We culled all of the videos from YouTube by searching for well-known animation titles. Annotations are placed on each video clip to indicate whether

it is safe or hazardous. The dangerous gang uses video recordings to document dream violence and sexually graphic content. Also, we want to provide this dataset to the research community for free.

3. We evaluate the performance of the CNN-BiLSTM architecture that we proposed. A validation accuracy of 95.66% was achieved by our multiclass video classifier. The job of inappropriate video clip online content finding is also evaluated and contrasted with a variety of other state-of-the-art AI and deep learning techniques. Finally, this approach may help any video sharing site either delete the video entirely or blur or conceal any part of the movie that contains harmful content. Additionally, it might be useful for developing parental control systems that can be used with browser extensions or plugins to block inappropriate material for children. What follows is an overview of the post's content: Area II delves into the related operations conducted at this research site. Our suggested system's methodology is detailed in Section III. Section IV contains the proposed system's hypothetical layout. Section V reviews and analyses the results obtained from the hypothetical setting, and Section VI wraps up the work and specifies a future scope for improvements.

EXISTING SYSTEM

a method for identifying illegal content in movies that coupled the extraction of audio functions based on periodicity with aesthetic elements. The most popular application of device locating algorithms is in classifiers. Liu et al. [38] used an assistance vector maker (SVM) approach with a Gaussian radial basis function (RBF) kernel to classify the periodicity-based sound and aesthetic segmentation features. They expanded the structure in later iterations by utilizing the energy envelope (EE), BoW-based audio representations, and aesthetic purposes.

PROPOSED SYSTEM

1. Based on BiLSTM deep learning, the system recommends a novel convolutional neural network (CNN) named EfficientNet-B7 for the identification and classification of unsuitable video clip online content.
2. The system provides 1860 minutes (111,561 seconds) of young children's anime videos (under 13 years old) as a ground truth video dataset. To get all of these videos, we searched popular cartoon characters on YouTube. A "safe" or "dangerous" course annotation is appended to every video clip. Dreams involving physical violence and sexually explicit online material are monitored in videos for the dangerous category. In addition, we want to share this dataset with other researchers by making it publically accessible.
3. Our suggested CNN-BiLSTM structure is evaluated by the system for its efficiency. A validation accuracy of 95.66% was achieved by our multi-class video classifier. Furthermore, a number of potential state-of-the-art architectures for machine learning and deep understanding are assessed and compared for the purpose of identifying inappropriate video online content.

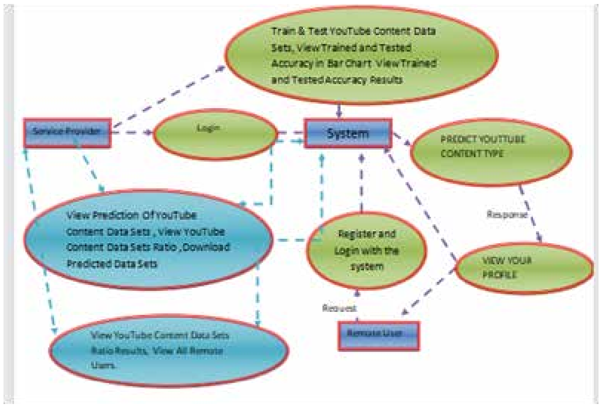
Business Modules

A valid user ID and password are required for the Service Provider to access this module. Once he has successfully logged in, he will be able to perform a number of operations, including: logging in, training and testing YouTube content data sets, viewing trained and tested accuracy in a bar chart, viewing trained and tested accuracy outcomes, viewing predicted data sets, downloading predicted data sets, viewing proportional outcomes for YouTube material data sets, seeing all remote users, and licence individuals.

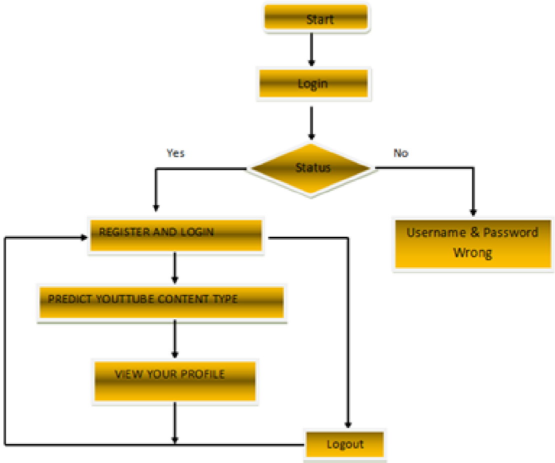
The admin may see a complete list of registered clients in this section. Here, the admin may see client details like name, email, and address, and they can also provide access to certain people.

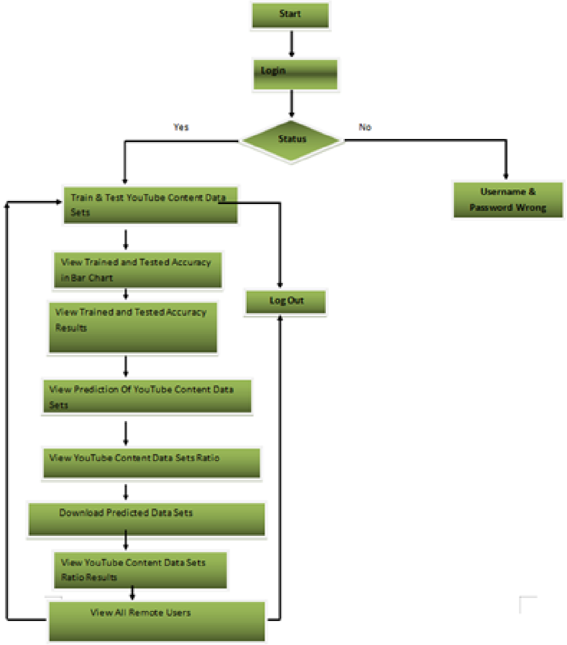
User Working Remotely

There are n different types of clients in this part. Before the customer can undertake any type of operation, they must register. Personal information is entered into the database when a person registers. Once the registration is complete, he will be prompted to provide the authorized user name and password. After logging in successfully, users will be able to do things like see their profile, make content type predictions for YouTube, and register and log in.


Train & Test YouTube Content Data Sets, View Trained and Tested Accuracy in Bar Chart View Trained and Tested Accuracy Results
Service Provider
Login
System
PREDICT YOUTTUBE CONTENT TYPE
Response
View Prediction Of YouTube Content Data Sets , View YouTube Content Data Sets Ratio ,Download Predicted Data Sets
Register and Login with the system
VIEW YOUR PROFILE
Request
Remote User
View YouTube Content Data Sets Ratio Results, View All Remote Users.

➢ **Flow Chart : Remote User**




Start
Login
Yes
Status
No
Train & Test YouTube Content Data Sets
Username & Password Wrong
View Trained and Tested Accuracy in Bar Chart
Log Out
View Trained and Tested Accuracy Results
View Prediction Of YouTube Content Data Sets
View YouTube Content Data Sets Ratio
Download Predicted Data Sets
View YouTube Content Data Sets Ratio Results
View All Remote Users

CONCLUSION

For the purpose of identifying and categorising child-unsuitable video content, this article proposes a new paradigm based on deep learning. To extract video clip qualities, transfer discovering in the EfficientNet-B7 style is used. This version uses the BiLSTM network to process the extracted video clip functions in order to identify the most effective video representations and carry out multi-course video classification. All analytical tests are conducted using a dataset of 111,156 cartoon video clips that were gathered from YouTube and manually annotated. As per the evaluation results, the Reliable Net-BiLSTM framework (with hidden devices D 128) that has been proposed performs better in terms of efficiency (accuracyD95.66%) than the other tested designs, which include Effective Net-FC, Reliable Net-SVM, Effective Net-KNN, Effective Net-Random Forest, and Effective Net-BiLSTM with focus mechanism-based models (with hidden devices D 64, 128, 256, and 512). Furthermore, when compared to other state-of-the-art designs, our BiLSTM-based framework had the greatest recall rate of 92.22%, surpassing all previous versions and methods. The proposed approach for discovering unsuitable video content for children, based on deep learning, has the following benefits:

It uses Ef_cientNet-B7 and a BiLSTM-based deep learning framework to refine the video clip at 22 frames per second while taking into account real-time difficulties. This helps filter the live-captured video clips.

2) Any video sharing platform can utilize it to either completely erase the movie or blur out distracting frames so they are undetectable.

3) It may also be useful for developing parental control features for use with browser extensions or plugins, which would allow for the immediate screening of child-harming content.

Furthermore, our method of identifying objectionable content for children on YouTube does not rely on YouTube video metadata, which may be quickly changed by dishonest uploaders to deceive users. In the future, we aim to connect the temporal stream that employs optical reduction frameworks with the spatial stream of

RGB frames in order to have a deeper understanding of the global depictions of video clips and thus improve the version efficiency. There are many different kinds of inappropriate content for children on YouTube, and we also want to improve the category identifiers to target them.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled “A DEEP LEARNING-BASED APPROACH FOR INAPPROPRIATE CONTENT DETECTION AND CLASSIFICATION OF YOUTUBE VIDEOS”, which gave us excellent resources and assistance to complete our assignment. I would like to express my profound gratitude to our chairman, director, deans, head of the CSE(AI&ML) department, guide, and teaching and non-teaching faculty members for their insightful advice and direction on every facet of our work.

REFERENCES

1. L. Ceci. YouTube Usage Penetration in the United States 2020,by Age Group. Accessed: Nov. 1, 2021. [Online]. Available:https://www.statista.com/statistics/296227/us-youtube-reach-age-gender/
2. P. Covington, J. Adams, and E. Sargin, ``Deep neural networks forYouTube recommendations,'' in Proc. 10th ACM Conf. RecommenderSyst., Sep. 2016, pp. 191_198, doi: 10.1145/2959100.2959190.
3. M. M. Neumann and C. Herodotou, ``Evaluating YouTube videos foryoung children,'' Educ. Inf. Technol., vol. 25, no. 5, pp. 4459_4475,Sep. 2020, doi: 10.1007/s10639-020-10183-7.
4. J. Marsh, L. Law, J. Lahmar, D. Yamada-Rice, B. Parry, and F. Scott,Social Media, Television and Children. Shef_eld, U.K.: Univ. Shef_eld,2019. [Online]. Available: https://www.stac-study.org/downloads/STAC_Full_Report.pdf
5. L. Ceci. YouTube_Statistics& Facts. Accessed: Sep. 01, 2021. [Online].Available: https://www.statista.com/topics/2019/youtube/
6. M. M. Neumann and C. Herodotou, ``Young children and YouTube:A global phenomenon,'' Childhood Educ., vol. 96, no. 4, pp. 72_77,Jul. 2020, doi: 10.1080/00094056.2020.1796459.
7. S. Livingstone, L. Haddon, A. Görzig, and K. Ólafsson, Risks and Safetyon the Internet: The Perspective of European Children: Full Findings andPolicy Implications From the EU Kids Online Survey of 9-16 Year Oldsand Their Parents in 25 Countries. London, U.K.: EU Kids Online, 2011.[Online]. Available: http://eprints.lse.ac.U.K./id/eprint/33731
8. B. J. Bushman and L. R. Huesmann, ``Short-term and long-term effectsof violent media on aggression in children and adults,'' Arch. PediatricsAdolescent Med., vol. 160, no. 4, pp. 348_352, 2006, doi: 10.1001/archpedi.160.4.348.
9. S. Maheshwari. (2017). On YouTube Kids, Startling Videos Slip Past Fil-ters. The New York Times. [Online]. Available: https://www.nytimes.com/2017/11/04/business/media/youtube-kids-paw-patrol.html
10. C. Hou, X. Wu, and G. Wang, ``End-to-end bloody video recognitionby audio-visual feature fusion,'' in Proc. Chin. Conf. Pattern Recognit.Comput. Vis. (PRCV), 2018, pp. 501_510, doi: 10.1007/978-3-030-03398-9_43.

Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection

Seelam Venkata Krishna Reddy,
Kailash Mali, Jangiti Rahul
B.Tech Students
Department of CSE(AI&ML)
CMR Technical Campus, Medchal
Hyderabad, Telangana

Kekkarla Madhu
Faculty
Department of CSE(AI&ML)
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ madhu.aiml@cmrtc.ac.in

ABSTRACT

This study assesses the effectiveness of two machine learning techniques: artificial neural networks (ANN) and support vector machines (SVM). Machine learning techniques will be employed to ascertain whether the request data has regular or atypical signatures. Demand is lowered if request data is found to contain usual or attack signatures by internet intrusion detection systems (IDS). This is required because, in the modern world, all services can be accessed online, and malevolent actors can exploit this to attack web servers or client computers. Using AI techniques, the IDS will learn every conceivable strike signature when new request signatures arrive. This knowledge will be used to assess whether the incoming request includes regular or attack trademarks. Here, we examine and contrast two artificial intelligence algorithms: artificial neural networks (ANN) and support vector machines (SVM). Our experimental results show that ANN outperforms the most advanced SVM at this time in terms of accuracy. Examining how well SVM and ANN work is the focus of this article. By utilising Relationship Based and Chi-Square Based function option formulas, the author has reduced the dataset dimension, eliminated irrelevant data, and loaded the model with important attributes. As a result of these features choice formulas, the dataset dimension will decrease and the forecast accuracy will increase.

KEYWORDS: *ANN, SVM, IDS, Attacks, UTM, IPS.*

INTRODUCTION

The prevalence of cybercrime is directly proportional to the exponential growth in internet use and the ease with which people may access various online resources. first two The first line of defence against a safety strike is intrusion detection. Research investigations are therefore focusing heavily on safety and security services such Intrusion Avoidance Systems (IPS), Unified Hazard Modelling (UTM), Firewall, and Invasion Discovery System (IDS). By gathering data and analysing it for potential security breaches, intrusion detection systems (IDS) may identify attacks from a variety of systems and networks [3]. There are two ways that network-based intrusion detection systems (IDS) evaluate data packets as they travel across a network. Anomaly based detection is a substantial field of study because, even now, it lags significantly behind signature based detection [4-5]. The fact that anomaly-based invasion detection has to cope with new attacks for which there is no precedent in order to detect the irregularity is one of its challenges. Therefore, in order for the system to discern between benign and malicious transmission, some sort of intelligence is required, and researchers have been uncovering artificial intelligence ways to do just that in recent years [6]. Still, intrusion detection systems aren't a panacea for all problems with security. For instance, in the event that there is a hole in the network's procedures or a poor identity and authentication system, IDS will not be able to patch it.

Research on invasion finding began in 1980, and the first edition was released in 1987 [7]. The field of intrusion detection innovation is still in its early stages and has not produced adequate results, despite significant investments made over the past few decades by businesses and academia [7]. Signature-based network intrusion detection systems have gained commercial success and broad acceptance by innovation-based companies worldwide; anomaly-based systems have not seen the same degree of success. Because of this, anomaly-based discovery is currently a focal point for intrusion detection system research and development [8]. Before a broad implementation of anomaly-based intrusion detection systems is contemplated, there are still significant difficulties that must be resolved [8]. However, there is a lack of recent study evaluating the effectiveness of breach detection employing supervised equipment learning methodologies [9]. An advancement in protecting certain networks and systems from hostile activity are anomaly-based network intrusion detection systems. Anomaly detection capabilities have enabled safety and security tools to emerge, but some important issues have not yet been resolved, despite the selection of anomaly-based network invasion discovery methods described in recent literature [8]. Linear regression, support vector machines (SVMs), genetic algorithms, k-nearest neighbour formulas, ignorant bayes classifiers, decision trees, and gaussian mix models are just a few of the anomaly-based approaches proposed [3,5]. Since it has already shown itself on several kinds of problems, support vector machines (SVMs) are among the most used learning formulas [10]. Although all of the aforementioned strategies may identify new attacks, they all have a high failure rate, which is a significant issue with anomaly-based discovery. This is because it might be challenging to extract reasons for normal, sensible behavior from training data sets [11]. Back propagation is still widely employed today to train Artificial Semantic Networks (ANNs), since it has been recognized since 1970 as the opposite setting of automated differentiation [12]. It is exceedingly difficult to assess the effectiveness of network intrusion detection systems in the absence of a comprehensive collection of network-based data [13]. The KDD MUG 99 dataset was used to test the majority of abnormality-based strategies that have been proposed in the literature [14]. Support vector machines (SVMs) and artificial neural networks (ANNs) are two machine learning techniques that were applied in this study on the well-known NSLKDD benchmark dataset for network intrusion [15].

RESEARCH STUDY

A macro-social exploratory study of the cyber-victimization rate across states [1]. This study looks at the relationship between cyber-theft victimisation and signs of macro-level possibilities. In line with the arguments put forward by criminal chance theory, patterns of net accessibility at the state level are used to quantify direct exposure to run the risk. To find out if cyber-victimization varied between states due to differences in social structure, we looked at a number of other structural characteristics of states. Where people get their internet connection is correlated with structural factors like unemployment and the percentage of the population that lives outside of major cities, according to the current research. Additionally, this study found a positive correlation between the proportion of people who solely use their home internet connection and cyber-theft victimisation rates at the state level. Theoretical considerations about these results are addressed.

[2] Use of limited recognised information in a step-by-step anomaly-based breach detection system, As the internet continues to grow and more people across the globe have access to online media, the prevalence of cybercrime is also on the rise. Cybercriminals target both people and businesses nowadays. You may protect yourself with a variety of tools, such as firewall programmes and Invasion Discovery Systems (IDS). In order to prevent packages from passing through unchecked, firewall software acts as a checkpoint. It may even partition the whole network's web traffic in the worst-case scenario. In contrast, intrusion detection systems automate network monitoring. Building intrusion detection systems is quite challenging due to the streaming nature of data in computer networks. Online dataset categorization is suggested as a solution to this issue in this research. This is achieved by using a naïve Bayesian classifier that is trained step-by-step. Moreover, active discovery enables issue solving with a small amount of identifiable data points, which are sometimes quite expensive to gather. Two groups, one dealing with offline activities and the other with online ones, make up the suggested approach. The first one describes data preparation while the final one shows the NADAL online method. We evaluate the suggested method against the NSL-KDD typical dataset-using step-by-step naive Bayesian classifier. The suggested

method has three benefits over the step-by-step ignorant Bayesian approach:(1) it overcomes the streaming data challenge;(2) it reduces the high expenditure of instance labelling; and(3) it boosts accuracy and Kappa. Therefore, the method works well for intrusion detection systems.

the third A method for assessing the security breach detection system via modelling and application, The goal of intrusion detection systems (IDSs) is to identify strikes either in progress or after they have already occurred. Research aimed to do two things: first, examine the system IDS; and second, reduce the impact of attacks. Undoubtedly, intrusion detection systems (IDSs) collect data on website traffic from many network or computer system resources and use it to make systems safer. On the other side, evaluating IDS is a vital task. Considering the components of a system is different from studying the system as a whole in terms of performance. We provide an approach to IDS evaluation in this research that relies on element efficiency determination. To begin, we have proposed an embedded systems-based equipment system to ensure the safe implementation of the IDS SNORT components. Next, we ran it through a test that mimics real-world website traffic and attacks using the Metasploit 3 Framework and Linux KALI (Backtrack). The obtained data demonstrates that the characteristics of these components have a significant impact on the IDS efficiency.

EXISTING SYSTEM

One of the primary challenges in evaluating the effectiveness of network intrusion detection systems is the absence of an extensive network-based data collection [3]. The KDD CUP 99 dataset was used to test most abnormality-based strategies that have been proposed in the literature. Here, we apply two machine learning methods to the well-known standard dataset for network invasion, NSLKDD: support vector machines (SVMs) and artificial neural networks (ANNs).

Interesting are the claims made and the contributions AI has made up to this point. Today, machine learning has presented us with a plethora of reality applications. It would seem that AI will definitely become global ruler in the near future. So, we tested the notion that machine learning techniques may overcome the challenge of discovering new attacks, sometimes known as zero-day attacks, that contemporary technology-enabled enterprises face. We were able to create a version of the monitoring maker that can identify undiscovered network website traffic using the data we collected from the observed traffic. The combination of the SVM and ANN learning algorithms yielded the best classifier in terms of both success rate and accuracy.

PROPOSED SYSTEM

As shown in Figure 1, the suggested system is composed of an algorithm for detecting attributes and an algorithm for selecting them. In order to assign a given set of circumstances to a certain class, attribute selection elements are responsible for extracting the most relevant functions or attributes. Using the results found in the function selection component, the finding formula component builds the necessary knowledge or expertise. The design learns and becomes qualified with the help of the training dataset. After that, the testing dataset is used to evaluate how well the intelligences were able to classify unknown data. Unattended device learning for anomaly detection in network website traffic Two supervised device-finding techniques, namely Sup SVMs and ANNs, are evaluated in this study. The presence of attack (abnormality) fingerprints in demand data may be detected using artificial intelligence algorithms. IDS (Network Invasion Detection System) is used to prevent malicious users from launching attacks on customer or web server devices. The system monitors request information and checks for attack signatures; if it finds any, the request is rejected. Nowadays, everything can be found on the internet. The IDS will be trained with all potential attack signatures using AI formulae, and then a train version will be created. When new demand trademarks appear, this design will be used to determine whether the new demand has typical or attack signatures. The research compares the effectiveness of support vector machines (SVMs) and artificial neural networks (ANNs), and the results show that the latter is more accurate than the former. In order to prevent all attacks, intrusion detection systems have developed procedures to examine each incoming request for signs of attacks. If the request appears to be from legitimate users, the request will be forwarded to the web server for processing. However, if the request contains attack trademarks, the request will be rejected and the information will be recorded in the dataset for future use in discovery. For intrusion detection systems (IDS) to detect these types of assaults, they must first be trained with all possible attack trademarks originating from malevolent individuals' requests. Only then can they create a training model. In order to determine whether a newly received request is part of the normal class or the strike course, IDS will apply the request to the specific train model. In order to train these designs and make predictions, a plethora of information mining prediction algorithms will be used. Reviewing the effectiveness of SVM and ANN is the focus of this research. To reduce the size of the dataset and improve the accuracy of the predictions, the author has used Relationship Based and Chi-Square Based function selection formulas in this algorithm. The feature option algorithms removed unnecessary data from the dataset and replaced it with design with vital functions.

WORKING METHODOLOGY

The creator of the habits test has made use of the NSL KDD Dataset; here are some sample documents containing request trademarks from that dataset. You can find the dataset I used, which is located in the 'dataset' folder, at the same location.

Here are some examples of variables found in datasets: size, protocol_type, service, flag, src_bytes, dst_bytes, land, incorrect_fragment, pushing, hot, num_failed_logins, logged_in, num_compromised, root_shell, su_attempted, num_root, num_file_creations, num_shells, quantity of access files, number of outbound commands, is_host_login, is_guest_login, and matter, servers, error rate, srv_serror_rate, rerror_rate, srv_rerror_rate, same_srv_rate, diff_srv_rate, srv_diff_host_rate, dst_host_count, dst_host_srv_count, dst_host_same_srv_rate, dst_host_diff_srv_rate, dst_host_same_src_port_rate, dst_host_srv_diff_host_rate, dst_host_serror_rate, dst_host_srv_serror_rate, dst_host_rerror_rate, dst_host_srv_rerror_rate, label.

Names of request trademarks are all capitalised and presented in a bright way.

no, tcp, ftp_data, SF,491, zero,0,0,0, zero,0,0,0, absolutely no, no, absolutely no, no, zero,0, absolutely no, zero, absolutely no,2,2, zero,0, zero,0,1,0,0, a hundred and fifty,25, normal,.17,0.03,0.17,0,0,0.05, normal, no.

0, tcp, private, S0,0,0, no, no, no,0,0, absolutely no, absolutely no, no,0,0, no, absolutely no, zero,0, absolutely no, absolutely no, Anamoly, 166,9,1,1, zero, zero.05, zero.06, no, 255, 9,0.04, zero.05, zero,0,1,1,0,0, zero.

The values of the signatures are the facts mentioned above, and the class tag that makes up the staying charge is either a standard request signature or an assault trademark. 'Neptune' is an attack name in the second file. In the same vein, the collection contains over 30 notable assault names.

Some variables in the dataset documents are still in string format; for example, tcp and ftp_data. These values aren't important for the prediction and may be removed using the PREPROCESSING Idea. Because our algorithm would fail miserably if fed attack names in text format, we want to instead provide each attack a number value. A new document called "easy. Txt" will be created for the purpose of creating the education version when all of this is carried out in the PREPROCESS activities.

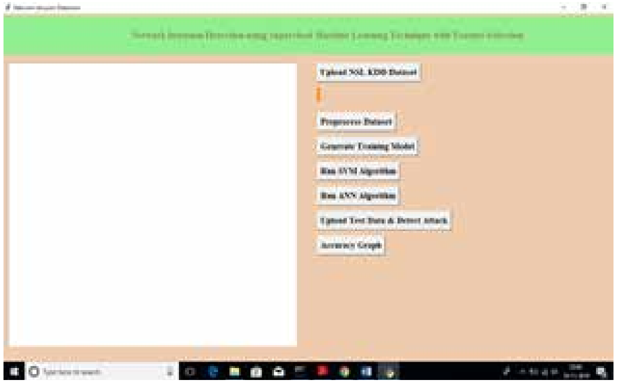
I am assigning number identifiers for each attack in the line below.

In the aforementioned traces, we can see that "day-to-day" has the id "absolutely no" and "anamoly" has the id "1" and exists for all attacks.

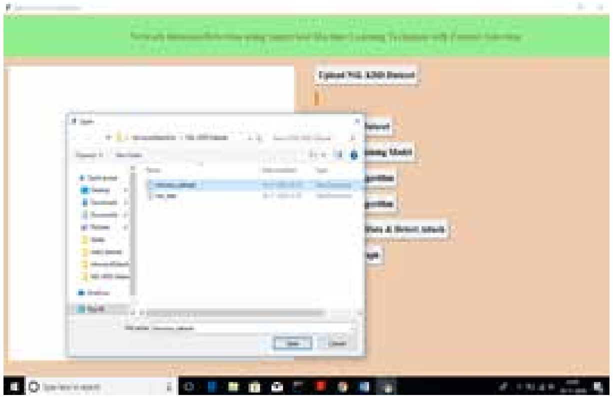
The two instructions below must be executed before any code.

OUTPUT EXPLANATION

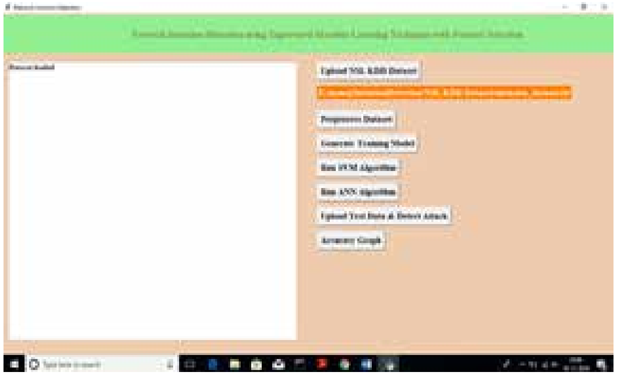
Double click on 'run.bat' file to get below screen


Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection
Upload NSL KDD Dataset
Preprocess Dataset
Generate Training Model
Run SVM Algorithm
Run ANN Algorithm
Upload Test Data & Detect Attack
Accuracy Graph

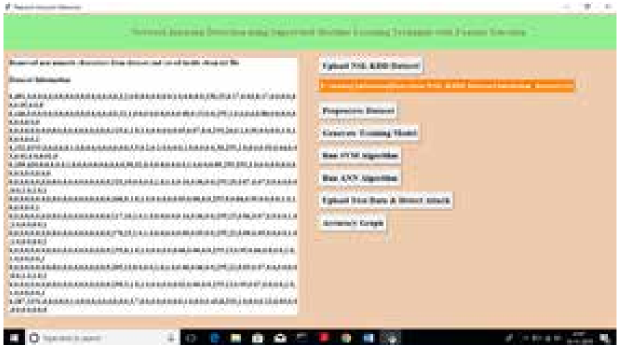
Press the "Upload NSL KDD Dataset" button on the previous screen to add the dataset.



After importing the dataset, the following screen appears underneath the one I was using to submit the "intrusion_dataset.Txt" report:

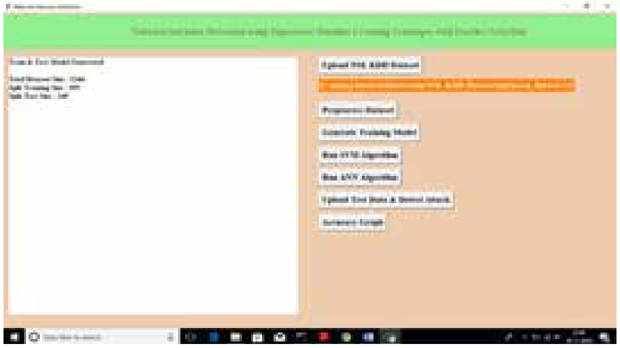


To clean the dataset by converting attack names to numbers and eliminating string values, choose the "Pre-process Dataset" button.

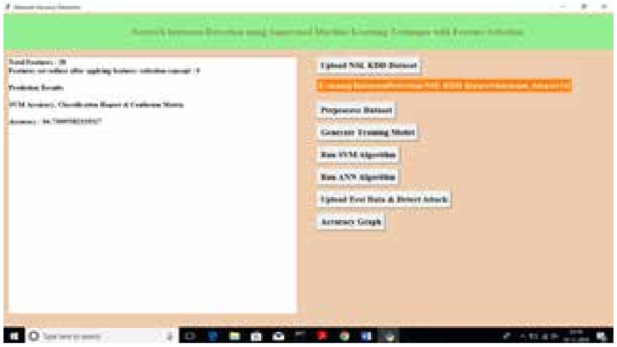


The next step is to pre-process the data by removing any text values and then converting the names of the attacks to numbers. The normal signature has an identifier of zero and the anomalous attack has an identifier of one.

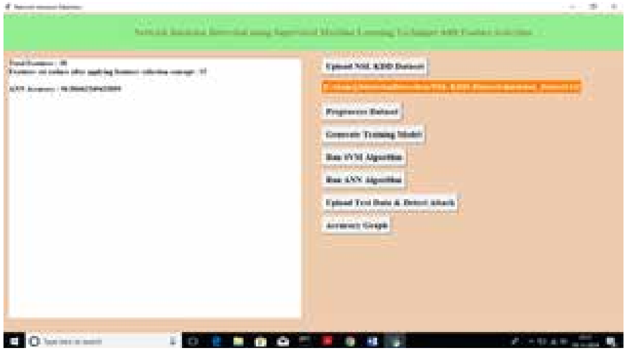
The next step is to use the 'Generate Training Model' button to divide the data into train and examine sets. This will create a version for making predictions using SVM and ANN.



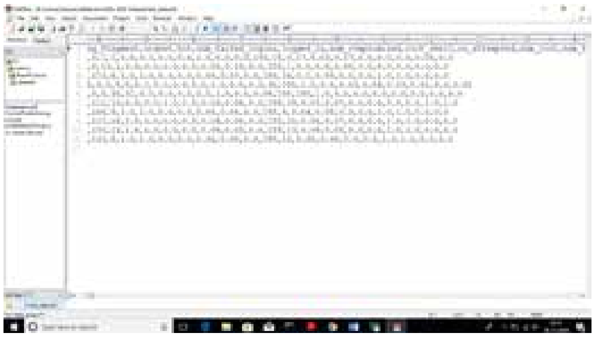
There are a complete of 1244 entries in the dataset, with 995 getting used for schooling and 249 for checking out, as seen in the above display screen. The next step is to construct an SVM model and determine its accuracy through clicking the "Run SVM Algorithm" button.



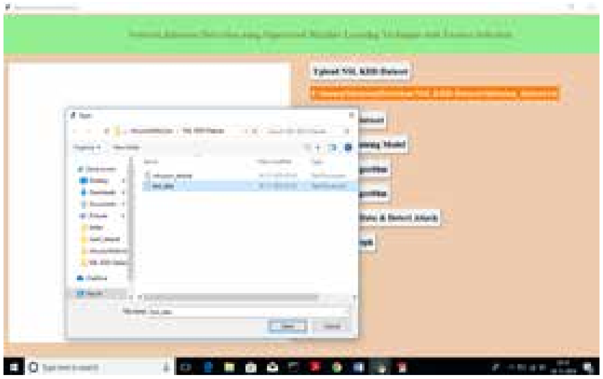
To find the ANN accuracy, click on "Run ANN Algorithm," as shown above; using SVM, we got an accuracy of 84.73%.



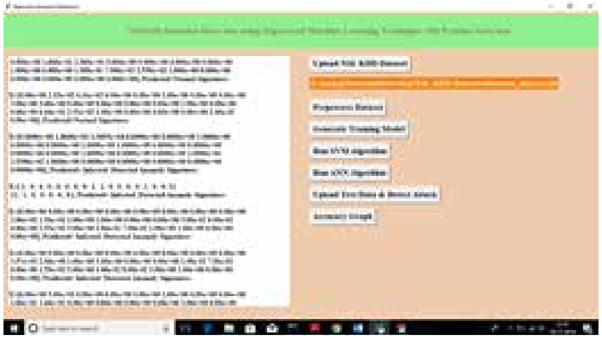
After achieving an accuracy rate of 96.88% in the previous display, we can proceed to submit test data and determine whether it is normal or has a strike by clicking the "Upload Test Data & Detect Strike" button. The programme will undoubtedly make a prediction and provide us with the results, and all of the test data is numeric. See a few documents derived from test data below.



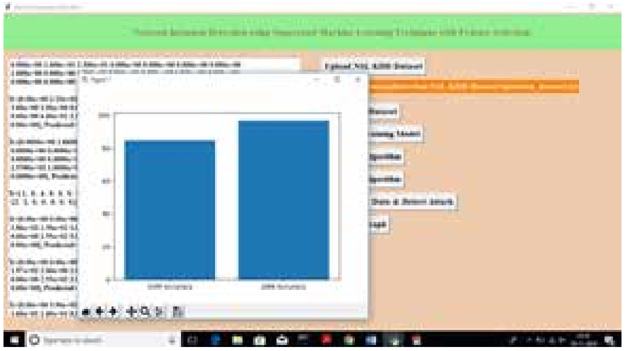
The programme will identify and provide us with results even when the test data above does not include either a 0 or a 1.



I am uploading the 'test_data' file, which includes the test records, to the screen above. After the prediction, I will receive the results below.



We expected each test document to show up in the above display as either "Regular Trademarks" or "contaminated," depending on the information entered. To see a graph comparing the accuracy of SVM and ANN, click the "Precision Chart" button.



The graph above shows that ANN outperformed SVM in terms of accuracy; the x-axis indicates the names of the methods, and the y-axis suggests how properly they achieved.

CONCLUSION

In order to determine the best model, we have provided a number of machine learning options that make use of different maker learning algorithms and function option strategies. Based on the results, the design that used ANN and wrapper feature selection was the most effective in correctly classifying network website traffic, with a detection rate of 94.02%. We anticipate that further research into the topic of building a detection system capable of detecting both known and new attacks will be prompted by these results. At this time, intrusion detection systems can only identify assaults that have already been discovered. Due to the high false positive rate of current systems, finding new attacks, often known as zero day attacks, is an active area of study.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "NETWORK INTRUSION DETECTION USING SUPERVISEDMACHINE LEARNING TECHNIQUE WITH FEATURE SELECTION", which gave us excellent resources and assistance to complete our assignment. I would like to express my profound gratitude to our chairman, director, deans, head of the CSE(AI&ML) department, guide, and teaching and non-teaching faculty members for their insightful advice and direction on every facet of our work.

REFERENCES

1. H. Song, M. J. Lynch, and J. K. Cochran, “A macro-social exploratory analysis of the rate of interstate cyber-victimization,” American Journal of Criminal Justice, vol. 41, no. 3, pp. 583–601, 2016.

2. P. Alaei and F. Noorbehbahani, “Incremental anomaly-based intrusion detection system using limited labeled data,” in Web Research (ICWR), 2017 3th International Conference on, 2017, pp. 178–184.

3. M. Saber, S. Chadli, M. Emharraf, and I. El Farissi, “Modeling and implementation approach to evaluate the intrusion detection system,” in International Conference on Networked Systems, 2015, pp. 513–517.

4. M. Tavallaee, N. Stakhanova, and A. A. Ghorbani, “Toward credible evaluation of anomaly-based intrusion-detection methods,” IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, no. 5, pp. 516–524, 2010.

5. A. S. Ashoor and S. Gore, “Importance of intrusion detection system (IDS),” International Journal of Scientific and Engineering Research, vol. 2, no. 1, pp. 1–4, 2011.

6. M. Zamani and M. Movahedi, “Machine learning techniques for intrusion detection,” arXiv preprint arXiv:1312.2177, 2013.

7. N. Chakraborty, “Intrusion detection system and intrusion prevention system: A comparative study,” International Journal of Computing and Business Research (IJCBR) ISSN (Online), pp. 2229–6166, 2013.

8. P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, “Anomaly-based network intrusion detection: Techniques, systems and challenges,” computers & security, vol. 28, no. 1–2, pp. 18–28, 2009.

9. M. C. Belavagi and B. Muniyal, “Performance evaluation of supervised machine learning algorithms for intrusion detection,” Procedia Computer Science, vol. 89, pp. 117–123, 2016

10. J. Zheng, F. Shen, H. Fan, and J. Zhao, “An online incremental learning support vector machine for large-scale data,” Neural Computing and Applications, vol. 22, no. 5, pp. 1023–1035, 2013.

11. M. Tavallaee, N. Stakhanova and A. A. Ghorbani, “Toward Credible Evaluation of Anomaly-Based Intrusion-Detection Methods,” in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, no. 5, pp. 516-524, Sept. 2010.

12. F. Gharibian and A. A. Ghorbani, “Comparative Study of Supervised Machine Learning Techniques for Intrusion Detection,” Fifth Annual Conference on Communication Networks and Services Research (CNSR ‘07), Fredericton, NB, Canada, 2007, pp. 350-358.

13. N. Moustafa and J. Slay, “UNSWNB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set),” 2015 Military Communications and Information Systems Conference (MilCIS), Canberra, ACT, Australia, 2015, pp. 1-6.

14. T. Janarthanan and S. Zargari, “Feature selection in UNSW-NB15 and KDDCUP’99 datasets,” 2017 IEEE 26th International Symposium on Industrial Electronics (ISIE), Edinburgh, UK, 2017, pp. 1881-1886.

15. M. Panda, A. Abraham and M. R. Patra, “Discriminative multinomial Naïve Bayes for network intrusion detection,” 2010 Sixth International Conference on Information Assurance and Security, Atlanta, GA, USA, 2010, pp. 5-10.

Students Performance Prediction in Online Courses using Machine Learning Algorithms

Bulusu Anagha Krishna, Mittapally Vamshi Krishna, V Shiva Kumar Reddy
B.Tech Student
Department of CSE (AI&ML)
CMR Technical Campus, Medchal
Hyderabad, Telangana

Jagiri Sushmitha
Assistant Professor
Department of CSE(AI&ML)
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ sushmitha.aiml@cmrtc.ac.in

ABSTRACT

Progress in Communicating and Detailing Thanks to advancements in information and communication technology, massive open online courses (MOOCs) have flourished and are now widely employed in online education. In order to encourage students to develop new cognitive skills, a variety of techniques have been used to provide interactive content including images, figures, and videos. Some of the world's most prestigious universities have begun offering massive open online courses (MOOCs) to students from all around the globe. Students' progress is assessed via the use of predetermined, computer-marked tests. In particular, when the student completes the online assessments, the computer immediately provides feedback. According to the study's authors, students' engagement and performance in the prior session have a bearing on how likely they are to succeed in an online course. The literature has not focused enough on the question of whether or how students' past performance and engagement on exams could affect their future success on those same examinations. Two anticipatory versions, focusing on students' assessment grades and final trainees' efficiency, have been developed in this research. Using the designs, we can isolate the factors influencing students' discovery success in massive open online courses (MOOCs). Both models provide accurate and feasible outcomes, as shown by the findings. For the students' analysis grades model, the most cost-effective RSME gain was 8.131 for RF, while for the final students' performance, GBM produced the most accurate results, with an average value of 0.086.

KEYWORDS: *MOOC, PIC, ML, DL, RSME, GBM, RF.*

INTRODUCTION

The Massive Open Online Courses (MOOCs) are among the most popular kinds of online education. The training courses offered by massive open online courses (MOOCs) make use of a variety of electronic tool items, including visual, audio, video, and plain text. Rather of reading lengthy plain-text publications, many students find that video clip speeches help them better comprehend course content. Interactive videos in massive open online courses (MOOCs) have the potential to ease students' minds, make them feel more at ease, and speed up their learning. the first two There are two main categories of massive open online courses (MOOCs): eXtended large open online courses (xMOOCs) and connectivity massive open online courses (cMOOCs). The principles of cognitive behaviourism are the foundation of the new paradigm that the xMOOCs are revealing. [4] The programme are structured similarly to the typical training course, with a final exam, a series of multiple-choice quizzes, and video lectures making up the curriculum. Once a week, students may see video lectures in which the training course instructor goes over the material from the previous online session. People may watch the film at their own leisure and pause it whenever they choose. In addition, by posting in discussion boards, students may socialise with the instructor and other participants. Discussion online forums play an important part in improving the course quality and making online sessions collaborative and attractive since instructors generally express problems, suggest task solutions, and react to student complaints via these forums. the third [5] A new iteration of massive open online courses (MOOCs) based on connectives theory of learning [3] [4] Trainees acquire the training course curriculum by asking questions and sharing this information with other participants using the collectivism approach; the instructor does not provide the actual knowledge material. Works Cited [3] [4] Assume for the sake of argument that massive open online courses (MOOCs) rely on a collaborative approach to discovery, with students working together to create a final output that is both reusable and remixable before sending it on to future students. Unlike xMOOCs, where college lecturers may utilise computer-marked analytical replies to gauge students' knowledge, cMOOCs make it hard to include expertise in assessing students' comprehension. In particular, when the student completes the online analysis, the computer system provides immediate feedback. Upon successful completion, the student will undoubtedly get their xMOOC accreditation. There is no official assessment in the cMOOCs. This is why massive open online courses (cMOOCs) are not an option for academic institutions. [5] [6]

The rapid development of AI in recent years has made it a viable option for screening and evaluating students' performance in online courses. There have been few efforts to evaluate the performance of the trajectories, despite the fact that some researchers have utilised machine learning to predict student success in [7]. [8] Consequently, educators are unable to monitor students' actual learning curves in real time. This paper conducts two sets of experiments. Students' assessment ratings are estimated using regression analysis in the first set of experiments. To forecast the trainee's outcome, we look at their present and previous actions, as well as their performance. The second round of trials aims to predict students' long-term performance using a monitored machine learning strategy. There are three types of prospect predictors that have been considered: behaviour functions, group traits, and temporal functions. The proposed revisions let teachers keep tabs on students' immediate performance and give fresh understanding of the most important learning activity. To the best of our knowledge, the online programme has only ever used two outcomes—"success" and "fail"—to evaluate student achievement. "Success," "fail," and "took out" are the three-class labels that our version uses to predict the performance.

Based on the notion of Product Action Theory, the Element Evaluation Design (FAM) was suggested as a way to forecast the trainee's performance in an Intelligent Tutoring System (ITS) by considering the degree of difficulty of evaluations. [9] [10] One facet of the relationship between students' performance and evaluation queries may be inferred from the difficulty level of assignments. The FAM specifies a collection of forecaster variables that include the amount of chances given to the student for each assignment, the time spent on each action, and the difficulty level of each inquiry or hidden variable in order to calculate the possibility of a student correctly completing a task. The findings show that the version may be much improved by including the hidden factors into the student performance estimations. [10]

SURVEY OF RESEARCH

Researchers have shown that a combination of Discovering Analytics (LAs) and machine learning provides a promising way to map trainee understanding, which might help them assess the impact of learner activities on their understanding success in massive open online courses (MOOCs). Scientists were able to conceptualise and assess data collected at each level of the student's understanding process, proving that AI may aid the teacher in providing friend details on the learning process. As a result, these programmes can provide an accurate rendition of the forecast. the eleventh [12] in [13] Students' social factors, exam scores, and first-analysis grades are used to forecast how well they would do in an online class. [14]

We introduced two models that can predict outcomes. Whether trainees obtained a standard or distinction certification was predicted in the first edition using logistic regression. The second design also used logistic regression to predict whether students met the qualifying criteria or not. The results showed that the amount of peer reviews is a significant factor in determining the quality of the results. In order to get a certificate, average test results were thought to be a reliable indicator. With a percentage of 92.6% for the first model and 79.6%

for the second design, particularly, the accuracy of differentiation and regular versions were recorded. [14]

Researchers at UMBC looked at the correlation between student performance and data collected from virtual learning environments (VLEs) [12]. It was LA was achieved by use of the Inspect My Activity (CMA) application. CMA is a LA tool that provides regular feedback on trainees' feelings and compares their VLE activities to others. Students who participated in the programme more often compared to those who did not to have a higher likelihood of receiving a grade of C or above [12].

PROPOSED SYSTEM

Synopsis of Contents

The OULAD dataset was extracted from the OULAD database, which stands for Open University UnderstandingAnalyticsDataset.Onlinetrainingcourses in a variety of subjects are available to undergraduate and graduate students from the Open University in the United Kingdom during the academic year 2013–2014. "Trainee Info" is the primary composite table that is linked to all the tables. The "student Info" table includes data that is pertinent to the demographic characteristics of students. [15] Pupil Analysis and "Evaluations" tables compile data pertaining to students' academic progress. For each module, you may find the necessary number, weight, and kind of assessments in the "Assessments" table. In most cases, the final exam follows a set of analyses that are included in each component. Both Tutor Marked Analysis (TMA) and Computer System Marked Analysis (CMA) are part of the evaluations. All tests (50%) and final exams (50%) are used to calculate the last ordinary quality. Trainee and analysis mark data is included in the "Student Analysis" table. [15] The "Trainee Registration" table contains information on the trainees' sign-up dates, regardless of whether they are included in a specific module or not. The total number of days is determined by keeping track of the number of separate days that students access the programmed until the training course concludes. While students enrolled in an Open University online course may view a module before the start of the training, they will no longer have access to the programme after the course has ended. Data collected from students' Virtual Knowing Atmospheres pertains to their interactions with technological devices.

This case study includes the results of a number of experiments designed to improve students' efficiency using the Model 2. The first experiment uses characteristics of students' dynamic behaviour to predict how well they would do in class, whereas the second uses characteristics of students' static behaviour. The issues manifest as regression and categorization. When aiming to forecast students' analytical abilities, the regression setup is considered; when aiming to forecast students' overall programme performance, the classification setup is used. Considered here is a multi-class problem whose goal is the success or failure of training courses. Instructors might be able to help students with poor ratings more quickly by using early quality prediction to find further discoveries goods and treatment support. [7] The learner is required to complete the final exam in addition to the five CMA analyses and six TMA assessments that have already been covered. There is a strict deadline for submitting the analyses. Our temporal analysis is based upon the TMA entrance date since the TMA assessment considers 45% of the final outcome while the CMA analysis only considers 5%. Students' efficiency is quickly asserted in the initial set of trials. As a result, the training course is divided into six time periods, with evaluation entry days relating to each. On the day of the assessment, the students' behaviour records are handed out. This evaluation takes into account the student's performance on previous assessments related to their communication acts.

We combined the child's behavioural tasks from all six time pieces into one single time piece to evaluate the trajectories of pupil efficiency in the second set of studies. As input variables, we make use of the behavioural, market, and temporal functions. Our calculations did not take into consideration the grades received on previous exams or the final test score, which are factors in determining the target course. Out of 4004 papers in the dataset, 28% fall into the "stop working" category, 40% into the "withdrawn" category, and 32% into the "pass" category.

The selection of features

We will only analyse the included option during the first set of experiments since our aim is to explore how child performance in the previous assessment affects the following evaluation. This case study makes use of Recursive Function Elimination (RFE). One of the most popular methods for selecting wrapper functions is reclusive feature elimination. You might think of RFE as an optimisation algorithm that uses resampling and reverse selection as its foundation. Up until it obtains a minimal set of characteristics, it continues to create the model iterative. There are a number of components that distinguish the data set into train and bootstrap samples. We choose the algorithms at each model based on their importance. We keep using the new design, which includes one of the most important forecasts, to test the probability of ranking functions until we're all exhausted.

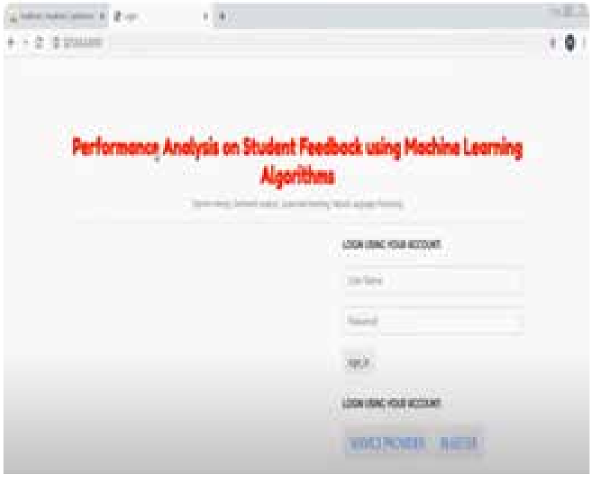


Fig.1. Home page

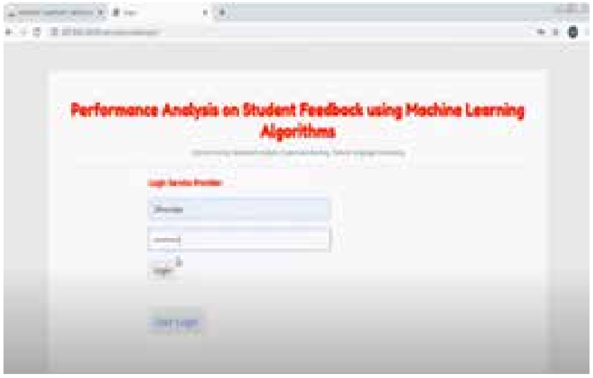

Performance Analysis on Student Feedback using Machine Learning Algorithms
Login Service Provider
Login
User Login

Fig.2. Admin login page

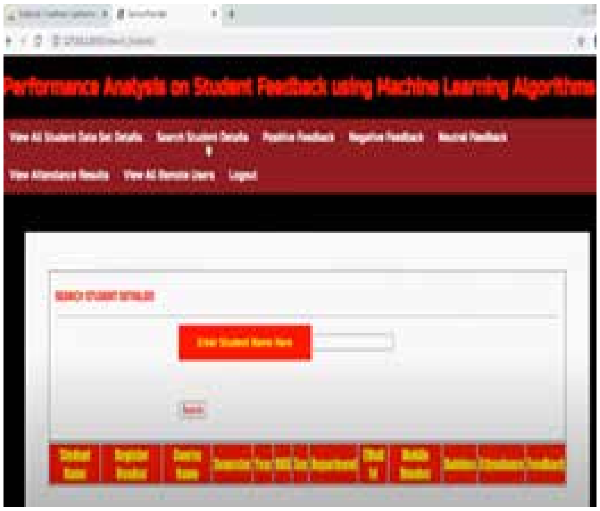


Fig.3. Search student details

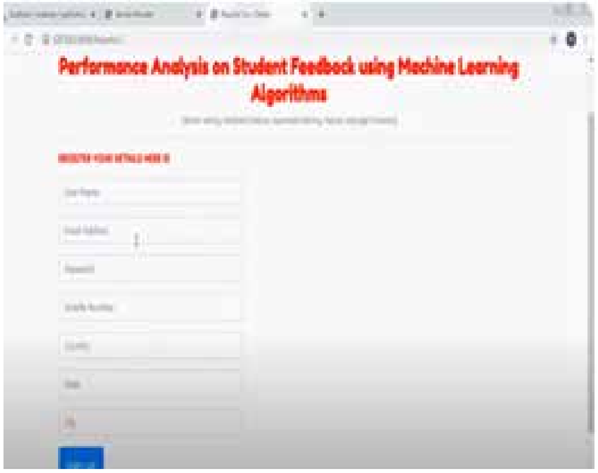

Performance Analysis on Student Feedback using Machine Learning Algorithms
REGISTER YOUR DETAILS HERE !!!
User Name
Email Address
Password
Mobile Number
Country
State
City

Fig.4. Registration details

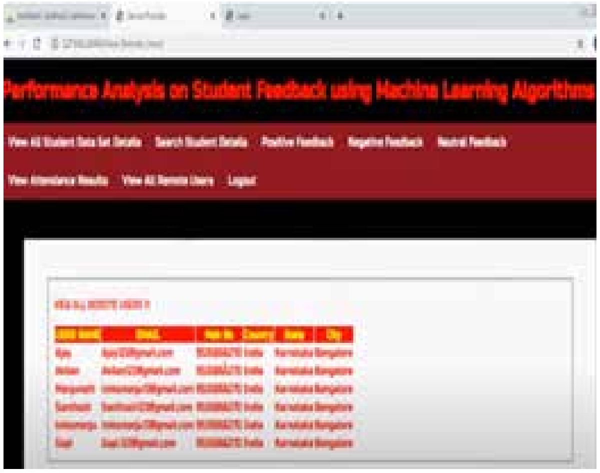


Fig.5. Users details

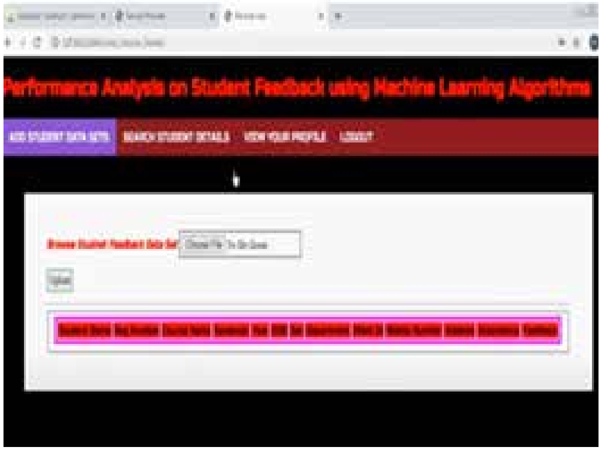

Performance Analysis on Student Feedback using Machine Learning Algorithms
ADD STUDENT DATA SETS
SEARCH STUDENT DETAILS
VIEW YOUR PROFILE
LOGOUT
Browse Student Feedback Data Set
Upload

Fig.6. upload the data set.

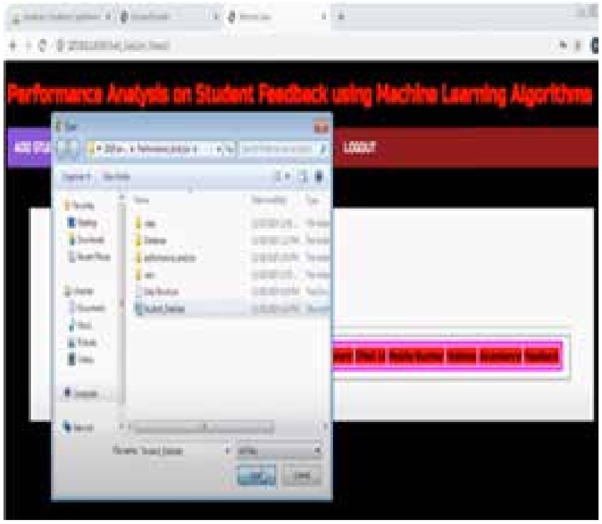

Performance Analysis on Student Feedback using Machine Learning Algorithms
LOGOUT

Fig.7. Load dataset

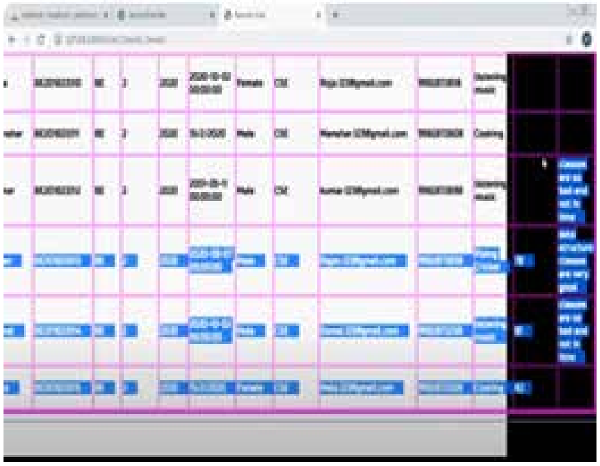


Fig.8. Student details

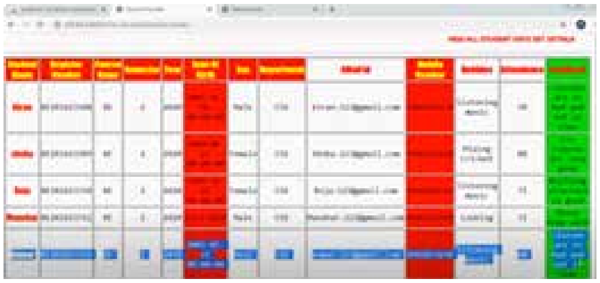


Fig.9. Output results

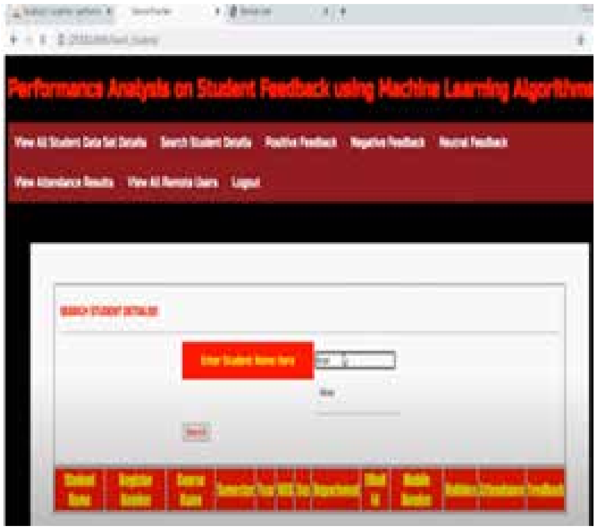


Fig.10 Output with student details

CONCLUSION

In this research, two sets of analyses were conducted using regression and category assessment. The results of designing assessments based on students' attributes show that, in solo programme, students' performance on one assignment is dependent on their performance on another. The study's authors draw the conclusion that, in a traditional classroom setting, students are more likely to drop out of subsequent classes if their prior grade point average (GPA) is poor. In terms of the reliability of past performance into future understanding achievement, traditional and online classrooms are similar.

Students' engagement with course materials is shown to have a major role in their overall achievement, according to the most recent predictive version of student performance. Also, since temporal functions aren't a part of regression assessment, the results demonstrate that long-term trainee efficiency achieves far superior accuracy than students' evaluations grading prediction design. A useful predictor that is highly associated with trainee efficiency is the day of student registration from the training course. The data does not include the latest date of students' activity before taking the assessments, which is a problem with the regression evaluation. Consideration of time elements on expecting of subsequent assessment grades has actually been suggested for the searches.

One potential avenue for future research is to utilise temporal features to predict how students will do on examinations. Potentially, more sophisticated machine learning will be used in place of time collection assessment when dealing with temporal attributes.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "STUDENTS PERFORMANCE PREDICTION IN ONLINE COURSES USING MACHINE LEARNING ALGORITHMS", which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department of CSE(AI&ML), Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. K. F. Hew and W. S. Cheung, "Students' and instructors' use of massive open online courses (MOOCs): Motivations and challenges," Educ. Res. Rev., vol. 12, pp. 45–58, 2014.

2. H. B. Shapiro, C. H. Lee, N. E. Wyman Roth, K. Li, M. Çetinkaya-Rundel, and D. A. Canelas, "Understanding the massive open online course (MOOC) student experience: An examination of attitudes, motivations, and barriers," Comput. Educ., vol. 110, pp. 35–50, 2017.

3. J. Renz, F. Schwerer, and C. Meinel, "openSAP: Evaluating xMOOC Usage and Challenges for Scalable and Open Enterprise Education.," Int. J. Adv. Corp. Learn., vol. 9, no. 2, pp. 34–39, 2016.

4. S. Li, Q. Tang, and Y. Zhang, "A Case Study on Learning Difficulties and Corresponding Supports for Learning in cMOOCs| Une étude de cas sur les difficultés d'apprentissage et le soutien correspondant pour l'apprentissage dans les cMOOC," Can. J. Learn. Technol. Rev. Can. l'apprentissage la Technol., vol. 42, no. 2, 2016.

5. S. Zutshi, S. O'Hare, and A. Rodafinos, "Experiences in MOOCs: The Perspective of Students," Am. J. Distance Educ., vol. 27, no. 4, pp. 218–227, 2013.

6. Z. Wang, T. Anderson, L. Chen, and E. Barbera, "Interaction pattern analysis in cMOOCs based on the connectivist interaction and engagement framework," Br. J. Educ. Technol., vol. 48, no. 2, pp. 683–699, 2017.

7. W. Xing and D. Du, "Dropout Prediction in MOOCs: Using Deep Learning for Personalized Intervention," J. Educ. Comput. Res. p.0735633118757015., 2018.

8. M. J. Gallego Arrufat, V. Gamiz Sanchez, and E. Gutierrez Santiuste, "Trends in Assessment in Massive Open Online Courses," Educ. Xx1, vol. 18, no. 2, pp. 77–96, 2015.

9. B. J. Hao Cen, Kenneth Koedinger and Carnegie, "Learning Factors Analysis – A General Method for Cognitive Model Evaluation and Improvement," in In International Conference on Intelligent Tutoring Systems, 2006, vol. 8, pp. 164–175.

10. K. R. Koedinger, E. A. Mclaughlin, and J. Stamper, "Automated Student Model Improvement," in Educational Data Mining, proceedings of the 5th International Conference on, 2012, pp. 17–24.

11. J. Sinclair and S. Kalvala, "Student engagement in massive open online courses," Int. J. Learn. Technol., vol. 11, no. 3, pp. 218–237, 2016.

12. J. Mullan, "Learning Analytics in Higher Education," London, 2016.

13. P. and K. Al-Shabandar, R., Hussain, A.J., Liatsis, "Detecting At-Risk Students With Early Interventions Using Machine Learning Techniques," IEEE Access, vol. 7, pp. 149464–149478, 2019.

14. S. Jiang, A. E. Williams, K. Schenke, M. Warschauer, and D. O. Dowd, "Predicting MOOC Performance with Week 1 Behavior," in Proceedings of the 7th International Conference on Educational Data Mining (EDM), 2014, pp. 273–275.

5. L. Analytics and C. Exchange, "OU Analyse : Analysing at - risk students at The Open University," in in Conference, 5th International Learning Analytics and Knowledge (LAK) (ed.), 2015, no. October 2014.

Defensive Modeling of Fake News through Online Social Networks

Dommeti Bhumeera, Bhaireddy Bhavana,
Gatla Sai Kiran
B.Tech Student
Department of CSE (AI&ML)
CMR Technical Campus, Medchal
Hyderabad, Telangana

Nagarapu Sateesh
Assistant Professor
Department of CSE(AI&ML)
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ sateesh.aiml@cmrtc.ac.in

ABSTRACT

In today's internet-driven world, where news can be found almost everywhere, people rely on online sources for their news. With the rise of social media platforms like Facebook and Twitter, misinformation spreads quickly among many users. This has serious consequences, such as the influence of biassed opinions on political election outcomes, and spammers use misleading headlines to generate clickbait ads. In this paper, we will use concepts from Artificial Intelligence, Natural Language Processing, and Artificial Intelligence to classify online newspaper articles into binary categories.

KEYWORDS: *Fake news, AI, Facebook, Twitter.*

INTRODUCTION

Due to the increasing amount of time spent online on various social media platforms, more and more people are opting to get their news from these sources rather than traditional wire services. [1] Within the framework of those social media platforms, the explanations for this shift in behaviour are crucial: (1) consuming information on social media sites is often faster and cheaper than traditional journalism (e.g., newspapers or television); and (2) sharing, reviewing, and discussing the information with friends or other readers is easier on social networking sites. Case in point: in 2016, 62% of American adults accessed news via social media, compared to 49% in 2012 who did the same. [1] The fact that social networks now surpass television as a major knowledge resource was also discovered. The quality of news on social networking sites is lower than that of traditional wire service, notwithstanding the benefits of these networks. On the other hand, many fake details, or news articles with intentionally false information, are created online for various reasons, including financial and political gain, because it is cheaper to provide information online and much easier to disseminate with social media websites. Preliminary estimates indicate that more than one million tweets have been associated with the "Pizzagate" hoax by the time the federal political election concludes. Given the prevalence of this new fad, the Macquarie thesaurus even named "phoney information" the 2016 word of the year [2]. The widespread dissemination of synthetic data has the potential to significantly negatively impact both individuals and society. In the 2016 U.S. presidential election, for instance, one of the most infamous pieces of fake news appeared to have far more Facebook shares than one of the most widely acknowledged pieces of legitimate mainstream news. This highlights the potential impact of fake news on the credibility balance of the news environment. In addition, false news sites promote bigotry and logical errors by inviting viewers to just approve them. Some evidence suggests that Russia has indeed used social media and fake accounts to disseminate false articles, as is typical of propagandists who manage counterfeit information to propagate political agendas or exert influence. Third, false information alters how people interpret and react to real news; for instance, some fake news is intentionally designed to make people wonder and be confused, which inhibits their ability to distinguish between fact and fiction. For the benefit of both the public and the details environment, and to mitigate the negative effects of false information. We need to come up with ways to spot fake news programmes on social media platforms fast.

LITERATURE SURVEY

In their research [3], Mykhailo Granik et. al. demonstrate a simple approach to detecting false news using a naïve Bayes classifier. This technique was deployed as a software program system and assessed against a data collection of Facebook information articles. They came from three major political news websites (Politico, CNN, and ABC Information) and three massive Facebook pages (one for each political party). A classification accuracy of around 74% was attained. False data has somewhat lower classification accuracy. The dataset's skewness might be to blame for this; just 4.9% of it contains false data. The authors Himank Gupta et al. [10] have up a framework that uses many machine learning techniques to handle hundreds of tweets in 1 second while addressing issues including accuracy lack, time lag (BotMaker), and excessive handling time. They began by extracting 400,000 tweets from the HSpam14 database. They should next specify which of the 150,000 tweets were spam and which of the 250,000 were not. They also obtained certain lightweight functions in addition to the top 30 terms from the Bag-of-terms version that provide the most detail increase. 4. Their accuracy was 91.65%, which was around 18% better than the previous answer. An innovative ML fake news detection approach was first suggested by Marco L. Della Vedova et. al. [11]. This method outperforms current techniques in the literature and increases its accuracy by up to 78.8 percent by integrating news data with social context variables. Additionally, they verified their approach using a real-world application and implemented it into a Facebook Messenger Chabot. The result was an impressive bogus information finding accuracy of 81.7%. Their goal was to determine whether a story was real or not. They started by outlining the datasets they used for testing, then presented the content-based technique they used, and finally suggested a way to combine it with a social-based strategy that was already present in the literature. The final dataset contains 15,500 messages pulled from 32 sites, including 14 pages devoted to conspiracy theories and 18 pages devoted to science. The number of likes on these pages exceeds 2,300,000, all thanks to more than 900,000 individuals. While 6,577 articles (42.4% of the total) are not hoaxes, 8,923 (57.6%) are. By predicting precision assessments in two credibility-focused Twitter datasets—CREDBANK, a group-sourced dataset of accuracy assessments for events in Twitter, and PHEME, a dataset of potential rumours in Twitter and journalistic analyses of their accuracy—Cody Buntain et. al. [12] develop a method for automating fake news detection on Twitter. Twitter posts pulled from BuzzFeed's fake news database are subject to this method. The results of crowd sourced and journalistic precision assessments are based on previous work, and a feature evaluation finds the traits that are most predictive of those outcomes. This approach is limited to the set of favourite tweets since they depend on identifying highly retweeted threads of discourse and using the attributes of these strings to categorise tales. Due of the low retweet rate, this strategy is only applicable to a small subset of Twitter debates. The goal of the study by Shivam B. Parikh et al. [13] is to provide light on how news stories are portrayed in today's diaspora, as well as the various types of news stories and the impact they have on readers. In the end, we discuss common fake news datasets and go into current methods for discovering false news, most of which rely heavily on text-based assessment. Finally, the report concludes with a list of four critical open research issues that might direct future research. This method provides examples of the detection of false news by examining the psychological components; it is an academic technique.

METHODOLOGY

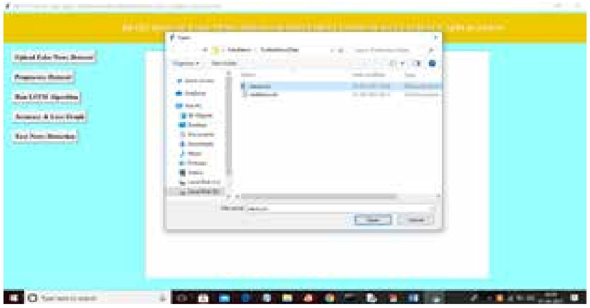
Technical linguistics, the study's methodology, and the findings and performance of the classifier are all detailed in this article. At the conclusion of the article, we go over the steps needed to transform the current system into an impact mining system. False news stories circulate on social media and share common language features including excessive use of unfounded exaggeration and unattributed quotation material. This article presents and discusses the findings of a fake news classifier efficiency study that used fake information recognition.

Objective

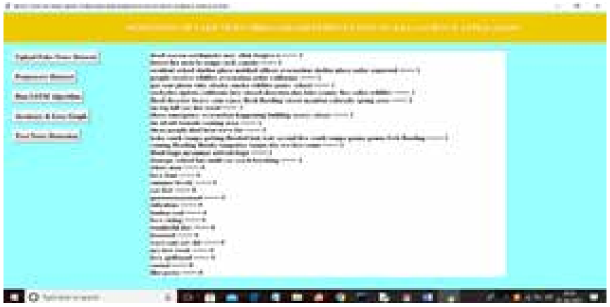
Multiple examples have shown how annoying phoney news can be. Recent research has shown that it may influence national and regional discourse and has a significant effect on public opinion. It has hurt businesses and people, and in the worst case scenario, a private citizen's reaction to a hoax led to their death. Many pupils are unable to distinguish between authentic and fake postings, and it has even caused some teens to reject the idea of media neutrality. Some even go so far as to say it had an impact on the 2016 US presidential election. Crawler militaries provide a wicked write-up enormous reach, but folks may circulate fake material purposefully or indiscriminately. In order to maximise effect, it is common practice to use not just faked articles but also fake, mislabeled, or misleading images. False news, according to some, is a "afflict" on the cultural infrastructure of the internet. A number of people are trying to stop it. Examples of such recommendations include a factor-based method put out by Farajtabar et al. and the usage of "peer-to-peer counter publicity" put forth by Haigh, Haigh, and Kozak.

SYSTEM UNDER PROPOSAL

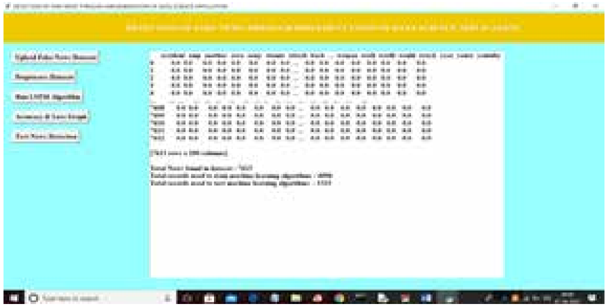
Using Natural Language Processing and an acknowledgment monitoring finding estimator, the author of this study proposes a method to detect fake news in record corpora or on social media. In order to determine a score, verbs, quotes, and name entity (also known as acknowledgment), the application will first upload news documents or posts. Then, using Natural Language Processing, the application will extract quotes, verbs, and name entities (such as companies or individuals' names) from the documents. By dividing the entire number of words in a phrase by the total number of verbs, name entities, and quotations, we can get a score using a supervised comprehension estimator. The information is considered real if the score is more than 0, and new if the score is less than 0.



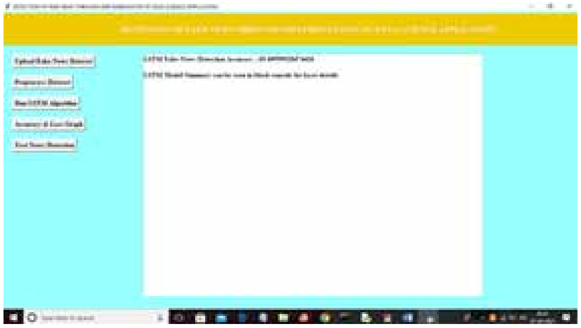
To load the dataset, pick the "information.Csv" report and click on the "Open" button on the previous page. This will convey up the following display.



After loading the dataset, we are able to see all the information textual content with a class label of zero or 1 within the text box. To transform the string records to a numeric vector, click on at the "Preprocess Dataset & Apply NGram" button. This will deliver us to the following web page.

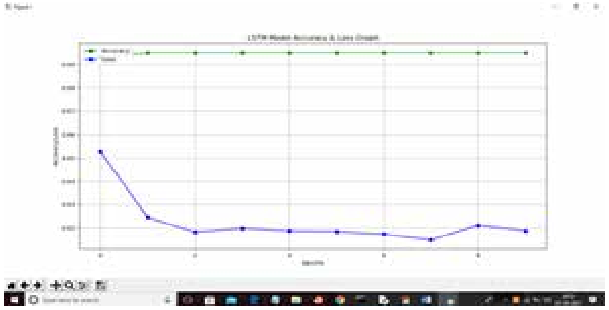


If a news word appears in any row in the above display, its column will be updated to reflect the word matter. Otherwise, if the word does not appear in any row, the column will be set to 0. The dataset is now prepared with numerical records; to educate over the dataset with LSTM, develop an LSTM design, compute precision and error rate, and so on, click the "Run LSTM Algorithm" button. The previous screen displayed a subset of the 7612 records in the dataset, and the bottom lines show that the dataset includes 7613 records in total. The application used 80% of the news documents for training, and 15% for testing.

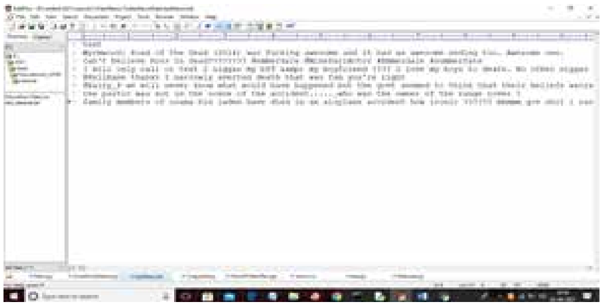


You can see the specifics of the LSTM layer inside the console below, and the created LSTM model with a prediction accuracy of sixty nine. Forty nine% is shown in the screen above.

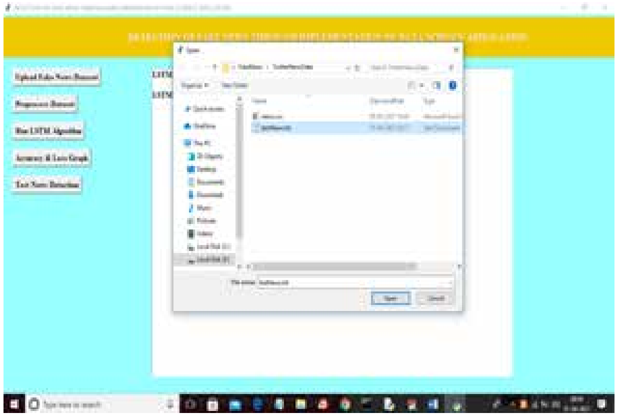
Various LSTM layers are constructed on the above screen to extract useful characteristics for prediction by filtering input data. To get the LSTM graph, choose "Accuracy & Loss Graph" from the current menu.



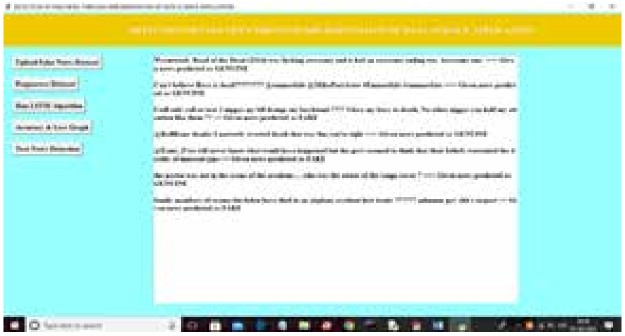
The x-axis within the above graph suggests the range of iterations or epochs, even as the y-axis shows the values of accuracy and loss. The green line shows the accuracy, and the blue line indicates the loss. As the epochs boom, the loss values drop, and the accuracy reaches 70%. To see how the app determines the veracity of test information phrases, click the "Test News Detection" button. The following take a look at news dataset consists of honestly text facts without any class labels; LSTM will then use this records to predict the class labels.



We can see that the take a look at information display only has one column, "TEXT," and that by applying the test news, we might also get the prediction result shown above.



Choose the "testNews.txt" file from the previous page, upload it, and then click the "Open" button to receive the following prediction result.



On the screen above, the dashed symbols come after the news content, and the software then makes a prediction about the veracity of the information based on the text. When the model is complete, the programme will use the LSTM to ascertain whether the bulk of the words in the provided data fall into the authentic or false category. Based on this data, it will choose which class label to use.

CONCLUSION

An incomplete method for detecting false information changed into advanced and discussed in this text. This examine gives something new to the sector by way of showcasing the outcomes of a comprehensive research initiative that commenced with qualitative observations and ended with a purposeful quantitative model. Also, this paper's look at is encouraging as it suggests that with handiest one extraction characteristic, system studying can classify huge amounts of fake news articles as an alternative well. Lastly, a greater state-of-the-art scheme for classifying each false news and direct quotations has to be the result of continuing to take a look at and improvement to find and assemble more grammar for classifying faux information.

Looking Ahead

Also, this paper's study is encouraging as it indicates that with most the effective one extraction feature, device mastering can classify big quantities of false news articles alternatively well. Lastly, a extra state-of-the-art scheme for classifying both fake news and direct quotations needs to be the result of continuing a look at and improvement to uncover and assemble more grammar for classifying fake information.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "DEFENSIVE MODELING OF FAKE NEWS THROUGH ONLINE SOCIAL NETWORKS", which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department of CSE (AI&ML), Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu, "Fake News Detection on Social Media: A Data Mining Perspective" arXiv:1708.01967v3 [cs.SI], 3 Sep 2017.
2. Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu, "Fake News Detection on Social Media: A Data Mining Perspective" arXiv:1708.01967v3 [cs.SI], 3 Sep 2017.
3. M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, 2017, pp. 900-903.
4. Fake news websites. (n.d.) Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Fake_news_website. Accessed Feb. 6, 2017.
5. Cade Metz. (2016, Dec. 16). The bittersweet sweepstakes to build an AI that destroys fake news.
6. Conroy, N., Rubin, V. and Chen, Y. (2015). "Automatic deception detection: Methods for finding fake news" at Proceedings of the Association for Information Science and Technology, 52(1), pp.1-4.
7. Markines, B., Cattuto, C., & Menczer, F. (2009, April). "Social spam detection". In Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web (pp. 41-48).
8. Rada Mihalcea , Carlo Strapparava, The lie detector: explorations in the automatic recognition of deceptive language, Proceedings of the ACL-IJCNLP .
9. Kushal Agarwalla, Shubham Nandan, Varun Anil Nair, D. Deva Hema, "Fake News Detection using Machine Learning and Natural Language Processing," International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-6, March 2019.
10. H. Gupta, M. S. Jamal, S. Madisetty and M. S. Desarkar, "A framework for real-time spam detection in Twitter," 2018 10th International Conference on Communication Systems & Networks (COMSNETS), Bengaluru, 2018, pp. 380-383.
11. M. L. Della Vedova, E. Tacchini, S. Moret, G. Ballarin, M. DiPierro and L. de Alfaro, "Automatic Online Fake News Detection Combining Content and Social Signals," 2018 22nd Conference of Open Innovations Association (FRUCT), Jyvaskyla, 2018, pp. 272- 279.

Agricultural Crop Recommendations based on Productivity and Season

Namana Balabharath, G. Manikanta, M. Kiran Kumar
B.Tech Student
Dept. of Computer Science and Engg. (Data Science)
CMR Technical Campus, Medchal
Hyderabad, Telangana

A. Mahendar
Associate Professor
Dept. of Computer Science and Engg. (Data Science)
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ mahi.adapa@gmail.com

ABSTRACT

Farming is the backbone of India's socioeconomic material resources. As a result of farmers' inability to accurately assess land suitability for crop production using conventional, non-scientific methods, a major problem has emerged in a country where almost 58% of the population works in agriculture. Despite careful consideration of soil type, planting time, and location, farmers did not always choose the best crops. A lot of farmers are committing suicide by giving up on farming and heading to cities in search of better economic opportunities. Our research has provided a method to help farmers choose plants by considering factors including planting time, soil, and location, which might alleviate this problem.

KEYWORDS: *Crop, Agriculture, Farmer, Wrong crop.*

INTRODUCTION

Nearly 58% of our nation's population relies on farming as a source of income [1]. The 2016–17 economic survey found that farmers in 17 states had an average monthly income of only Rs.1700/–, leading to farmer suicides and the conversion of farmland to non-farm uses. Furthermore, almost half of all farmers would prefer that their children and grandchildren live in urban areas rather than continue the family business. The reason being, farmers often choose the incorrect crop for their soil type, plant it in the wrong season, and make other similar mistakes when it comes to choosing plants [2]. The farmer may not have had much experience making this decision, since he or she may have inherited the property. Less yield is an inevitable consequence of choosing the wrong plant. After that, it becomes very difficult for the family to subsist if they rely only on these incomes. Prospective researchers are discouraged from collaborating on conducting national studies due to the lack of readily available and accurate information. We have the means to implement a system that will help us anticipate plant sustainability issues and provide solutions based on AI models that are well-informed on critical financial and environmental factors. The proposed method takes into account both the customer's location relative to the state and environmental factors like rainfall and temperature, as well as soil type, pH value, and vitamin and mineral concentration, in order to recommend the best crop. Furthermore, if the farmer chooses the correct crop, they will also get a return projection. The objective is to create a long-term model that accurately predicts how long plants will survive in a given environment, taking into account the specific soil type and weather conditions [4].

Second, to protect the farmer from financial ruin, suggest the finest local flora.

Third, using data from the previous year, provide a study of the income of several plants [5].

The suggested system is powered by AI, which is one of the many uses of AI. AI enables systems to learn and adapt instantly, even when not explicitly specified by a developer. After then, the program's accuracy will be improved without any human intervention [6]. Many researchers are devoting their time and energy to this area in the hopes of providing farmers with the information they need to choose the best plant for their needs by taking into account a variety of aspects, including physical, ecological, and economical ones. It is believed that a Synthetic Neural Network can choose the plant with the highest production rate [1]. Prior to culture, the plants were evaluated using algorithms such as K Nearest Neighbours Regression and Choice Tree Learning-ID3 (Iterative Dichotomiser 3). [9] The arbitrary woodland formula and BigML were used for the analysis of crop characteristics. [10] In order to protect plants from the effects of water stress, AI algorithms were used, which led to the development of a set of decision criteria used for plant state prediction. Predicting plant costs using machine learning approaches and providing real-time suggestions via smart systems were also used. This position has included researching various AI algorithms applications in agricultural production systems. Suggestions on plant management were provided by further AI-enabled technologies. Improved agricultural yields are possible with the use of deep-understanding approaches. A dependable return projection technique is explored in this research using real-time monthly weather. In order to implement the aforementioned method of projection, a non-parametric analytical design and non-parametric regression methods were used.

LITERATURE SURVEY

1) Choosing the Right Plants A technique that makes the most of the crop return price by using an AI approach The economic development and food security of agro-based nations are greatly impacted by agricultural planning. A major consideration in agricultural planning is the selection of plants. Factors like manufacturing cost, market value, and official policies all have a role. In order to get the agricultural sector ready, several researchers have looked at data approaches and artificial intelligence techniques for predicting crop output prices, weather conditions, soil types, and plant classifications. If limited land resources allow for more than one plant to be planted simultaneously, then the choice of crop becomes more complex. In order to solve the problem of plant selection, maximise the use of the internet yield price of crops throughout the year, and achieve optimum national economic growth, this study proposed a method called the Plant Option Approach (CSM). Crops' web yield rates may be increased by the suggested method.

2) Chemical Prediction and Efficient Plant Return for Agricultural Economic Climate Improvement with the Use of Exploring Data There is risk in the agricultural sector. Climate, geography, biology, politics, and finances are all factors that affect crop yields. It is possible to quantify the risks posed by these factors by using appropriate mathematical or statistical methods. In reality, precise information on the characteristics of crop yield history is crucial modelling input that aids farmers and government organisations in making decisions regarding the establishment of suitable policies. Technological developments in computing and data storage have really made available enormous amounts of data. The challenge has been in de-expertizing this raw data, which has prompted the development of new techniques like data mining that can connect the dots between data comprehension and the plant return estimate. In order to determine whether meaningful connections may be found, this study set out to assess these novel information mining techniques by applying them to the many variables that make up the data source.

Dirt Characteristic Forecasting and Category Methods for Analysing Dirt Data Technological developments like information mining and automation have benefited agricultural research. There is a plethora of domain-specific information mining software and off-the-shelf data mining systems available today, and data mining is in widespread usage; yet, the application of data mining to agricultural soil datasets is still underdeveloped. The massive amounts of data that are now often gathered alongside crops need to be assessed and used to their maximum potential. The goal of this study is to apply data mining techniques to assess a soil dataset. Soil categorization using multiple publicly accessible methods is its main focus. Another important goal is to use the regression approach and the use of the automated soil sample categorization to predict untested qualities.

Fourthly, Intelligent Farming using Machine Learning In India's economic landscape, farming is a key player. However, a structural shift is causing a crisis in India's agricultural industry right now. The only way out of this jam is to bring in more farmers and make farming a lucrative business so they can keep making crops. This term paper is an effort in that direction; it will use AI to assist farmers make better decisions about their farms. utilising supervised maker figuring out formulae, this research focuses on crop prediction utilising weather scenarios and plant yield from historical data. Along with that, an online app has been developed.

EXISTING SYSTEM

Approximately 58% of our nation's population relies on agriculture as a primary source of income. The average monthly income of a farmer in 17 states is Rs.1700/-, according to the 2016-17 economic survey. This leads to farmer suicides and the conversion of agricultural land for non-agricultural use. Furthermore, almost half of all farmers (48%) want to leave farming to their children and grandchildren and instead live in urban areas. Why is this? Well, it's because farmers often choose the incorrect plants for the soil, don't water them enough, let them grow for too long, etc. The farmer may not have had much experience making the decision since the land might have been purchased from someone else. Choosing the wrong crop will always result in lower yields. It will be very difficult to get by if everyone in the family relies on this income. A forest-based random formula was stored in the previous system. yet the exact recommended harvest is impossible to predict.

PROPOSED SYSTEM

With the use of machine learning, the proposed system aims to improve the accuracy of yield price by advising farmers on the best crops to plant depending on factors such as soil type, sowing season, climate, physical, ecological, and economic factors, and perennially popular plants. This new development allows the systems to learn and adapt on the fly, even when the programmer hasn't explicitly set them up. Eliminating human intervention will improve the program's accuracy. In order to help farmers pick the best crops by taking into account all of the relevant aspects, including physical, environmental, and economical ones, several scientists are doing research into plant selection guidelines. system. The aforementioned projection tool was being used using a non-parametric analytical version in conjunction with nonparametric regression methodologies. This task involves directly feeding the system a plethora of datasets obtained from official government websites and Kaggle. Various versions of the equipment are trained using the dataset that was retained after the pre-processing stage to get the highest potential accuracy.

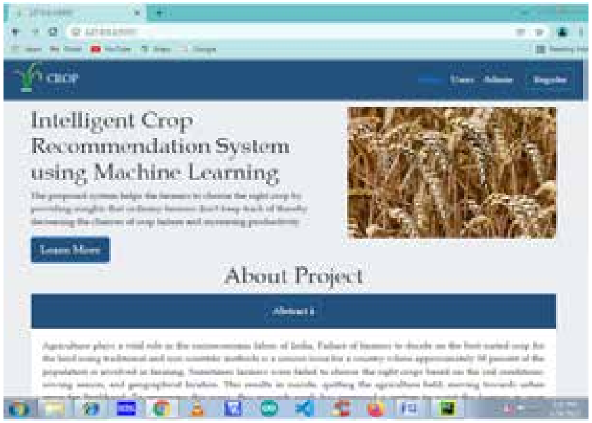
MODULES DESCRIPTION

Person: A person may register their initials. For future communications, he needed a valid user email and cellphone upon registration. The administrator may activate the user the moment a client registers. The client will be able to access our system after the administrator has triggered the person. Our dataset columns may be used as a basis for customers to provide their datasets. Data must be expressed as integers or floats in order for algorithms to run. We have a ph here. For the purpose of screening, there is a repository dataset of weather issues. Using our Django application, users may also contribute new data to an existing dataset. Users may initiate the data cleansing process by clicking the "Data Preparations" button on the websites. It is guaranteed that the cleaned-up data and its necessary graphics will be shown.

If the administrator wants to log on, he may use his credentials. Customers who have signed up may be enabled by the admin. Once he activates, our system will only allow the user to log in. In the browser, the administrator may see the aggregate statistics. The ROC contour, confusion matrix, and accuracy of the algorithms are all under his purview as well. The following is also a bar graph showing the comparative accuracy. The administrator will be able to see the websites' overall accuracy after all algorithm execution is complete.

Preprocessing of Data: Information objects (also referred to as records, factors, vectors, patterns, occurrences, instances, samples, monitoring, or entities) make up what is known as a dataset. Data objects are described by a number of characteristics that record the commonalities between things, such the mass of a physical item or the time when an event occurred, among others. Variables, attributes, regions, features, and measures are common names for functions. This forecast's data pre treatment takes use of methods including collecting characteristics for prediction at several levels, removing missing details, adjusting default values as needed, and eliminating sound from the data.

The dataset is subjected to five artificial intelligence classifiers, including Logistic Regression (LR) with pipe, Assistance Vector Maker (SVM), Decision Tree (DT), and Random Woodland (RF), after which the cleaned information is divided into 60% training and 40% test, according to the split requirement. In order to find out how accurate the classifiers were, the complexity matrix was used. The best classifier might be the one that achieves the most precision.



Intelligent Crop Recommendation System using Machine Learning

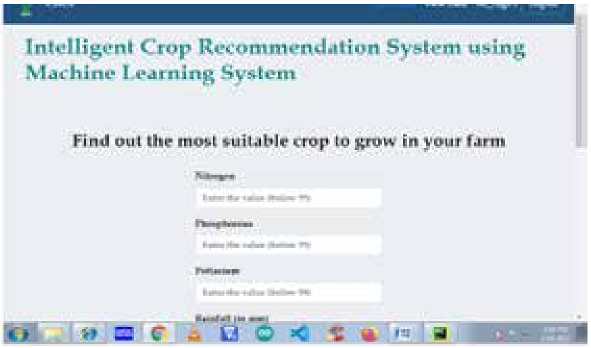
The proposed system helps the farmers to choose the right crop by providing insights that ordinary farmers don't keep track of thereby decreasing the chances of crop failure and increasing productivity

About Project

Abstract ↓

Agriculture plays a vital role in the socioeconomic fabric of India. Failure of farmers to decide on the best-suited crop for the land using traditional and non-scientific methods is a serious issue for a country where approximately 58 percent of the population is involved in farming. Sometimes farmers were failed to choose the right crops based on the soil conditions, sowing season, and geographical location. This results in suicide, quitting the agriculture field, moving towards urban [illegible]

Fig.1. Home page



Intelligent Crop Recommendation System using Machine Learning System

Find out the most suitable crop to grow in your farm

Nitrogen

Enter the value (Below 99)

Phosphorous

Enter the value (Below 99)

Potassium

Enter the value (Below 99)

Rainfall (in mm)

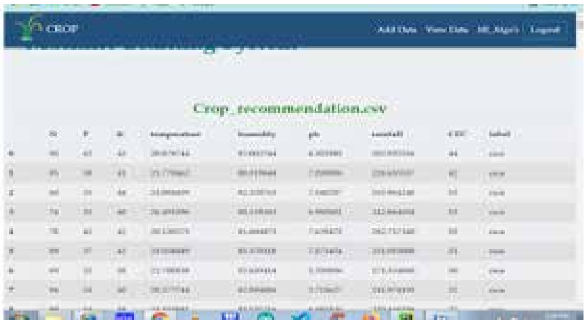
Fig.2. Crop grow materials

Intelligent Crop Recommendation System using Machine Learning System

Datasets Here are:

1. Crop_Recommendation.csv
2. Crop_Sustainability.csv

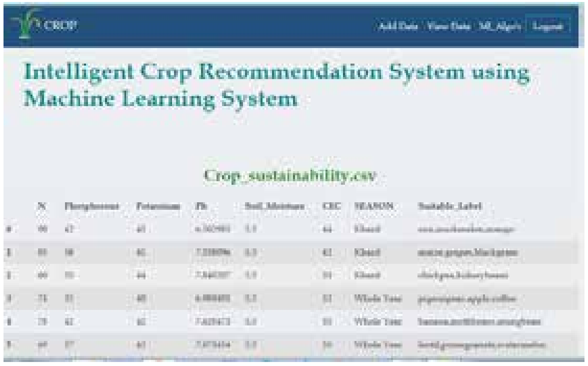
Fig.3. Dataset



Crop_recommendation.csv

	N	P	K	temperature	humidity	ph	rainfall	CEC	label
0	90	42	43	20.879744	82.002744	6.502985	202.935536	44	rice
1	85	58	41	21.770462	80.319644	7.038096	226.655537	42	rice
2	60	55	44	23.004459	82.320763	7.840207	263.964248	53	rice
3	74	35	40	26.491096	80.158363	6.980401	242.864034	53	rice
4	78	42	42	20.130175	81.604873	7.628473	262.717340	55	rice
5	69	37	42	23.058049	83.370118	7.073454	251.055000	51	rice
6	69	55	38	22.708838	82.639414	5.700806	271.324860	50	rice
7	94	53	40	20.277744	82.894086	5.718627	241.974195	55	rice

Fig.4. Crop recommendation

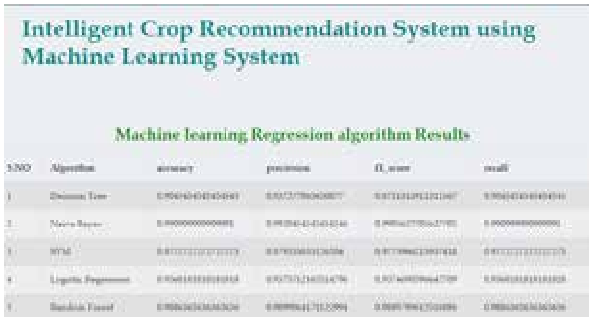


Intelligent Crop Recommendation System using Machine Learning System

Crop_sustainability.csv

	N	Phosphorous	Potassium	Ph	Soil_Moisture	CEC	SEASON	Suitable_Label
0	90	42	43	6.502985	5.5	44	Kharif	rice,muskmelon,orange
1	85	58	41	7.038096	5.5	42	Kharif	maize,pigeon,blackgram
2	60	55	44	7.840207	5.5	53	Kharif	chickpea,kidneybeans
3	74	35	40	6.980401	5.5	53	Whole Year	pigeonpeas,apple,coffee
4	78	42	42	7.628473	5.5	55	Whole Year	banana,mothbeans,mungbean
5	69	37	42	7.073454	5.5	50	Whole Year	lentil,pomegranate,watermelon

Fig.5. Crop sustainability



Intelligent Crop Recommendation System using Machine Learning System

Machine learning Regression algorithm Results

S.NO	Algorithm	accuracy	precision	f1_score	recall
1	Decision Tree	0.9045454545454545	[illegible]	[illegible]	0.9045454545454545
2	Naive Bayes	0.9900000000000001	[illegible]	[illegible]	0.9900000000000001
3	SVM	0.9772727272727273	[illegible]	[illegible]	0.9772727272727273
4	Logistic Regression	0.9545454545454546	[illegible]	[illegible]	0.9545454545454546
5	Random Forest	0.9886363636363636	[illegible]	[illegible]	0.9886363636363636

Fig.6. ML based algorithm results

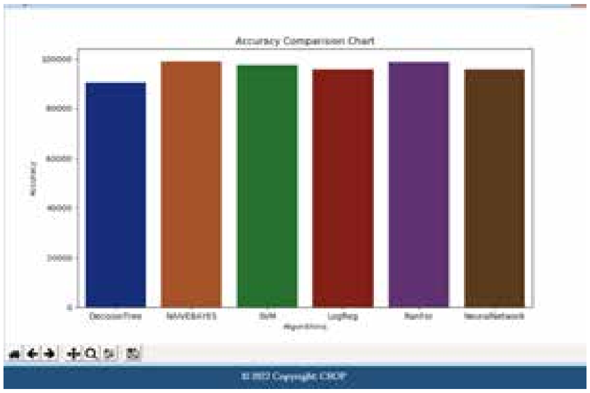


Fig.7. Output results

CONCLUSION

In order to reduce the likelihood of plant failure and increase efficiency, the proposed method aids farmers in selecting the optimum plant by providing insights that typical farmers overlook. Furthermore, it prevents them from incurring losses. It is the intention of the developers to eventually integrate a web interface with a mobile app so that millions of farmers throughout the country may get plant growing recommendations.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "AGRICULTURAL CROP RECOMMENDATIONS BASED ON PRODUCTIVITY AND SEASON", which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department Of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. R. Kumar, M. P. Singh, P. Kumar, and J. P. Singh, "Crop Select ion Method to maximize crop yield rate using machine learning technique", 2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy, and Materials (ICSTM), Chennai, 2015, pp. 138-145, DOI: 10.1109/ICSTM.2015.7225403.
2. H. Lee and A. Moon, "Development of yield prediction system based on real-time agricultural meteorological information", 16t h International Conference on Advanced Communication Technology, Pyeongchang, 2014, pp. 1292- 1295, DOI: 10.1109/ICACT.2014.6779168.
3. T.R. Lekhaa, "Efficient Crop Yield and Pesticide Prediction for Improving Agricultural Economy using Data Mining Techniques", International Journal of Modern Trends in Engineering and Science (IJMTES), Volume 03, Issue 10, 2016.
4. Jay Gholap, Anurag Ingole, JayeshGohil, ShaileshGargade and Vahida At t ar, " SoilData Analysis Using ClassificationTechniques and Soil Attribute Prediction", International Journal of ComputerScience Issues, Volume 9, Issue 3,2012.
5. S. R. Rajeswari, ParthKhunteta, Subham Kumar, Amrit Raj Singh, Vaibhav Pandey, "Smart Farming Prediction using 76 Machine Learning", International Journal of Innovative Technology and Exploring Engineering, 2019, Volume-08, Issue07.
6. Z. Doshi, S. Nadkarni, R. Agrawal, and N. Shah, "AgroConsultant : Intelligent Crop Recommendation System Using Machine Learning Algorithms", Fourth International Conference on Computing Control and Automation (ICCUBEA), Pune,India, 2018, pp. 1-6, DOI: 10.1109/ICCUBEA.2018.8697349.
7. S. Pudumalar, E. Ramanujam, R. H. Rajashree, C. Kavya, T. Kiruthika and J. Nisha, "Crop recommendation system for precision agriculture," Eighth International Conference on Advanced Computing (ICoAC), 2017, pp. 32-36, DOI: 10.1109/ICoAC.2017.7951740.
8. Konstantinos G. Liakos, PatriziaBusato, DimitriosMoshou, Simon P earson and Dionysus Bocht is, "Machine Learning in Agriculture: A Review", Article on sensors, 2018, pp 1 -29, doi:10.3390/s18082674.
9. M. T. Shakoor, K. Rahman, S. N. Rayta and A. Chakrabarty, "Agricultural production output prediction using Supervised Machine Learning techniques", 1st International Conference on Next Generation Computing ions (NextComp), Mauritius, 2017, pp. 182-187, DOI: 10.1109/NEXTCOMP.2017.8016196.

A Spam Transformer Model for SMS Spam Detection

Jamuna Singh B, Goutham Saketh Cheeti, Dasari Sai Teja
B.Tech Student
Dept. of Computer Science and Engg. (Data Science)
CMR Technical Campus, Medchal
Hyderabad, Telangana

K Murali Kanthi
Associate Professor
Dept. of Computer Science and Engg. (Data Science)
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ murali.kanthi @gmail.com

ABSTRACT

Protocol for Short Message Service One kind of smart phone attack that might affect the security and privacy of mobile individuals is spam. The reason for this is because these types of attacks use social engineering techniques to deceive customers into giving up personal information. This study proposed a Malay-specific SMS spam detection framework based on Naïve Bayes. There are a lot of text spam detection systems out there, but artificial intelligence is among the best. Furthermore, the current discovery architecture that employs a machine learning method fails miserably when it comes to SMS sent in the Malay language. This is due to the fact that Spam detection characteristics are not based on the Malay language. Data gathering, pre-processing, classification, discovery, and three types of attribute selection are all part of this system. The outcome demonstrates that the categorization achieves a satisfactory level of accuracy, above 90%.

KEYWORDS: *SMS, Naive bayes, Classification, Data collection.*

INTRODUCTION

When it comes to mobile phone contact tools, one alternative is Short Message Solutions (SMS). However, scammers may use SMS to trick people. [1] While there is anti-Spam software that can be downloaded and installed on mobile devices, it does not yet have the capability to detect spam written in Malay. The official and informal communication in Malaysia is conducted in Malay, the country's primary language. Sending out SMS messages to subscribers at random is how text spam operates. Unwanted content, such a link to an online store or advertisement, is included in the message [2]. In most cases, the recipient is charged for each unwanted SMS and must respond in order to stop receiving them. The consumer is even responsible for paying to cancel the SMS. It seems that the customer's smartphone's security and privacy have been compromised. Content-Based Filtering, Whitelists or Blacklists, Machine Learning, Matching Patterns, and Artificial Immune Systems are some of the detection strategies that have been implemented in SMS spam discovery research [3]. This research presents a Malay language text spam detection system that includes the following steps: data collection, pre-processing, feature selection based on Malay language, classification, and detection. This article presents the results of many experiments that were conducted to evaluate the method and framework that were suggested. Additionally, Ignorant Bayes classification approach was used to validate the findings in each framework step. Since most existing research focuses on English SMS Spam functions and very little on Malay language SMS functions, this framework would undoubtedly aid in providing a Malay language SMS Spam feature for future work in spam detection [4].

The Short Message Service (SMS) allows for the transmission of text messages between mobile phones. Spam refers to unsolicited communications or scrap of text [1]. In addition to making and receiving audio calls, people nowadays use their mobile phones for a plethora of other financial tasks, such as checking balances, sending and receiving emails, accessing Facebook, and online shopping. All of these activities necessitate the use of sensitive information, such as passwords, PINs, savings account numbers, bank or debit card numbers, etc. Not only that, but a lot of individuals also save sensitive information, including the numbers of friends and family members, as well as images of themselves and their identification on their phones. Cybercriminals may get access to people's personal information by sending them spam SMS. The intrusion of spam SMS may annoy and irritate mobile users [2]. Lost productivity and wasted network data transmission are the results of spam SMS [3]. The Indian government instituted the National Client Preference Windows Registry (NCPR) to cut down on unwanted calls, however it doesn't screen spam text messages [1]. We aim to classify sms messages as either spam or legitimate using text classification algorithms, which are extensively utilised in spam filtering systems [4]. Each document is assigned a group from a set of preset categories [5]. One use of supervised learning is text categorization. So, gathering labelled data is necessary for building a classifier in message classification at the moment. According to our data, legitimate mails were labelled as pork. There are many similarities between email spam filtering and SMS spam filtering, which means that many of the problems with SMS spam detection are really improvements over email spam detection [6]. [7] For the purpose of screening spam emails, the Naïve Bayes classifier is the method of choice. [8] For the purpose of detecting spam SMS, this study uses the Naïve Bayes formula.

AN OVERVIEW OF PROPOSED SYSTEM

Analysing the Effectiveness of Different AI Techniques for E-Mail Spam Classification

Venkata Sri Vinitha and D. Karthika Renuka are the authors. In the year 2020, because it is cheap and easy, a lot of people use email for business and basic communication. This efficacy exposes exposed email to various spamming attacks. These days, e-mail users are mostly worried about spam. The purpose of these spams is to cause unnecessary network bandwidth usage by delivering malicious links to hazardous websites or malicious components such as executable files to attack client computers. In order to filter out spam emails, this paper explains various AI strategies like the J48 classifier, Adaboost, K-Nearest Next-door Neighbour, Ignorant Bayes, Artificial Semantic Network, Support Vector Equipment, and Random Woodlands algorithm. It does this by utilising various email datasets. There is a summary of the current state of affairs about the accuracy price of several available tactics, and a comparison of various spam email categorization methods is also provided.

Using Machine Learning Algorithms for the Determination of Email Spam

Nikhil Kumar, Sanket Sonowal, and Nishant In the year 2020, The rapid growth of internet users has coincided with an increase in email spam, which has become a serious problem in recent times. Phishing and fraud are among the unethical and illegal uses of these. Distributing malicious links in spam emails, which may compromise both our and your systems, is unacceptable. Spammers prey on unsuspecting victims since it is simple for them to set up a false identity and email account; in their spam emails, they pretend to be someone they are not. Since it is necessary to identify fraudulent spam emails, this project will use AI techniques to do just that. In this paper, we will discuss machine learning formulas and then apply all of these algorithms to our data sets until we find the one with the highest accuracy and precision for e-mail spam detection.

Extracting Attributes and Classifying Spam Emails

Authors: Nhamo Mtetwa and Muhammad Ali Hassan [2018] When compared to other forms of written communication, emails are much more common and favoured. Spam is an issue with email. Various goals motivate the sending of these spam emails, but the most common ones are advertising and fraud. It poses a lot of issues for online culture, yet it's cheap to distribute. In order to create a spam filtering system, this study explores the usage of several feature removal techniques in conjunction with two supervised machine learning classifiers. The classifiers are tested using four performance metrics on two publicly accessible spam email datasets. Using two separate datasets has obvious advantages, and we stress the need of properly connecting function removal and classifier.

An AI-Based Approach to Email Spam Classification: A Proposed Data Science Research Method

The authors of this work are Aakash Atul Alurkar, Sourabh Bharat Ranade, Shreeya Vijay Joshi, and Siddhesh Sanjay Ranade from 2017.

Since the majority of people now have access to the internet, one of the biggest problems plaguing our internationally connected communication networks is the proliferation of spam emails. In the past, spam filtering and hiding services relied on manually searching for certain keyword phrases and blacklisting sites that were known to transmit spam. When it comes to classifying emails as spam or junk, however, these methods have some distinct limitations. The proposed method employs machine learning techniques in an effort to detect spammy patterns of repeated keyword phrases. Other framework elements, such as the domain name, header, and Cc/Bcc, are also used by the system to suggest the email category. When plugging each parameter into the machine learning formula, it would be treated as a function. An already-trained version with a feedback mechanism to differentiate between correct and incorrect results will undoubtedly make up the machine learning architecture. This method offers an alternative technique for implementing a spam filter.

Research on AI Classifiers for Identifying Spam

Shrawan Kumar Trivedi wrote this.

Email engagement is in demand in today's society, yet unsolicited emails make it difficult to have productive conversations. The current research focuses on creating a spam category design that incorporates approaches employing ensembles of classifiers and methods without them. The goal of this project is to develop a sensitive and dependable classification model that can distinguish between ham and spam emails, and then use that model to lower the false positive rate while maintaining excellent accuracy. Insatiably wealthy In order to find relevant features in the Enron email dataset, a step-by-step attribute search method has been developed. A variety of maker finding classifiers, including Bayesian, Naïve Bayes, SVM (assistance vector maker), J48 (decision tree), Bayesian with Adaboost, and Naïve Bayes with Adaboost, have been compared. The relevant classifiers are statistically evaluated and assessed using metrics like training duration, False Favourable Rate, and F-measure (precision). It has been shown that SVM is the best classifier to utilise after examining all of these factors together. Its accuracy is good and its misleading favourable price is minimal. Although architecture takes a long time to train SVMs, this is not a major issue because of the results on other parameters.

PROPOSED METHODOLOGY

We separated the dataset into a training set and an examination set after we cleaned it up. In order to train the Naïve Bayes classifier, educational data was gathered. The examination dataset was used to evaluate the efficiency of the competent classifier. A. Summary of the Dataset

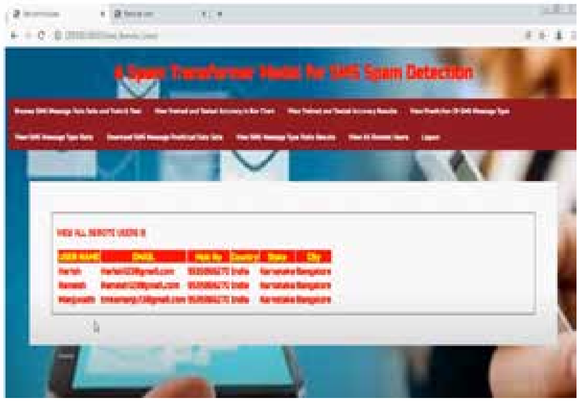
The verbal Spam Collection v. 1 dataset was used [9]. It was from [10] that we obtained this dataset. There are 5572 SMS messages in this collection, all of which were marked as spam or junk. Two columns, labelled v1 and v2, make up the data set. There are merely two values in the first column, v1, that indicate whether the content in the second column, v2, is spam or legitimate. You may access the dataset in a comma-separated values (CSV) format. The following sources provided the messages used to compile this dataset: Caroline Tag's PhD thesis, the Grumbletext website, and the NUS text corpus (NSC) Large Text Spam Corpus v. 0.1 version. Within this sample, 4825 pieces of text were classified as spam and 747 as ham.

Fig.1. Home page

B. Preprocessing the Data

Column v1 is now called course while column v2 is called text. We mixed the dataset after column relabeling in order to reduce over fitting. The dataset was cleaned after being shuffled. The dataset was cleaned by changing all the content to lowercase and removing any instances of stop words, numerals, spelling mistakes, or URLs.

Fig.2. Registration page



A Spam Transformer Model for SMS Spam Detection
Browse SMS Message Data Sets and Train & Test
View Trained and Tested Accuracy in Bar Chart
View Trained and Tested Accuracy Results
View Prediction Of SMS Message Type
View SMS Message Type Ratio
Download SMS Message Predicted Data Sets
View SMS Message Type Ratio Results
View All Remote Users
Logout

VIEW ALL REMOTE USERS !!!

USER NAME	EMAIL	Mob No	Country	State	City
Harish	harish123@gmail.com	[illegible]	India	Karnataka	Bangalore
Ramesh	ramesh123@gmail.com	[illegible]	India	Karnataka	Bangalore
Manjunath	tmksmanju13@gmail.com	[illegible]	India	Karnataka	Bangalore

Fig.3. User details

Bayes Classifier with Ignorance

There was an immediate separation of the dataset into a training set and an assessment set after data preparation. Out of 5572 texts in the sample, 747 have been classified as spam and 4825 as ham. Two datasets were created from the data. The educational dataset included 4,00 SMS messages, 3,461 of which were deemed spam and 539 as pork. The remaining 1572 SMS messages were categorised as either spam or ham in the test dataset. First, a model or classifier is constructed for the category; next, it is used to anticipate the course identifiers. the eleventh At first, we extracted words with frequencies below 5 from the training dataset and transformed them into a file term matrix. "No" was substituted for "Yes" at the paper term matrix's entrance 0 and for many other non-zero accesses. Only the values "Yes" and "No" were present in this file word matrix. Using the course tags of text from the training dataset and this paper word matrix, the Naïve Bayes classifier was trained. Similarly, a file word matrix was also generated for the text messages in the test dataset and used by the Naïve Bayes classifier to forecast the course labels of the SMS messages.

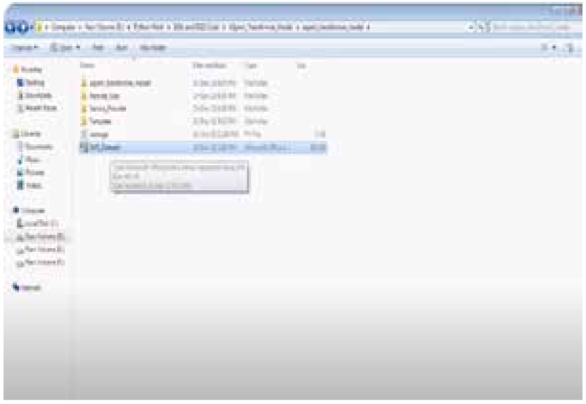
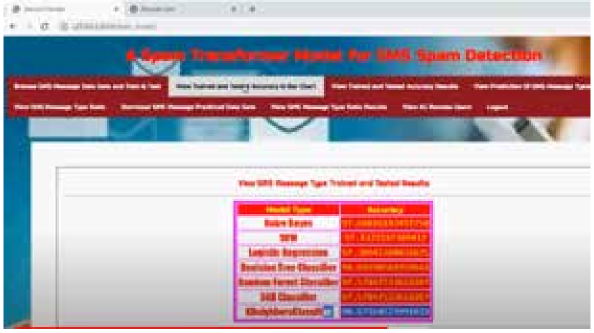


Fig.4. Upload data set



A Spam Transformer Model for SMS Spam Detection
Browse SMS Message Data Sets and Train & Test
View Trained and Tested Accuracy in Bar Chart
View Trained and Tested Accuracy Results
View Prediction Of SMS Message Type
View SMS Message Type Ratio
Download SMS Message Predicted Data Sets
View SMS Message Type Ratio Results
View All Remote Users
Logout

View SMS Message Type Trained and Tested Results

Model Type	Accuracy
Naive Bayes	[illegible]
SVM	[illegible]
Logistic Regression	[illegible]
Decision Tree Classifier	[illegible]
Random Forest Classifier	[illegible]
SGD Classifier	[illegible]
KNeighborsClassifier	[illegible]

Fig.5. Accuracy details

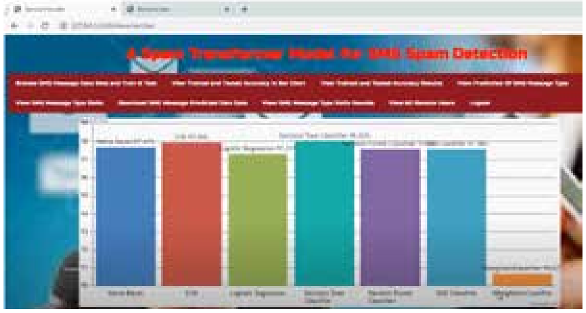


Fig.6. Output results

CONCLUSION

The current framework for detecting SMS attacks can only identify attacks that target certain attributes. Finding Malay text spam has led to the presentation of spam characteristics in Malay, which aids in the detection of SMS spam in Malaysia. The suggested structure may be used with five (5) message mining algorithms to identify these attacks. Using Ignorant Bayes, the data mining device testing demonstrated satisfactory results. The method's accessibility in current AI devices and its widespread use by other scientists in text spam strike identification formed the foundation for its selection. In conclusion, existing research shows that it is important to use AI strategies to find Malay SMS spam, as most studies focus on English and this creates a barrier to finding Malay SMS spam. The confidentiality and privacy of mobile device users are being undermined by the growing amount of SMS spam on mobile phones. While there are devices that can detect and filter out SMS spam, they don't work very well when it comes to Malay language text spam. There has to be greater functionality to identify Malay language SMS spam attacks.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "A SPAM TRANSFORMER MODEL FOR SMS SPAM DETECTION", which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department Of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. S. Chhabra, "Fighting spam, phishing and email fraud," UNIVERSITY OF CALIFORNIA RIVERSIDE, 2005.
2. K. Yadav, P. Kumaraguru, A. Goyal, A. Gupta, and V. Naik, "SMSAssassin: crowdsourcing driven mobile-based system for SMS spam filtering," Proceedings of the 12th Workshop on Mobile Computing Systems and Applications. ACM, Phoenix, Arizona, pp. 1– 6, 2011.
3. H. Shirani-Mehr, "SMS Spam Detection using Machine Learning Approach," 2014.
4. Cloudmark, "Annual Security Threat Report 2014," Cloudmark, San Francisco, USA, 2014.
5. E. Vall, #233, and P. Rosso, "Detection of near-duplicate user generated contents: the SMS spam collection," Proceedings of the 3rd international workshop on Search and mining user-generated contents. ACM, Glasgow, Scotland, UK, pp. 27–34, 2011.
6. H. Najadat, N. Abdulla, R. Abooraig, and S. Nawasrah, "Mobile SMS Spam Filtering based on Mixing Classifiers," 2014.
7. K. Yadav, S. K. Saha, P. Kumaraguru, and R. Kumra, "Take Control of Your SMSes: Designing an Usable Spam SMS Filtering System," in Mobile Data Management (MDM), 2012 IEEE 13th International Conference on, 2012, pp. 352–355.
8. T. M. M. and A. M. Mahfouz, "SMS Spam Filtering Technique Based on Artificial Immune System," IJCSI Int. J. Comput. Sci. Issues, vol. 9, no. 2, 2012.
9. Q. Xu, E. Xiang, J. Du, J. Zhong, and Q. Yang, "SMS Spam Detection using Content-less Features," Intell. Syst. IEEE, vol. PP, no. 99, p. 1, 2012.
10. M. Taufiq Nuruzzaman, C. Lee, M. F. A. bin Abdullah, and D. Choi, "Simple SMS spam filtering on independent mobile phone," Secur. Commun. Networks, vol. 5, no. 10, pp. 1209–1220, 2012.
11. M. Z. R. que and M. Farooq, "SMS Spam Detection By Operating On Byte-Level Distributions Using Hidden Markov Models (HMMS)," Virus Bulletin Conference September 2010. 2010.
12. S. J. Delany, M. Buckley, and D. Greene, "SMS spam filtering: Methods and data," Expert Syst. Appl., vol. 39, no. 10, pp. 9899–9908, 2012.
13. Tiago A. Almeida and J. M. G. Hidalgo, "SMS Spam Collection Data Set," 2012. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection.
14. I. Androulidakis, V. Vlachos, and A. Papanikolaou, "FIMESS: filtering mobile external SMS spam," in BCI, 2013, pp. 221–227.
15. T. Charninda, T. T. Dayaratne, H. K. N. Amarasinghe, and J. Jayakody, "Content based hybrid sms spam filtering system," 2014.

Toward Detection and Attribution of Cyber-Attacks in IoT-Enabled Cyber-Physical Systems

**Karishma Singh, Peteri Shashank,
Thatipamula Mukesh**
B.Tech Student
Dept. of Computer Science and Engg. (Data Science)
CMR Technical Campus, Medchal
Hyderabad, Telangana

B Shankar Nayak
Professor
Dept. of Computer Science and Engg. (Data Science)
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ bsnayak546@gmail.com

ABSTRACT

Protecting cyber-physical systems (CPS) that are enabled by the Internet of Things (IoT) may be challenging since security solutions designed for information technology (IT) and operational technology (OT) systems may not work as well in a CPS environment. Consequently, this article lays forth a two-tiered attack detection and acknowledgment framework developed for CPS, and more specifically for use in an ICS. In order to detect attacks in unbalanced ICS settings, a decision tree is paired with a one-of-a-kind ensemble deep representation-learning model at the initial level. An attack recognition set deep semantic network is created at the second degree. The proposed layout is tested using data from actual state pipes and water treatment systems. Compared to other competing techniques with comparable computing complexity, the suggested model performs better in search results.

KEYWORDS: *IOT, CPS, IT, OT, ICS, ML, DNN.*

INTRODUCTION

The term "framework" refers to any physical asset that can be used to support or create solutions for a society or a company. This includes things like roads, bridges, airports, subways, parks, public buildings, energy plants, hospitals, public spaces, and community centres. In contrast, essential frameworks include digital and physical centres and services that provide the groundwork for a nation's security, stable economy, and people's health and safety [1]. Its responsibilities include supplying basic services, including water and food as well as energy and gas, as well as healthcare, banking, and transportation [2]. Industrial Control Systems (ICS) are an integral part of any critical infrastructure, including SCADA and other control systems that monitor and regulate data acquisition and processing as well as other control functions [3]. ICS may control the flow of power from the grid to a house or the supply of natural gas to a power plant. Cyber systems are the backbone of vital sectors since almost every single one of them uses IT to support fundamental company operations [4]. Because of their great value, critical infrastructure cyber systems have become targets for attack, and disruptions to these systems have had significant monetary, political, and societal consequences. Software is an integral part of cyber systems; its development encompasses a wide range of tasks, including the execution of new functionalities, analysis of needs, and bug fixes (Sebastian and Stephan, 2018) [5]. A number of important advancements in maritime have occurred in the realm of universal satellite and data connectivity. Safe navigation, communication, emergency response, and traffic management are just a few of the many vital services that rely on the Global Navigation Satellite System (GNSS) [6]. However, ships may be diverted off their intended route and cause accidents, groundings, and environmental catastrophes if GPS signals are altered or tampered with (Dennis et al., 2017) [7].

Critical infrastructure components like dams and power plants are part of cyber-physical systems (CPS), which heavily include Internet of Things (IoT) devices. Internet of Things (IoT) devices, sometimes called Industrial Internet of Things (IIoT), are often integrated into an Industrial Control System (ICS) in such environments. The ICS is responsible for the reliable functioning of the facilities. System control and data acquisition (SCADA) systems, distributed control systems (DCS), and systems using programmable logic controllers (PLCs) and Modbus protocols are all examples of what is often known as industrial control systems (ICS) [7], [8]. However, the attack surfaces and threats of cyber lawbreakers targeting ICS or IIoT-based systems are increased when they are linked to public networks. The 2010 Stuxnet attack, which reportedly targeted Iranian centrifuges used for nuclear enrichment and severely damaged them, is one such example [1, 2]. The 2011 failure of a water facility in Illinois due to an incident involving a pump is another example [3]. In 2015, a further effort called BlackEnergy3 struck the electrical systems of Ukraine, causing some 230,000 people to lose electricity [4]. Also in April 2018, three US gas pipeline firms were the targets of successful cyber-attacks that knocked their electronic customer contact systems down for days [1]. Protection solutions developed for IT and OT systems may not be directly applicable to ICSs, despite the fact that they are quite mature. Because of the close integration of cyber systems with the regulated physical environment, this may be the case, for instance. For this reason, it is crucial to assess physical behaviour and maintain a system operating accessible via system-level safety and security measures. [1] When it comes to protecting ICS systems, transparency, honesty, and discretion take precedence above privacy, stability, and accessibility, which is the usual order for most IT/OT systems. [5] When cyber-attacks against industrial control systems (ICS) are successful, they may have far-reaching, even catastrophic, consequences for people and the planet because of how closely related the variables of the underlying control gaps and physical operations are. In light of this, it is more important than ever to implement stringent safety and security measures to detect and prevent breaches into ICS. [1] Methods based on signatures and anomalies are popular for strike detection and attribution. The limitations of signature-based and anomaly-based detection and attribution techniques have prompted attempts to propose hybrid-based approaches. [6] The continual network updates that lead to different Breach Discovery System (IDS) typologies make hybrid based approaches unreliable, even if they are effective at recognising rare triggers. [7] Other than this, traditional attack attribution and detection approaches rely heavily on analysing network information, such as IP addresses, transmission ports, web traffic length, and packet duration's. Machine learning (ML) and deep neural network (DNN) based strike detection and acknowledgment methods have therefore recently seen a resurgence of attention. Another way to classify attack detection technologies is as either host-based or network-based. Typical methods for detecting strikes in network traffic include distributed neural networks (DNN), fuzzy logic, single-class or multi-class Support Vector Machines (SVMs), and overt clustering. These methods swiftly identify malicious assaults by analysing data from real-time online traffic. On the other hand, skilled strikes and complex assaults may go undetected by attack detection systems that rely only on host and network data.

SURVEY OF RESEARCH

An important concept embedded within a wider spectrum of networked items and digital sensors is the Net of Things (IoT), as stated by Tobby (2017). There has been an explosion of applications made possible by this new technology, which signifies a sea shift in how people use the Internet and presents both benefits and challenges, particularly with regard to critical infrastructure. Some examples of Internet of Things (IoT) products that hackers have used to compromise security systems include printers, thermostats, and videoconferencing equipment. Improvements in agriculture, industry, power generation and distribution, smart homes, energy management systems, intelligent vehicles, online traffic systems, road and bridge sensors, and other areas have all benefited from Internet-enabled frameworks. The unchecked growth of the Internet of Things (IoT) poses several concerns, including threats to personal data security and privacy, disruptions to communications networks, and increased demand for electricity, despite the fact that it has opened up many performance-enhancing opportunities. The efficacy of Internet connection also increases vulnerability to safety and security offence via the exploitation of IoT information, which leads to fraudulent breaches of the networks that support vital infrastructure. Even if an ICS is air-gapped and hence a closed system, it is still susceptible to attacks carried out via physical accessibility, such as infected portable devices. Many industrial control systems (ICSs) are now online, leaving them open to a variety of threats as contemporary technology continues to advance. The interconnection of various infrastructures is being facilitated by computers and communications, which are essential in and of themselves. Critical facilities are more susceptible to cyber attacks due to their reliance on computer systems and networks, which makes them more susceptible to disruptions in one network potentially affecting other networks.

The integration, monitoring, or management of CPS processes is carried out by a computing core, and Arash and Stuart (2015) state that CPS provides cyber-based instructions for controlling physical components. A CPS creates a control loop for each physical component by combining actuators, control refining systems, sensor units, and interaction cores. Programmable logic controllers, distributed control systems, and supervisory control and data acquisition (SCADA) are the main components of a control and data system (CPS). From monitoring a country's electricity circulation to operating sensors inside a factory, SCADA systems gather and manage geographically scattered assets. They play an important role in many important systems, including water distribution networks, oil refineries, and electrical power grids. Contrarily, DCS is in charge of the controllers that are physically located in the same area and are working together to complete a certain job. When it comes to industrial processes and components, both SCADA and DCS use PLC devices. It is common practice for operators to setup PLCs from a Windows-based manufacturer. Process monitoring and the establishment of control criteria are only two examples of the many controlling duties performed by operators using SCADA and DCS. The article by Lange et al. (2016) notes that a company's ability to achieve its goals is highly dependent on the CIS that support those goals. As a result, the efficiency and fulfilment of the linked goal capacity are impacted by online strikes on CIS. Delivering power from suppliers to consumers is the basic objective of an electrical grid. Their connection to CIS serves as a means of monitoring and control. A variety of network services spanning various devices and sub networks inside a facility might affect an application's separability, efficiency, and dependability. Threat actors may take advantage of security holes in web apps or desktop software to compromise sensitive customer data or intellectual property, which increases the danger of software applications released into the public. The BYOD pattern's inherent vulnerabilities have not been adequately addressed by consistent, positive policies. As the number of mobile devices continues to rise, there is a greater risk of both accidental and malicious security breaches, which is a major concern when it comes to personal information. Consequently, it is becoming an issue for businesses to ensure that the software being downloaded to such devices is secure. This is crucial since systems like Android by Google undertake little testing of apps for security flaws before letting users download them from their app store.

EXISTING SYSTEM

The relative summary indicated that the LR method performed the lowest, with a recall of 0.4744, while the RF formula had the most successful strike detection (recall: 0.9744). The ANN algorithm ranked fifth with a recall of 0.8718. The authors also mentioned that 12.82% of attacks were unnoticed by the ANN, and that 0.03% of normal samples were mistakenly considered attacks. The ML algorithms LR, SVM, and KNN are particularly vulnerable to data imbalances since they treat multiple attack samples as normal samples. Put simply, they are unfit for use in industrial control system (ICS) attack detection. To identify cyber-attacks on gas pipelines, the authors provided a KNN method in [12]. They over sampled the dataset to achieve equilibrium, which reduced the consequence of utilising an imbalanced dataset in the method. By applying the KNN to the balanced dataset, they achieved 97% accuracy, 0.98 precision, 0.92 recall, and 0.95 f-measure. In order to create a two-step system for abnormality detection, the authors of [13] offered a Rational Analysis of Information (LAD) method for extracting rules and patterns from the data collected by the sensing units. First, we determine whether the system is stable or unstable; second, we find out if an attack has occurred.

Along with DNN, SVM, and CNN, they evaluated the suggested LAD method's performance. Though the lad technique performed better in recall and f-measure, these studies showed that the DNN was more accurate overall.

Attractive Features of the Current System

Models that include process and physical data may compete with system monitoring even in the absence of supervision, since they do not depend on comprehensive understanding of cyber-threats. Strong defences may not be enough to thwart a highly intelligent attacker with the resources to do so, such as a nation-state that has developed a persistent danger star. To add insult to injury, the majority of current approaches model only the typical behaviour of a system and label any deviation from this pattern as abnormal, ignoring the fact that ICS data comes from an imbalanced source. This could be because both real-world scenarios and available datasets have very few attack samples.

PROPOSED SYSTEM

Operating as a cyber-physical system, a critical transportation framework incorporates the Net of Points based cordless sensing unit network. Nevertheless, because to the inherent cyber vulnerabilities of IoT devices and the absence of control barriers that may protect it, the new kind of transportation infrastructure that is enabled by the Internet of Things is vulnerable to cyber-physical attacks in the sensing region. No one checks the Internet of Things (IoT) enabled transportation infrastructure for cyber-physical attacks because conventional threat assessment methods treat the cyber and physical domains as two distinct environments. This leaves stakeholders, including drivers, civil engineers, and security and safety designers vulnerable. This research proposes a novel approach to risk analysis for cyber-physical attacks compared to wireless sensor networks based on the Internet of Things. With this approach, new cyber-physical characteristics—such as risk resources (like aims), vulnerabilities (like a lack of verification methods), and types of physical repercussions (like casualties)—can be identified and proposed. The degree and relevance of these qualities are multiplied to calculate cyber-physical hazard. An Internet of Things (IoT) enabled bridge is tested in cyber-physical attack scenarios using Monte Carlo simulations and level sensitivity analysis. According to the results, there are 76.6% of replacement scenarios with high-risk factors, and removing control hurdles in both the physical and cyber domains may reduce the cyber-physical hazard by 71.8%. In addition, cyber-physical hazard stands up when the significance of the factors considered during risk assessment is overlooked.

Individuals or organisations attempting to integrate the cyber domain name into risk assessment methods for their systems will find the method very appealing.

THE BENEFIT OF THE ADVISED

These days, a lot of people are curious in the Internet of Things (IoT) because it lets people improve their lives and keep up with the latest tech in the cyber-physical world. In terms of the technology they're based on and the storage document types they utilise, the IoT side tools are diverse. Before actually transmitting data, these devices must authenticate each other using extremely secure shared authentication methods. One of the most important parts of peer-to-peer communication is mutual authentication. These devices, which are limited in resources, are able to authenticate with each other thanks to secure session secrets. A device may be accredited and granted access to shared resources after successful verification. In order to prevent data privacy breaches that might jeopardise honesty and secrecy, it is necessary to validate a device that requests data transmission.

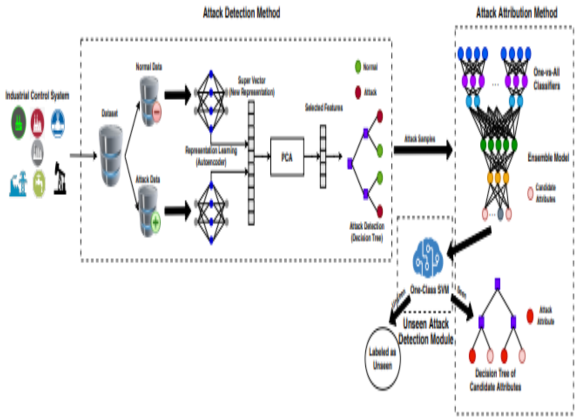


Fig.1. System diagram

METHODOLOGY

Both the portrayal finding and the discovery phases make up the planned assault discovery. A DNN variant that learnt bulk course patterns but missed minority class characteristics was produced by applying a conventional unsupervised DNN to an unbalanced dataset. To address this issue, some researchers have attempted to balance the dataset by adding new instances or eliminating certain samples before feeding the data to a DNN. Nevertheless, creating or removing samples is not a practical choice for ICS/IIoT security applications. It is challenging to verify produced samples in a real network because to the sensitivity of ICS/IIoT systems; this is because the produced assault instances may be harmful to the network and have significant environmental or human life-threatening effects. Recognition of the produced instances also takes a long time. Also, because most ICS/IIoT datasets include strike samples that are less than 10% of the dataset and since deleting 80% of the dataset gets rid of much of the dataset knowledge, removing the regular information from a dataset isn't the best solution. Researchers in this paper came up with a novel deep representation discovery method to train DNNs to handle imbalanced datasets without introducing new samples, altering existing ones, or correcting any of the previously mentioned difficulties. Each of the two stacked vehicle encoders in this configuration was in charge of finding patterns from a single course; they were not being observed. The output of every given design accurately reflects its inputs since all of them aim to eliminate abstract patterns from one course while ignoring others. There were three input/output layers in the stacked auto encoders, each containing an encoder and a decoder. A higher-dimensional region with 800 dimensions, a 400-dimensional area, and finally a 16-dimensional space were all mapped to the input representation by the encoder layers. The encoder functions of an automobile encoder. On the other hand, the decoder layers aimed to recreate the input representation by mapping the 16-dimensional new representation to the 400-dimensional, 800-dimensional, and input representations. The car encoder's decoder purpose. Using trial and error, we were able to choose these hyper parameters that provide the best performance in f-measure while keeping the construction complexity to a minimum.

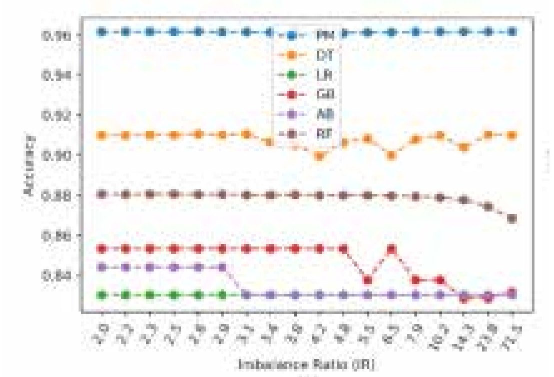


Fig.2. Output results with attacks.

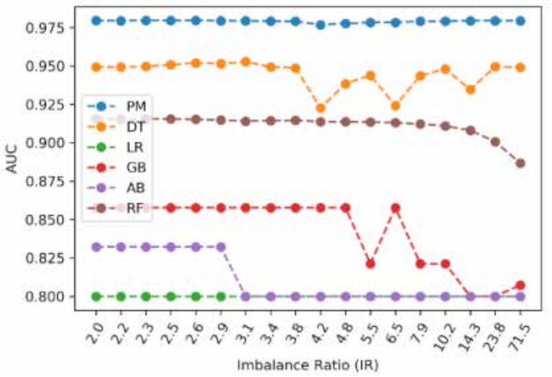


Fig.3. Ratio of imbalance.

CONCLUSION

This research presented a novel architecture for imbalanced ICS data that utilises two-stage ensemble deep learning for attack detection and acknowledgment. In order to locate the strike samples, the strike detection phase uses deep depiction learning to transfer the instances to the new higher-dimensional room and then employs a DT. This part of the process is good at finding attacks that were previously undiscovered and can withstand skewed datasets. With each attack feature learned separately, the strike acknowledgment step employs a series of one-vs-all classifiers. As shown, the whole thing comprises a complex DNN with a fully connected and partially linked component that can correctly identify cyber-attacks. The proposed structure has a complex design, but the training and screening stages are computationally simple compared to other DNN-based approaches in the literature. Specifically, they are O(n 4) and O(n 2), where n is the number of training instances. In addition, compared to earlier efforts, the suggested structure has much superior recall and f-measure when it comes to finding and attributing example prompts. In order to improve the detection of anomalies that the discovery portion cannot detect, such as creating a consistent account for the whole system and its attributes, the construction of a cyber-threat finding component is an important aspect of the future growth.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "TOWARD DETECTION AND ATTRIBUTION OF CYBER-ATTACKS IN IOT-ENABLED CYBER-PHYSICAL SYSTEMS", which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department Of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. F. Zhang, H. A. D. E. Kodituwakku, J. W. Hines, and J. Coble, "Multilayer Data-Driven Cyber-Attack Detection System for Industrial Control Systems Based on Network, System, and Process Data," IEEE Transactions on Industrial Informatics, vol. 15, no. 7, pp. 4362–4369, 2019.
2. R. Ma, P. Cheng, Z. Zhang, W. Liu, Q. Wang, and Q. Wei, "Stealthy Attack Against Redundant Controller Architecture of Industrial CyberPhysical System," IEEE Internet of Things Journal, vol. 6, no. 6, pp. 9783–9793, 2019.
3. E. Nakashima, "Foreign hackers targeted U.S. water plant in apparent malicious cyber attack, expert says." [Online]. Available: https://www.washingtonpost.com/blogs/checkpointwashington/post/foreign-hackers-broke-into-illinois-water-plant-controlsystem-industry-expert-says/2011/11/18/gIQAgmTZYN blog.html
4. G. Falco, C. Caldera, and H. Shrobe, "IIoT Cybersecurity Risk Modeling for SCADA Systems," IEEE Internet of Things Journal, vol. 5, no. 6, pp. 4486–4495, 2018.
5. J. Yang, C. Zhou, S. Yang, H. Xu, and B. Hu, "Anomaly Detection Based on Zone Partition for Security Protection of Industrial Cyber-Physical Systems," IEEE Transactions on Industrial Electronics, vol. 65, no. 5, pp. 4257–4267, 2018.
6. S. Ponomarev and T. Atkison, "Industrial control system network intrusion detection by telemetry analysis," IEEE Transactions on Dependable and Secure Computing, vol. 13, no. 2, pp. 252–260, 2016.
7. J. F. Clemente, "No cyber security for critical energy infrastructure," Ph.D. dissertation, Naval Postgraduate School, 2018.
8. C. Bellinger, S. Sharma, and N. Japkowicz, "One-class versus binary classification: Which and when?" in 2012 11th International Conference on Machine Learning and Applications, vol. 2, 2012, pp. 102–106.
9. I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org
10. Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1798–1828, 2013.
11. M. Zolanvari, M. A. Teixeira, L. Gupta, K. M. Khan, and R. Jain, "Machine Learning-Based Network Vulnerability Analysis of Industrial Internet of Things," IEEE Internet of Things Journal, vol. 6, no. 4, pp. 6822–6834, 2019.
12. I. A. Khan, D. Pi, Z. U. Khan, Y. Hussain, and A. Nawaz, "HML-IDS: A hybrid-multilevel anomaly prediction approach for intrusion detection in SCADA systems," IEEE Access, vol. 7, pp. 89 507–89 521, 2019.
13. T. K. Das, S. Adepu, and J. Zhou, "Anomaly detection in industrial control systems using logical analysis of data," Computers & Security, vol. 96, p. 101935, 2020.
14. J. J. Q. Yu, Y. Hou, and V. O. K. Li, "Online False Data Injection Attack Detection With Wavelet Transform and Deep Neural Networks," IEEE Transactions on Industrial Informatics, vol. 14, no. 7, pp. 3271–3280, 2018.
15. M. M. N. Aboelwafa, K. G. Seddik, M. H. Eldefrawy, Y. Gadallah, and M. Gidlund, "A machine-learning-based technique for false data injection attacks detection in industrial iot," IEEE Internet of Things Journal, vol. 7, no. 9, pp. 8462–8471, 2020.

An Efficient Spam Detection Technique for IoT Devices using Machine Learning

Peddi Naimimisha, Guddeti Baby Saroja,
Peteri Shashank
B.Tech Student
Dept. of Computer Science and Engg. (Data Science)
CMR Technical Campus, Medchal
Hyderabad, Telangana

K. Srinivas
Professor
Dept. of Computer Science and Engg. (Data Science)
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ phdknr@gmail.com

ABSTRACT

Connected via wired or wireless networks, the millions of devices that make up the Internet of Things (IoT) send and receive data. Internet of Things (IoT) devices generate massive amounts of data using a wide range of techniques, the data quality of which is characterised by its speed with respect to time and location, and by no little rise in volume. In this context, ML algorithms may be crucial in protecting IoT systems, ensuring security based on biotechnology, and making odd discoveries to enhance functionality. Meanwhile, hackers often use learning algorithms to target smart IoT-based devices' vulnerabilities. In light of these considerations, we propose in this brief essay a method for ensuring the security of IoT equipment via the detection of spam using ML. To achieve this goal, it is proposed to use Spam Discovery in IoT with an AI framework. Here, a large number of input feature sets are used to evaluate five different ML versions using different metrics. All of the models incorporate the tweaked input functions while calculating a spam score. The reliability of the IoT gadget is shown by its score across many criteria. In order to validate the suggested strategy, data collected from REFIT Smart Homes is used. The results demonstrate how much more efficient the suggested strategy is compared to the other current plans.

KEYWORDS: *REFIT, IOT, ML, Power, Spam detection.*

INTRODUCTION

Interconnection of Things (IoT) It enables the integration and use of physical items from different locations. Implementing such community management and monitoring requires robust privacy and defence mechanisms, which might be challenging in such a setting [1]. The goal of Internet of Things (IoT) security measures is to prevent infiltration, eavesdropping, spam, malware, hacking, phishing, and denial-of-service (DoS) assaults.

The scope and nature of the threat dictate the measures needed to secure Internet of Things devices. Safety websites are compelled to work together due to user activities. The location, kind, and intended use of Internet of Things (IoT) devices dictate the precautions to take in order to keep sensitive data secure. In an intelligent organisation initiative, Internet of Things (IoT) smart security cameras may capture different specifications for careful study and assessment [2]. Due diligence is required for fully internet-connected devices as the vast majority of IoT devices rely on networks. The efficient implementation of safety and personal privacy features by Internet of Things (IoT) devices put up in a company's workplace is not rare. To prevent the disclosure of statistics and to guarantee a certain level of personal privacy, wearable, for example, collect data on a person's health and fitness and transmit it to a paired phone. Market research indicates that between 25 and 30 percent of workers connect their own Internet of Things (IoT) devices to the company's internal network [3]. With the proliferation of the Internet of Things comes the target market, which includes both allies and adversaries.

Nevertheless, since ML presents new entry points for attacks, IoT devices take a defensive stance by defining critical parameters inside security protocols to toggle between computing, privacy, and safety. This process is challenging since it is also tough for an IoT system with limited resources to estimate the current network and attack history [4].

Part A. Financial Transactions The following payments are detailed in this document, expanding upon earlier discussions. The SPAM Detection Scheme has been tested with five separate gadgets that are conscious of fashion [5].

2) To calculate the pastiest score for each version, a formula is suggested and thereafter used for intelligent discovery and selection.

The integrity of IoT instruments is assessed using unique rating ranges, taking into account the degree of pastiest derived in the preceding step.

Organisation B, Following through is crucial for the rest of the task. Important panels are discussed in the second part. The suggested synopsis is covered in Section 3.

RELATED STUDY

Web spam detection is centred upon this suggestion to stop IoT devices from causing hazardous actions. We looked at several systems that relied on form to find spam from Internet of Things devices. We want to resolve challenges with home-based Internet of Things (IoT) devices. The suggested technique, on the other hand, considers all relevant design characteristics before verifying it using machine learning models.

There are a lot of phases that make up the process that gets you to the end result.

1) Creating the function: When given the right timeframes and properties, maker proficiency algorithms perform as expected. We are all aware that instances are statistics of real-world rates collected from real-world, globally dispersed intelligent entities. One step in the feature engineering process is the elimination or selection of attributes.

Function reduction: This technique is used to reduce the amount of data. The goal of attribute reduction is to simplify attributes by reducing their complexity. Over processing, huge memory needs, and processing power are all reduced by this cutting-edge technology. A number of distinct methods exist for removal. One of the most used is principal component analysis (PCA) [5]. However, PCA and the following IoT parameters are the methods used in this approach.

Time for evaluation: The data set for the experiments includes the statistics recorded throughout the course of the 18 months. We looked at one month's worth of papers to ensure even greater accuracy and outcomes. In light of this fact, the weather is the primary determinant for IoT tool operation, and the most dissimilar month has been considered.

Software for the web: The only things that can run them without an Internet connection are protected. Devices included in the statistics collection include the following: television, peak container collection, DVD player/recorder, high-fidelity system, electric heating system, refrigerator, dishwasher, toaster, coffee maker, pot, electric heating system, dryer, DAB radio, home computer, display Devices such as a computer, printer, router, heater, freezer, electric heater, light, alarm clock, lava lamp, video player, television, set-top box, CD player, and centre

- A Choice of function: One of the most crucial aspects of characteristics is computed during this phase. Its purpose is to determine the weight of each position. This line of thinking uses entropy-based full elimination for feature choice.

The primary principle of filtering is degeneration, which is a system of rules that determines the weights of discrete qualities by looking at the link between certain features and continuous traits. Uncertainty about the symmetrical ratio of revenues and profits is one of three aspects where this deterioration entirely filters information. These capabilities are expressed using the Statistics syntax. Feature that is beneficial (technique, points, equipment). Disputed relationships including system, facts, and division. Device, process, or information uncertainty The reasoning for the attributes that are described here.

a) Method: This section provides a synopsis of the steps used to compile the recommendations.

(b) Specifics: It's a collection of research study papers outlining the attributes that will be considered.

c) System: this is the yardstick by which degeneration is measured. The cost of a record is automatically borne by it.

PROPOSED SYSTEM

- The SPAM detection strategy is double-checked using five distinct models of equipment efficacy.
- Every model used to identify specificity and make a practical selection should have its specificity rating computed using a specific formula.
- The dependability of IoT gadgets is assessed using unique score metrics, based on the specificity level computed in the previous stage.
- Oversaw the process of identifying Methods: Patterns used to categorise the region in order to detect assaults include support vector machines (SVMs), semantic networks (NN), K-closest communities (K-NN), and random woodland areas (RBAs). Threats to IoT devices may be detected by these models as DoS, DDoS, invasion, and malware attacks.
- Methods using artificial intelligence that are not being monitored: In a label-free environment, these techniques beat opposite number strategies. Forming groups is how it works. We employ multivariate correlation analysis to detect denial-of-service attacks in IoT devices.
- Improving Tools for Procedures: These blueprints let the Internet of Things gadget choose crucial specs and safety procedures for certain assaults via trial and error. General verification performance and malware identification have both benefited from the use of the Q research.

SIMULATION RESULTS

The Generalised Bayesian Linear Model (BGLM) is a constant, asymptotically green, asymptomatically normal, single-mode document option for exponential circles of family members. The main focus of Bayesian approaches is on these crucial components.

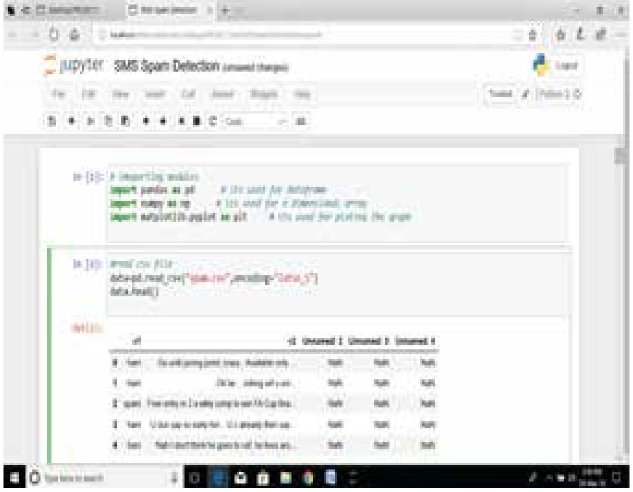
The inclusion of prior information is the first step. Ideally, the data shown above is distributed according to a specification's probability and is quantitatively differentiable in circulation.

Secondly, a probability function is linked to the pre-programmed value. Impacts are represented by the residential shell property.

Thirdly, a later distribution of the developed specified worth is the outcome of combining the main function with the potential feature.

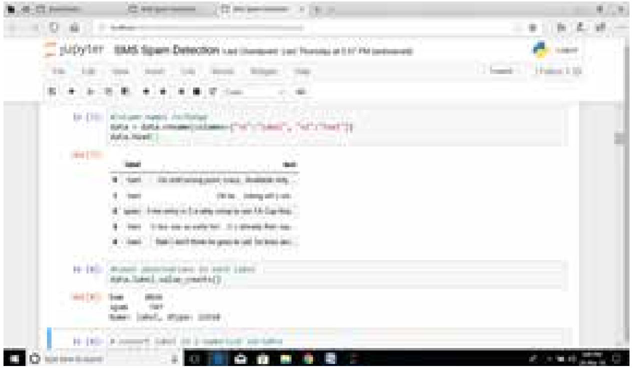
A population parameter for the potential values was experimentally circulated using simulations obtained from the post-distribution.

Fifth, extremely simple data is used to summarise the analytical circulation of the following simulations.



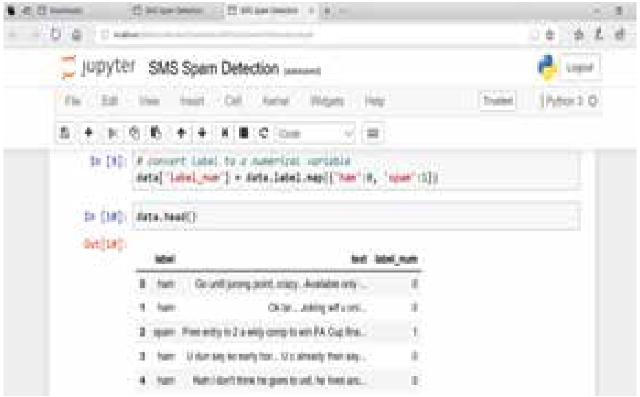
```
In [1]: # importing modules
        import pandas as pd      # its used for dataframe
        import numpy as np      # its used for n dimensional array
        import matplotlib.pyplot as plt     # its used for plotting the graph

In [2]: #read csv file
        data=pd.read_csv("spam.csv",encoding="latin_1")
        data.head()

Out[2]:
     v1                                                 v2  Unnamed: 2  Unnamed: 3  Unnamed: 4
0   ham  Go until jurong point, crazy.. Available only ...         NaN         NaN         NaN
1   ham                      Ok lar... Joking wif u oni...         NaN         NaN         NaN
2  spam  Free entry in 2 a wkly comp to win FA Cup fina...         NaN         NaN         NaN
3   ham  U dun say so early hor... U c already then say...         NaN         NaN         NaN
4   ham  Nah I don't think he goes to usf, he lives aro...         NaN         NaN         NaN
```

Fig. 1. DATA set


jupyter SMS Spam Detection Last Checkpoint: Last Thursday at 5:57 PM (autosaved)
Logout
Trusted
Python 3

```
In [3]: #column names changing
        data = data.rename(columns={"v1":"label", "v2":"text"})
        data.head()

Out[3]:
   label                                               text
0    ham  Go until jurong point, crazy.. Available only ...
1    ham                      Ok lar... Joking wif u oni...
2   spam  Free entry in 2 a wkly comp to win FA Cup fina...
3    ham  U dun say so early hor... U c already then say...
4    ham  Nah I don't think he goes to usf, he lives aro...

In [4]: #count observations in each label
        data.label.value_counts()

Out[4]: ham     4825
        spam     747
        Name: label, dtype: int64

In [5]: # convert label to a numerical variable
```

Fig. 2.SMS Spam detection.



```
In [9]:  # convert label to a numerical variable
         data['label_num'] = data.label.map({'ham':0, 'spam':1})

In [10]: data.head()

Out[10]:
     label                                              text  label_num
  0  ham     Go until jurong point, crazy.. Available only ...          0
  1  ham                         Ok lar... Joking wif u oni...          0
  2  spam  Free entry in 2 a wkly comp to win FA Cup fina...          1
  3  ham     U dun say so early hor... U c already then say...          0
  4  ham     Nah I don't think he goes to usf, he lives aro...          0
```

Fig. 3. Spam detection in OUTPUT

CONCLUSION

The spam specifications of Internet of Things (IoT) devices and their use by style-conscious devices are uncovered by the suggested framework. Experiments use a pre-processed IoT dataset that was created using the function engineering approach. Any Internet of Things (IoT) technology has a spam score since it uses a framework to test out different domain name designs. In the smart home, it improves the conditions needed to operate Internet of Things devices.

We want to make IoT devices even more secure and dependable in the future by considering their surroundings and weather conditions.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled “AN EFFICIENT SPAM DETECTION TECHNIQUE FOR IOT DEVICES USING MACHINE LEARNING”, which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department of Computer Science and Engineering, Guide and Teaching and Non-Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. Wu F, Zhao S, Zhang YZ, 2020), “A new coronavirus associated with human respiratory disease in china “265–269.

2. Medscape Medical News, “The WHO declares public health emergency for novel coronavirus”.

3. Wang J et al., 2020, “Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study,pp.507–513

4. World health organization: https://www.who.int/new-room/g-adetail/ q-a-corronaviruses#:/text=symptoms. Accessed 10 Apr 2020

5. Wikipedia coronavirus Pandemic data: https://en.m.wikipedia. org/wiki/ Template:2019%E2%80%9320_coronavirus_pandemic_data. Accessed 10 Apr 2020

6. H. Oh and H. Eun, H. Lee, 2013, “Conditional privacy preserving security protocol for nfc applications,”, pp. 153–160.

7. K. Venayagamoorthy and R. V. Kulkarni and G., 2009, “Neural network based secure media access control protocol for wireless sensor networks,”, pp. 1680–1687

8. Lin, D. Niyato and M. A. Alsheikh, 1996, “Machine learning in wireless sensor networks: Algorithms, strategies, and applications,” pp. 1996– 2018, 2014.

9. E. Guven and A. L. Buczak, 2015, “A survey of data mining and machine learning methods for cyber security intrusion detection,”, pp. 1153–1176.

10. A. Feizollah and F. A. Narudin, 2016, “Evaluation of machine learning classifiers for mobile malware detection,” pp. 343–357.

HIV-Infected Patients' Chronic Kidney Disease Stage Identification using Machine Learning

K Venkata Naga Aditya, T. Vaishnavi Sagar, K Karthik
B.Tech Student
Dept. of Computer Science and Engg. (Data Science)
CMR Technical Campus, Medchal
Hyderabad, Telangana

B. Ramji
Assistant Professor
Dept. of Computer Science and Engg. (Data Science)
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ ramji.cse@cmrtc.ac.in

ABSTRACT

Silent illness is chronic kidney disease. It is difficult to diagnose or assess the severity of chronic kidney disease (CKD) since its symptoms, when present, are often vague and nonspecific, unlike those of other chronic illnesses. It is generally possible to prevent the worsening of chronic renal disease with early diagnosis and therapy. Significant morbidity and death are outcomes of chronic kidney disease (CKD), which progresses over time. Because the kidneys are responsible for equilibrium, chronic kidney disease (CKD) may impact almost every bodily function. The only way to halt the course of chronic kidney disease (CKD), Early detection and treatment can improve results and maintain a high quality of life. Regardless of the underlying cause, renal damage or a Glomerular Filtration Rate (GFR) < 60 mL/min/1.73 m2 over a period of three months or longer, may also be characterised as chronic kidney disease (CKD).Our objective is to create a convolutional neural network (CNN) model that can classify chronic kidney disease (CKD).

KEYWORDS: *CKD, CNN, GFR, Lung disease.*

INTRODUCTION

Nowadays, chronic kidney disease (CKD) poses a significant threat to public health and welfare. Lab tests can detect chronic kidney disease, and there are treatments that can slow or stop its progression, increase survival, improve quality of life, decrease the risk of cardiovascular disease, and decrease the problems connected by a decrease glomerular filtration rate (GFR). Chronic kidney disease (CKD) may be brought on by a myriad of factors, including not drinking enough water, smoking, eating the wrong foods, not getting enough sleep, and many more. There were 753 million cases of this illness in the globe in 2016, with 417 million women and 336 million men afflicted. The majority of cases are discovered in their latter stages, which often results in renal failure. Evaluation of urine using serum cretonne level is the foundation of the current medical diagnostic system. This function makes use of a wide variety of medical procedures, including screening and ultrasonically approaches. Patients with hypertension, a history of heart disease, a current or previous medical condition, or a family history of kidney disease are all candidates for screening. One part of this plan is to measure the albumin-to-creatinine ratio (ACR) in the first urine illustration in use first object in the morning. Another part is to estimate the GFR from the serum creatinine level. This study delves into artificial intelligence approaches such as SVM and ACO, which enhance prediction accuracy via function reduction and feature selection.

The kidneys increase the risk of hypertension and cardiac arrest by producing vocative hormones via the renin-angiotensin system in response to fluid overload, which in turn increases the risk of hypertension. Although uremic toxins may attenuate this effect to some extent, people with CKD are more prone than the general population to develop atherosclerosis and cardiovascular disease as a consequence. Those whose diagnoses include both chronic kidney disease and cardiovascular disease are much worse than those whose diagnoses include just cardiovascular disease.

Symptoms of uremia, which may include drowsiness, pericarditis, and encephalopathy, develop when urea builds up, after azotemia. Because of its high concentration in the blood, urea is expelled in eccrine sweat in large quantities, which causes it to crystallise on the skin as it evaporates ("uremic frost").

When there is an excess of potassium in the blood, a condition known as hyperkalemia, symptoms might include fatigue and even fatal irregular heartbeats. Once the glomerular filtering rate drops below 20–25 mL/min/1.73 m2, the kidneys' ability to eliminate potassium is significantly reduced, and hyperkalemia often develops. Acidemia (which causes potassium to move out of cells) and insulin deficiency may worsen hyper kalemia in chronic kidney disease.

LITERATURE SUEVEY

By 2020, according to J. Snegha The approach suggested in [10] makes use of a plethora of data mining techniques, such as the Back breeding semantic network and the Random Forest algorithm. Here, they compare the two algorithms and conclude that, thanks to the usage of a supervised learning network known as a feed-forward neural network, the Back Proliferation method produces superior results. In [Mohammed Elhoseny, 2019], a technique for chronic kidney disease was detailed that make exploit of ACO and thickness-based function selection. Wrapper methods are used by the system for feature options.

(Chakraborty, Baisakh) AI approaches such K-Nearest Neighbour, Decision trees, Random Forests, Naïve Bayes, Support Vector Machines, Multi-Layer Perceptrons, and Logistic Regression Formula were proposed for the development of CKD prediction systems in [9]. We put them to use and compare how well they work with the outcomes of the precision, recall, and precision tests. The last stage before implementing this method is Random Woodland.

The method proposed by [Arif-Ul-Islam, 2019] makes use of Boosting Classifiers, Ant-Miner, and J48 Decision Tree for disease forecasting. This research aims to do two things: first, to assess the efficacy of improving formulas for CKD discovery; and second, to obtain policies that demonstrate the links between the characteristics of CKD. The experimental results reveal that Ada Boost's performance was much lower than Logit Boost's.

An approach to chronic kidney disease (CKD) prediction that makes use of extreme finding machines and ACO was proposed by S. Belina V. in 2018. The MATLAB device is used for classification, and ELM has minor optimisation constraints. According to the Sigmoid additive class of SLFNs, this approach is superior. In 2018, Siddheshwar Tekale A machine learning system that employs choice tree SVM techniques was described in [8]. We found that SVM gives the best outcome after comparing the two methods. Because its prediction procedure is less time-consuming, doctors can assess patients in much less time.

A system that makes use of the Back Propagation Neural Network technique for forecasting was described by Nilesh Borisagar in 2017. Levenberg, Scaled Conjugate, Bayesian regularisation, and the resistant back proliferation method are all covered here. The application function is implemented using Matlab R2013a. Levenberg and Bayesian regularisation are not as effective as scaled conjugate gradient and are resistant back breeding when it comes to training time.

In 2017, Guneet Kaur Using Hadoop's data mining algorithms, the authors of [7] proposed a method to predict CKD. Two data mining classifiers, such as KNN and SVM, are used by them. Here, the anticipatory analysis is performed using the data columns that were picked by hand. When compared to KNN, the SVM classifier provides superior accuracy in this setup.

6. Foretelling the Prognosis of Chronic Kidney Disease Utilising Deep Idea Network in the Year 2021. Predicting kidney-related disorders is achieved by the use of a modified Deep Idea Network (DBN) as a category formula, with the activation feature Softmax and the loss function Specific Cross-entropy. the dataset is retrieved from the machine learning database at UCI and pre-processing is carried out to address the missing values. When compared to the current designs, the suggested version achieves much greater performance, with a precision of 98.52%. Correct CKD predictor and classifier are therefore provided by the suggested version.

In 2021, a two-stage semantic network was used to predict chronic renal disease. The two-stage neural network is the basis for this version's prediction. It combines the attribute testing method in the very high dimensional structure with the deep knowing method. The proposed approach may help identify potentially useful biomarkers and the stage of chronic kidney disease. Due to the absence of versions for handling very high-dimensional datasets, efficiency is still limited.

8. Chronic Kidney Disease Prognosis with the Use of Semantic Networks and Machine Learning Architectures (2022) 1. Regression tree, K-Nearest Next-door neighbour, SVM, and hybrid model... It makes use of mathematical data. predict the illness using a hybrid model and significantly improve the accuracy of the prediction. It does, however, rely on a patient's historical CKD information.

9: Prognosis, Course, and Outcomes of Chronic Kidney Disease in the Elderly (2022). The DNN model is the foundation. Results, Progression, and Prognosis of Chronic Kidney Disease in the Elderly are Unveiled by the Model. In HIV patients, however, it inhibits the progression of CKD, especially in cases when protein is detected.

10. Regarding the Application of Artificial Intelligence to Chronic Kidney Disease Prognosis (2022). The fundamental elements are as follows: a chi-square automatic interaction detector, logistic regression, a synthetic semantic network, C5.0, a linear support vector machine with penalty L1 and charge L2, and assorted trees. In addition, the GINI coefficient, placement beneath the contour, F-measure, recall, and accuracy have all been calculated. However, a significant amount of time is needed for it.

11. A System for Intelligent Diagnosis and Classification of Kidney Diseases (2022) The design's feature-based prediction model for renal ailment detection is top-notch. Numerous equipment learning techniques, including k-nearest neighbors formula (KNN), synthetic semantic networks (ANN), support vector machines (SVM), naïve bayes (NB), and others, were used to test the prediction versions. Chi-Square test feature selection and Recursive Attribute Removal (RFE) techniques were also used. But this was the point at which the hybrid approach truly faltered.

12. A Method Using Artificial Intelligence to Identify Chronic Kidney Disease (2022). When it comes to data imputation and medical sample diagnosis, the suggested CKD analysis method is feasible. This version is applicable for small datasets only, after the establishment of missing values in the data using KNN imputation without supervision.

13. Artificial intelligence techniques for the prediction of chronic renal disease (2022). Prediction models include decision trees, assisted vector machines, and random woodland (RF) (DT). The data used is sourced from the UCI Database, which contains 400 data sets characterised by 25 variables. the practical considerations of data gathering and emphasises the merit of combining domain knowledge with random forest classifier and tree classifier to achieve great accuracy with little feature bias. Both the function selection and the design's efficiency are lacking.

EXISTING SYSTEM

Predicting chronic kidney disease using machine learning techniques isn't perfect. Artificial intelligence designs provide much less accurate results on ultra audio photographs. The primary basis for diagnosing chronic renal sickness is mathematical data, which may be intrusive and dangerous when trying to identify these illnesses. The urnie test and elevated blood pressure led to this medical diagnosis. The mathematical data of the patient's medical records allows for the diagnosis of chronic renal disease. For a full dataset, a large amount of space is required. Numerical datasets often include noisy information and missing values, which makes the process of normalization laborious and time-consuming.

PROPOSED SYSTEM

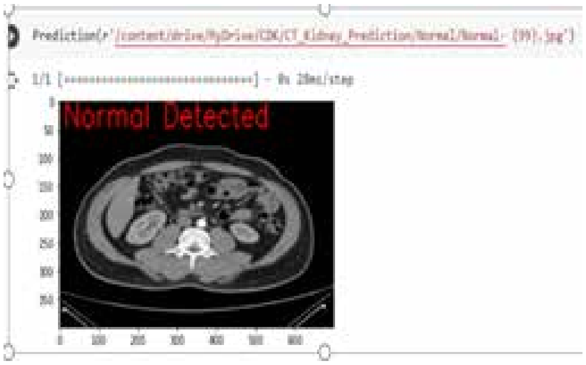
In Proposed system we find that most of the prior CKD prediction models either deal with a method with poor accuracy or a restricted application variety when assigning missing values. Therefore, we provide a method to increase the number of ckd analysis designs' applications in this work. The accuracy of the design has also been enhanced. The suggested work's compensation is a method of picture filtering based on CKD category and semantic networks. Pictures of Input-Labeled Ultrasound Pre-processing involves cleaning up image data, including removing distortion, so it may be used for further data refinement. A series of convolutional layers is applied to the pictures.

METHODOLOGY

Using the CNN model and its layers, our proposed architecture would undoubtedly detect the chronic kidney disease. The kidney ultrasound images will be evaluated by the system, and the results will indicate whether the kidney is sick or not. Based on the training approach, the system will be able to classify chronic kidney conditions into four groups using those ultrasound pictures. If the system determines that the photo we provided is normal, the model will display Typical Detect on the photo. Similar to the other three categories, they will also undergo refinement (rock, cyst, normal). This is how our proposed design would employ deep learning to identify the chronic kidney ailment.

Case study 1: The result is accurate.

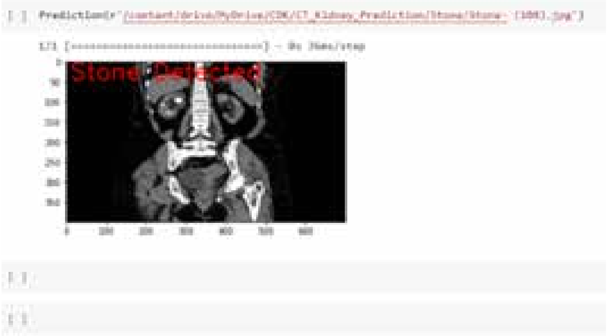
User: Regular Forecasted Results: Standard Actual Results: Consistent.



```
Prediction(r'/content/drive/MyDrive/CDK/CT_Kidney_Prediction/Normal/Normal- (99).jpg')
1/1 [==============================] - 0s 28ms/step
```

Test case 2: Result: Correct

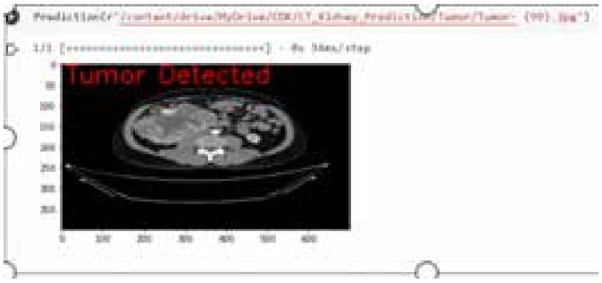
Input: Stone Expected Output: Stone Actual Output: Stone



```
[ ] Prediction(r'/content/drive/MyDrive/CDK/CT_Kidney_Prediction/Stone/Stone- (100).jpg')
1/1 [==============================] - 0s 36ms/step
```

Test case 3: Result: Correct

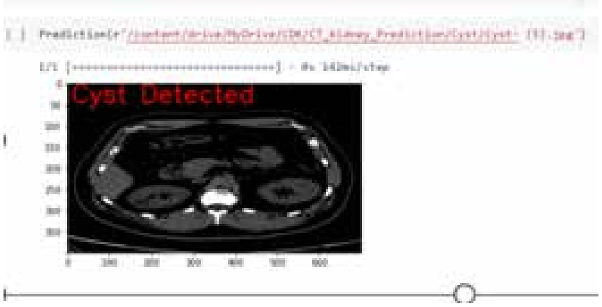
Input: Tumor Expected Output: Tumor Actual Output: Tumor



```
Prediction(r'/content/drive/MyDrive/CDK/CT_Kidney_Prediction/Tumor/Tumor- (99).jpg')
1/1 [==============================] - 0s 38ms/step
```

Test case 4: Result: Correct

Input: cyst Expected Output: cyst Actual Output: cyst

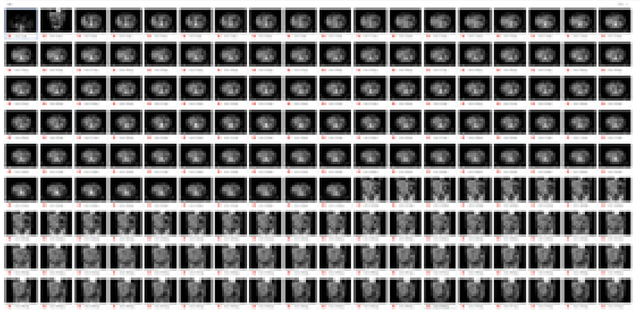


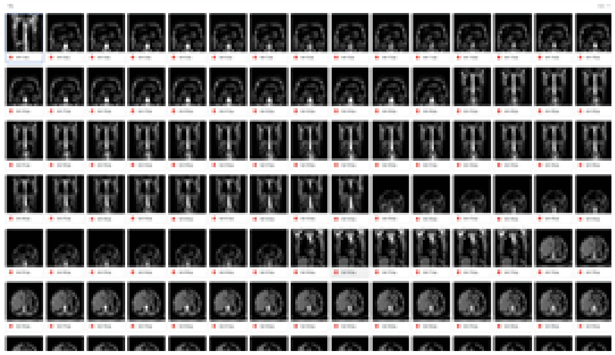
```
[ ] Prediction(r'/content/drive/MyDrive/CDK/CT_Kidney_Prediction/Cyst/Cyst- (9).jpg')
1/1 [==============================] - 0s 142ms/step
```

SCREEN SHOTS

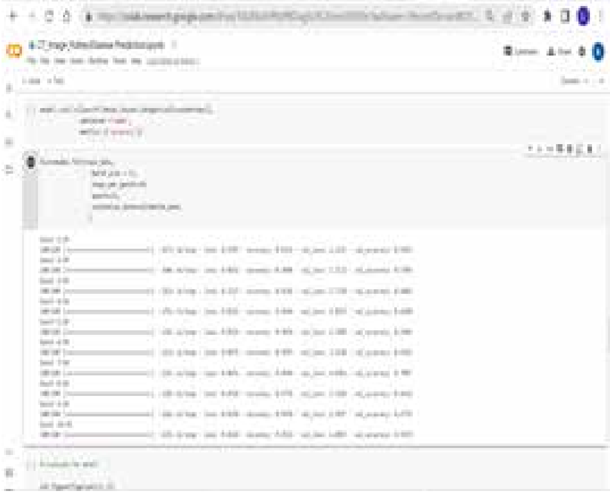
Test

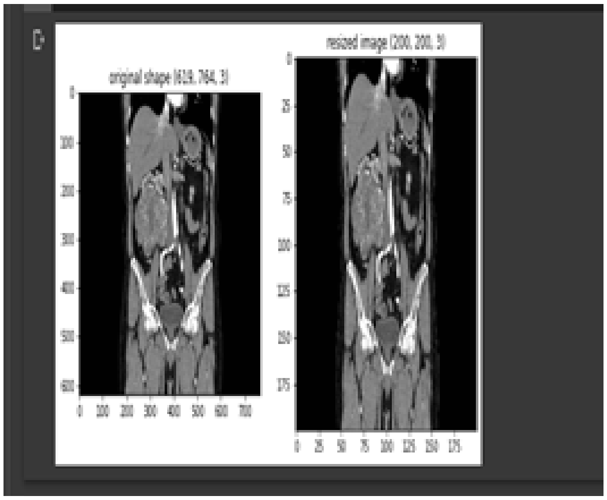
data

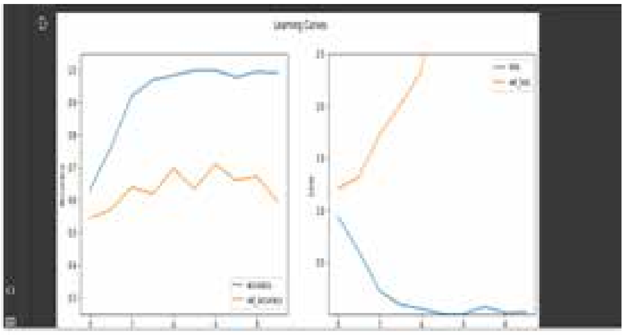




Training Result







CONCLUSION

Persistent kidney illness creates indolently, with several individuals identified late and a certain reason never developed in a significant variety of individuals. It has numerous multi-system problems, dramatically impairing the lifestyle and reducing the life span of sufferers. Therefore, the prevention and early discovery of persistent kidney disease is of utmost value. As a result, we developed a deep discovery design to find the kidney illness in early stage and we got the precision of 96 per.

ACKNOWLEDGMENT

We express our gratitude to CMR Technical Campus for providing excellent facilities and assistance to enable us to complete our work on the paper "Chronic Kidney Disease Stage Identification in HIV INFECTED PATIENTS USING Machine Learning." I truly appreciate the helpful advice and direction that our Chairman, Director, Deans, Head of the Department, Department of Computer Science and Engineering, Guide, and Teaching and Non-Teaching Faculty Members have provided for every facet of our job.

REFERENCES

1. Fadil Iqbal1, Aruna S. Pallewatte2, Janaka P. Wansapura, "Texture Analysis of Ultrasound Images of Chronic Kidney Disease", 2017 International Conference on Advances in ICT for Emerging Regions (ICTer): 299 – 303.

2. Chi Hu1 ,Xiaojun Yu1*, Qianshan Ding2 , Zeming Fan1 , Zhaohui Yuan1 ,Juan Wu1and Linbo Liu3, "Cellular-Level Structure Imaging with Micro-optical Coherence Tomography (μOCT) for Kidney Disease Diagnosis", 2019 the 4th Optoelectronics Global Conference.

3. Ahmad Amni Johari MohdHelmyAbdWahab Aida Mustapha , "Two-Class Classification: Comparative Experiments for Chronic Kidney Disease", 2019 4th International Conference on Information Systems and Computer Networks (ISCON) GLA University, Mathura, UP, India. Nov 21-22, 2019.

4. Rahul Gupta1 ,Nidhi Koli2 , Niharika Mahor3 , N Tejashri4, "Performance Analysis of Machine Learning Classifier for Predicting Chronic Kidney Disease", 2020 International Conference for Emerging Technology (INCET) Belgaum, India. Jun 5-7, 2020.

5. 1 Akash Maurya,2 Rahul Wable,3 RasikaShinde ,4 Sebin John ,5 Rahul Jadhav, 6 Dakshayani.R, "Chronic Kidney Disease Prediction and Recommendation of Suitable Diet plan by using Machine Learning", 2019 International Conference on Nascent Technologies in Engineering (ICNTE 2019).

6. Dr. Uma N Dulhare Professor, CSED, MJCET Hyderabad, India Uma.dulhare@mjcollege.ac.in Mohammad Ayesha PG Student, CSED, MJCET Hyderabad, India. "Extraction of Action Rules for Chronic Kidney Disease using Naïve Bayes Classifier", 978-1-5090-0612-0/16/$31.00 ©2016 IEEE.

7. Yedilkhan Amirgaliyev Institute of Information and Computing Technologies (IICT), Almaty, Kazakhstan amir_ed@mail.ru Shahriar Shamiluulu Faculty of Engineering and Natural Sciences, Suleyman Demirel University, Kazakhstan shahriar.shamiluulu@sdu.edu.kz Azamat Serek Faculty of Engineering and Natural Sciences, SuleymanDemirel University, Kazakhstan. "Analysis of Chronic Kidney Disease Dataset by Applying Machine Learning Methods".

8. Mubarik Ahmad, Vitri Tundjungsari, Dini Widianti, Peny Amalia, Ummi Azizah Rachmawati, "Diagnostic Decision Support System of Chronic Kidney Disease Using Support Vector Machine".

9. Sheng-Min Chiu1*, Feng-Jung Yang2 , Yi-Chung Chen3 , Chiang Lee1, "Deep learning for Etiology of Chronic Kidney Disease in Taiwan", 2nd IEEE Eurasia Conference on IOT, Communication and Engineering 2020.

10. S. Ramya and Dr. N. Radha, "Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms", International Journal of Innovative Research in Computer and Communication Engineering, Volume 4, Issue 1, January 2016, pp 813-820.

Emotion Recognition by Textual Tweets Classification using Voting Classifier LR-SGD

Jakkani Shriya, Jagtap Praveen,
Panjala Harini
B.Tech Student
Dept. of Computer Science and Engg. (Data Science)
CMR Technical Campus, Medchal
Hyderabad, Telangana

A. Veerender
Assistant Professor
Dept. of Computer Science and Engg. (Data Science)
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ veerender57@gmail.com

ABSTRACT

Due to the abundance of user-generated content on social media sites, point of view extraction has become a challenging task. In order to gather views regarding products, trends, and national politics, people use Twitter, a micro blogging network. Applying belief evaluation, a method for gauging how different individuals feel and think about a given topic, to tweets allows one to see how the public feels about particular news stories, legislation, social movements, and personalities. It is now possible to do opinion mining without manually scanning tweets by using variations of artificial intelligence. The policies, products, and events that federal governments and organisations showcase could benefit from their findings. By dividing tweets into happy and sad categories, seven ML models are used for emotion detection. The suggested ballot classifier (LR-SGD) with TF-IDF produced the best results (79% accuracy and 81% F1 score) in a thorough comparative efficiency study. In order to confirm the stability of the suggested approach on two more datasets, one with binary data and the other with multi-class data, and to achieve long-term results.

KEYWORDS: *TF, LR, ML, SGD, NLP, DL, Tweets, F1 score.*

INTRODUCTION

Computer vision, pattern recognition, and automatic emotion identification have lately become very essential in expert systems, with applications in many different fields. Twitter and other social networking sites have recently produced massive volumes of data in many formats, including structured, unstructured, and semi-structured. One recent example is the COVID-19 pandemic [1], which shows how false information spreads on social media platforms may have a significantly greater impact than a real disaster like a pandemic [2].

It is necessary to investigate in order to identify widely held beliefs with precision. Such jobs need precise natural language processing (NLP) methods and ML architectures for text categorization. Twitter allows its users to have a bird's-eye perspective of their data and examine it from all angles. Due to its audible character, effective techniques for automatically identifying message data are crucial. Twitter sentiment category has been the subject of several research in the past. [1] Since Twitter is both a fast and effective micro-blogging platform, its users are able to send short messages that are referred to as tweets. As a powerful tool in social networks, Twitter is one of the most in-demand apps in the world [2].

Twitter allows users to establish free accounts, which might lead to a large audience. Twitter is the best medium for business and marketing since it allows users to connect with famous people and other plentiful and well-known figures; this makes the acquisition of both celebrities and marketers' products much more enjoyable. Every famous person uses Twitter to connect with their fans and engage with them. Even for fans, this system is among the best options. It can compose a blog post or link on the web [4], granted that it is free and also opens as advertising, but its note range is small at 140 letters each post. Clusters of personal advertisements that seem like other social media websites aren't an issue. The time it takes to post a tweet on Twitter is little since everyone who is following that service will instantly get it.

Businesses and marketers may use this site to evaluate the many important functional perspectives. Thanks to this, they may get a quick reply from their fans. Surprisingly, many companies are increasing their partnerships with the goal of acquiring Twitter admirers. By informing followers about new services, goods, websites, blogs, books, and more, Twitter is useful for fans. Users of Twitter might then click on the link to positively contribute to produced goods, see the products offered, and earn pro_t. People can easily follow to get updates and news, businesses can tweet or re-tweet, users can choose who they want to send tweets to, how to promote their messages, and how to use it to manage their finances and other affairs. Academy, Super Bowl, and Grammy Awards, among other major sporting and home entertainment events, use it to create a lot of hype across the world [5].

There is a rise in competition across many goods on Twitter. On social media platforms like Twitter, people love to express their opinions on certain products. In order to advertise their things more effectively and generate more revenue, item owners are willing to spend more money on social media platforms [6]. A product's owner may improve the product's quality, marketing strategy, and delivery methods when customers provide feedback on the product. Customer reviews may serve as a kind of feedback for business owners or vendors. An analysis expert group is required to classify the client belief from the evaluations due to the massive amount of data collected in this manner. Machine learning and ensemble finding classifiers are necessary for precise consumer perspective classification since human error is inherent in belief analysis. In this study, we compare and contrast several types of emotion detection devices that use Tf and TF-IDF to classify tweets. This project aims to assess the performance of popular ML classifiers on Twitter datasets and offers a new classifier (LR-SGD) [8]. We compare and contrast several machine learning-based classifiers for emotion recognition using the Twitter dataset. These classifiers include support vector machines (SVMs), decision tree classifiers (DTCs), naive bayes (NBs), random forests (RFs), slope boosting machines (GBMs), and logistic regression (LRs).

A tweet-classifying voting classifier (VC) that uses LR and SGD to beat TF-IDF [10]. Using the suggested architecture on two different datasets, one with a binary component (containing courses measuring disgust or non-hatred) and another with a multi-class component (including product testimonials with ratings ranging from 1 to 5—further confirms its stability.

EXISTING SYSTEM

The findings of the sentiment evaluation developed by means of Sarlan et al. [2] classified customers' reviews expressed in tweets as either superb or bad when they extracted a huge quantity of tweets using a prototype. There have been elements to their studies. Using cutting-edge methodologies and strategies in sentiment analysis, the first element is primarily based on a literature analysis. Prior to its introduction, the utility's requirements and operations were exact inside the 2nd segment.

Alsaeedi and Zubair Khan [3] performed research that examined the consequences of numerous varieties of sentiment evaluation carried out to the Twitter dataset. Various strategies and findings on set of rules overall performance have been contrasted. Supervised ML, lexicon-based, and ensemble techniques had been used. The authors used 4 special processes, which includes supervised gadget learning for Twitter sentiment analysis and ensemble techniques for Twitter sentiment evaluation. A lexicon-primarily based method is getting used for Twitter sentiment analysis.

Numerous academics have investigated vocabulary-based totally strategies for emotion categorization. With the help of area-precise lexicon introduction, Bandhakavi et al. [4] extracted capabilities primarily based on emotions.

Disadvantages

- The existing model which is ensemble of LR and SGD is not applied on both dataset and the results.
- Voting Classifier(VC) is not a cooperative learning which engages multiple individual classifiers.

PROPOSED SYSTEM

- The dreams of the proposed system have been performed through the use of several ML methodologies. Various methodologies and techniques have been used to investigate the adaptability of the studies. When it comes to accuracy, recall, precision, and F1-score, the Voting classifier, an ensemble of Logistic Regression and Stochastic Gradient Descent, a long way outshines all other ML models that were implemented into the dataset.
- The experiment utilised a Twitter dataset that changed into withdrawn from the Kaggle repository. Datasets are pre-processed as way of getting rid of any extraneous records. The records were then divided into a schooling set and a trying out set. A percentage of 70% was allocated to the schooling set, at the same time as 30% changed into place aside for the take a look at set. The training set is then subjected to characteristic engineering strategies. Various device learning classifiers are taught on one set after which evaluated on another. This test's assessment parameters are as follows: (a) Accuracy, (b) Recall, (c) Precision, and (d) F1-rating.
- One advantage of the advised approach is that it tries to gauge how nicely famous ML classifiers perform on Twitter datasets by using presenting a voting classifier (LR-SGD).
- The hidden styles within the dataset can be higher understood with using facts visualisation. In order to have a better know-how of the dataset, it's useful to visualize the attributes' capabilities.

IMPLEMENTATION

Service Supplier

A legitimate man or woman call and password are required for the Service Provider to log in to this module. Following a successful login, the person is granted entry to to three features, such as: logging in, schooling and checking out facts units, seeing educated and examined accuracy in a bar chart, viewing trained and tested accuracy results, predicting emotions from statistics set info, locating the emotion prediction ratio on statistics units, and extra. Access all far-flung users, see the consequences of the emotion prediction ratio, and download the educated facts sets.

View and Grant Access to Users

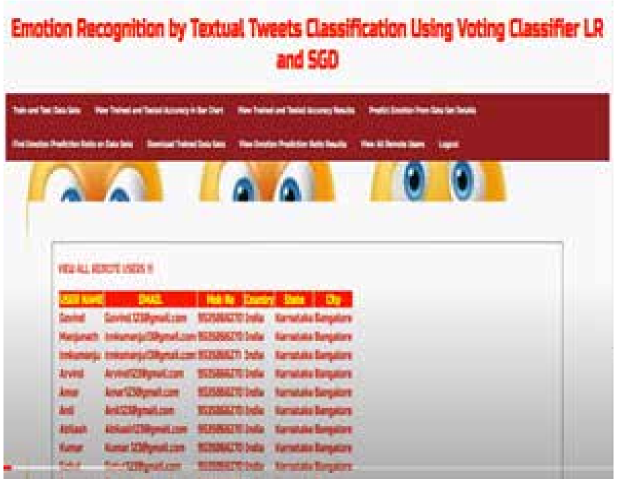
The admin may additionally see a complete listing of registered customers on this phase. Here the administrator may also see all the data of the user, such as their smartphone wide variety, email, and cope with, and that they also can authorise new customers.

User in a Remote Location

Numerous customers (n) are found in this module. Users should check in earlier than they'll do any operations. It is viable to add a person's statistics to the database after they sign in. After efficiently registering, he'll need to log in the use of the call and password of the prison man or woman. Once logged in, users may also do things like see profiles, search for and expect emotions, post tweet records units, and register and login.

Emotion Recognition by Textual Tweets Classification Using Voting Classifier LR and SGD
Login Service Provider
User Login

Fig.1. Home page



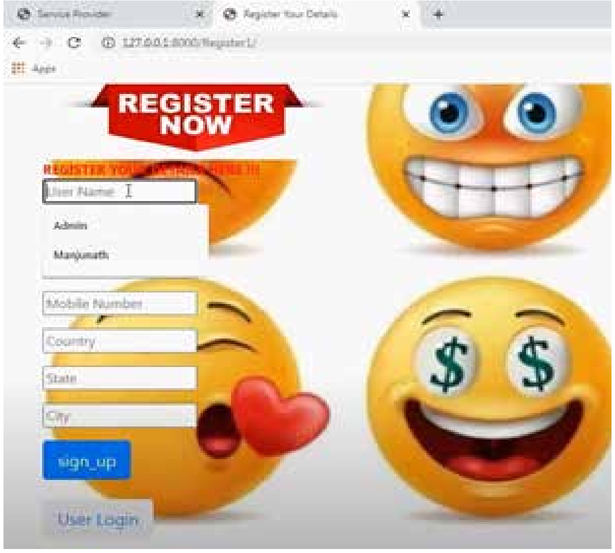
Emotion Recognition by Textual Tweets Classification Using Voting Classifier LR and SGD

Train and Test Data Sets
View Trained and Tested Accuracy in Bar Chart
View Trained and Tested Accuracy Results
Predict Emotion From Data Set Details
Find Emotion Prediction Ratio on Data Sets
Download Trained Data Sets
View Emotion Prediction Ratio Results
View All Remote Users
Logout

VIEW ALL REMOTE USERS !!!

USER NAME	EMAIL	Mob No	Country	State	City
Govind	Govind123@gmail.com	[illegible]	India	Karnataka	Bangalore
Manjunath	tmksmanju123@gmail.com	[illegible]	India	Karnataka	Bangalore
tmksmanju	tmksmanju123@gmail.com	[illegible]	India	Karnataka	Bangalore
Arvind	Arvind123@gmail.com	[illegible]	India	Karnataka	Bangalore
Amar	Amar123@gmail.com	[illegible]	India	Karnataka	Bangalore
Anil	Anil123@gmail.com	[illegible]	India	Karnataka	Bangalore
Abhiash	Abhiash123@gmail.com	[illegible]	India	Karnataka	Bangalore
Kumar	Kumar123@gmail.com	[illegible]	India	Karnataka	Bangalore
[illegible]	[illegible]123@gmail.com	[illegible]	India	Karnataka	Bangalore

Fig.2. User details login page



REGISTER YOUR DETAILS HERE !!!

User Name
Admin
Manjunath
Mobile Number
Country
State
City
sign_up
User Login

Fig.3. Registration page

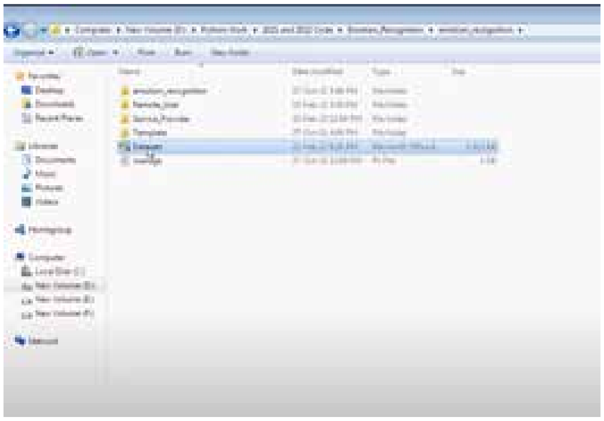
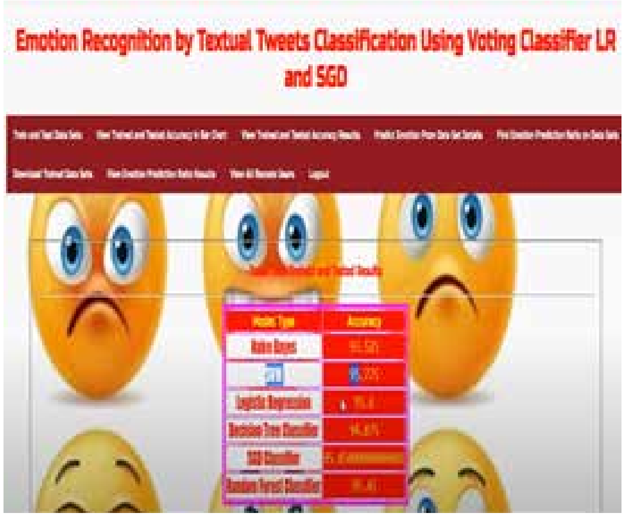


Fig.4. Dataset details



Emotion Recognition by Textual Tweets Classification Using Voting Classifier LR and SGD

Train and Test Data Sets
View Trained and Tested Accuracy in Bar Chart
View Trained and Tested Accuracy Results
Predict Emotion From Data Set Details
Find Emotion Prediction Ratio on Data Sets
Download Trained Data Sets
View Emotion Prediction Ratio Results
View All Remote Users
Logout

View Trained and Tested Results

Model Type	Accuracy
Naive Bayes	[illegible]
SVM	[illegible]
Logistic Regression	[illegible]
Decision Tree Classifier	[illegible]
SGD Classifier	[illegible]
Random Forest Classifier	[illegible]

Fig.5. Output graphs with different algorithms.

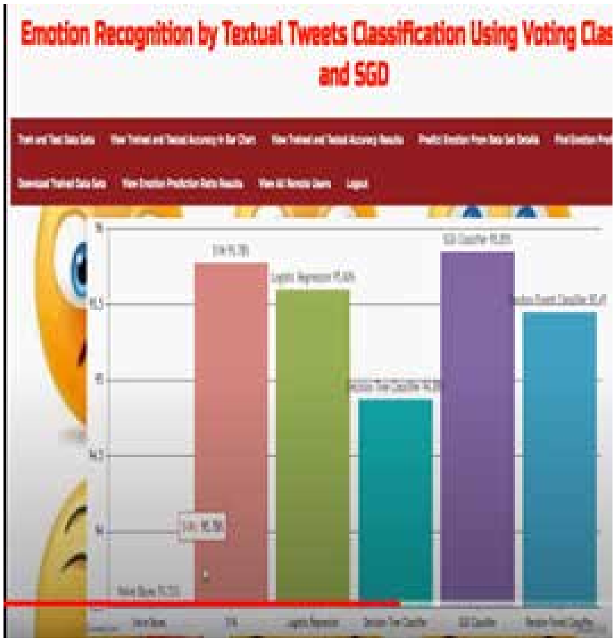


Fig.6. Output results

CONCLUSIONS

In order to perceive if a tweet's author is thrilled or sad, this research provided a brand new vote casting classifier that combines LR and SGD. Through powerful pattern recognition and model averaging, our studies proven that version performance may be more desirable. Starting with SVM, moving directly to RF, GBM, LR, DT, NB, and VC(LR-SGD), seven device studying fashions are examined via experiments. Both TF and

TF-IDF, that are function illustration methods, have been used in this paintings. Our proposed vote casting classifier, VC(LR-SGD), outperformed all models at the twitter dataset when trained the use of TF and TF-IDF. The counseled version outperforms all others whilst examined with TF-IDF, attaining an excellent 84% consideration, 79% accuracy, and an 81% F1-rating. The counseled technique has been rapidly tested on two extra datasets and has produced reliable consequences. More characteristic engineering techniques could be used in comparison and ensemble version mixtures may be explored in destiny study with the purpose of improving performance. Also, we can study some new strategies for dealing with snarky comments.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "EMOTION RECOGNITION BY TEXTUAL TWEETS CLASSIFICATION USING VOTING CLASSIFIER LR-SGD", which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head of the Department, Department of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. N. F. F. da Silva, E. R. Hruschka, and E. R. Hruschka, ``Tweet sentiment analysis with classi_er ensembles,'' Decis. Support Syst., vol. 66, pp. 170_179, Oct. 2014.
2. C. Kariya and P. Khodke, ``Twitter sentiment analysis,'' in Proc. Int. Conf. Emerg. Technol. (INCET), Jun. 2020, pp. 212_216.
3. A. Alsaeedi and M. Zubair, ``A study on sentiment analysis techniques of Twitter data,'' Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 2, pp. 361_374, 2019.
4. A. Bandhakavi, N.Wiratunga, D. Padmanabhan, and S. Massie, ``Lexicon based feature extraction for emotion text classi_cation,'' Pattern Recognit. Lett., vol. 93, pp. 133_142, Jul. 2017.
5. J. Capdevila, J. Cerquides, J. Nin, and J. Torres, ``Tweet-SCAN: An event discovery technique for geo-located tweets,'' Pattern Recognit. Lett., vol. 93, pp. 58_68, Jul. 2017.
6. T. Alsinet, J. Argelich, R. Béjar, C. Fernández, C. Mateu, and J. Planes, ``An argumentative approach for discovering relevant opinions in Twitter with probabilistic valued relationships,'' Pattern Recognit. Lett., vol. 105, pp. 191_199, Apr. 2018.
7. W. Chen, Y. Zhang, C. K. Yeo, C. T. Lau, and B. S. Lee, ``Unsupervised rumor detection based on users' behaviors using neural networks,'' Pattern Recognit. Lett., vol. 105, pp. 226_233, Apr. 2018.
8. H. Hakh, I. Aljarah, and B. Al-Shboul, ``Online social media-based sentiment analysis for us airline companies,'' in New Trends in Information Technology. Amman, Jordan: Univ. of Jordan, Apr. 2017.
9. R. Xia, C. Zong, and S. Li, ``Ensemble of feature sets and classi_cation algorithms for sentiment classi_ cation,'' Inf. Sci., vol. 181, no. 6, pp. 1138_1152, Mar. 2011.
10. M. Umer, S. Sadiq, M. Ahmad, S. Ullah, G. S. Choi, and A. Mehmood, ``A novel stacked CNN for malarial parasite detection in thin blood smear images,'' IEEE Access, vol. 8, pp. 93782_93792, 2020.

Predicting Stock Marketing Trends using Ml and Dl Algorithms via Continuous and Binary Data a Comparative Analysis

Kalva Deepika, Nayini Uday Kiran,
Ramala Santosh, Gaddam Devika
B.Tech Student
Dept. of Electronics and Communication Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

G. Srikanth
Professor and HoD
Dept. of Electronics and Communication Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ gimmadisrikanth79@gmail.com

ABSTRACT

The art of stock price prediction is both intriguing and challenging. The power economic situation is a measure of the developed nations' economic position. Because it often delivers quick profits with low-risk rates of return, the securities market is now regarded as an illustrious trading field. Many people believe that the stock market is the ideal place for data miners and business researchers due to its massive and ever-changing data resources. We used a non-linear regression technique and the k-nearest neighbour algorithm to forecast stock prices for a company's stock data so that users, administrators, decision-makers, and financiers could make well-informed investment decisions. To train the module, this algorithm takes into account the daily high, low, open, and close prices of a stock, as well as the quantities of that stock. Afterwards, in order to screen the module, an initial stock value is taken from the person and provided as a test variable. You can be sure that the component will provide you the predicted closing value of that supply. A visualisation graph drawn between the actual and expected closing values of the supply may be used to communicate the disparities between the two sets of numbers. As a consequence of the findings showing that the kNN formula is strong with a low error percentage, the results were both realistic and inexpensive. In addition, based on the data on actual stock rates, the predicted results were quite near to the actual supply rates.

KEYWORDS: *KNN, Testing stock, Stock price data.*

INTRODUCTION

Forecasting the future movements of stock prices is a challenging and time-consuming field of recent firm research study hobbies. The stock price projection of stock market movements is a topic of great interest to researchers, businesses, and interested people who believe that the future is dependent on current and historical facts (Kim, 2003). However, financial data is considered to be complicated data that requires anticipation or prediction. According to Fama's effective market hypotheses (EMH), which he put out in 1990, predicting market value is considered challenging. Market prices represent all accessible information, according to the EMH, which is seen as connecting economic data with the financial market. The EMH also confirms that rate changes are simply a result of newly available information. Supply is difficult for creators to predict, according to the EMH, but it is always stable. In addition, there is evidence that stock prices do not aim for a random walk and that additional data is required for stock prediction. In order to ascertain supply movements, data mining technology is used to examine massive amounts of business and economic data. If the current data and how they interact need to be monitored via time measurement, then temporal stock exchange mining is necessary to provide the extra capabilities. In order to predict the future values of supply, stock predictions employ a mix of basic data, pure technical data, and collected knowledge. The technical data is based on past stock performance, whilst the fundamental data represents the company's mission and the state of the market. Every unknown entity of a company's stock value may have its future value estimated using historical data by integrating data mining category approaches with supply prediction. This prediction employs a number of category strategy approaches, including k-Nearest Neighbours (kNN), decision tree induction, regression, genetic algorithms, and semantic networks. A data set is often partitioned into a training set and a screening set when using a classification strategy. In order to compare the given examination item with the training data collection, kNN uses similarity measures. An information entity is a document that represents one of n possible functionalities. For kNN to forecast a course label for an unknown document, it finds the k recodes from the training set that are the most similar to the unknown documents.

LITERATURE SURVEY

The study by Sayavong Lounnapha et al. in 2019 at the IEEE conference focused on a method for predicting stock prices using a convolutional semantic network. The extraordinary self-learning capability of convolutional neural networks is the subject of this paper's proposals for a stock price forecasting version. We display and assess the data set that links the operations of CNNs with the Thai securities market. Findings show that the Convolutional Neural Networks-based model successfully detects and predicts stock exchange rate pattern changes, which provides strong evidence for supply rate forecasting. The forecast's accuracy is determined to be enhanced, and it may also be encouraged in the finance industry.

[2] Improving Profitability with DNN-Based Supply Rate Predictions (IEEE 2019—Soheila Abrishami et al., A lot of academics have taken an interest in the prediction of financial time series because of how important it is to the industry. In order to predict the value of a portion of the stocks listed on the NASDAQ market, this article focuses on developing a deep learning system that uses a range of information. For multi-stepahead, this version—trained on very little data for a specific supply—precisely approximates the final value of that supply. It uses time series data design to distribute the inventive functions with the beginning functions and includes a vehicle encoder to eliminate noise. A Stacked LSTM Autoencoder is given these additional characteristics so that it may evaluate the supply's final value in several steps ahead of time. To make matters better, a profits maximisation strategy makes use of this appraisal to help choose when to purchase and sell a certain company. Based on the results, it seems that the suggested framework is the best of the best when it comes to logical accuracy and performance for time series projection.

The third in their 2019 IEEE paper, Ferdiansyah et al. provide an LSTM-method for predicting the price of bitcoin: a study conducted on the Yahoo Money Securities Market. One kind of investment on the stock market right now is bitcoin, which is a cryptocurrency. The stock market is vulnerable to a wide variety of dangers. Bitcoin is one kind of cryptocurrency that has been steadily rising in value over the last few years, only to have its value plummet unexpectedly on occasion, with no discernible impact on the stock market. Because bitcoin's value fluctuates, automated systems are required to predict its performance on the stock market. Methods for developing LSTM-based setting prediction bit-coin stock exchange forecasts are the focus of this investigation. Using RMSE (the Origin Mean Square Mistake), the study tries to ascertain the findings before verifying them. In every case, the RMSE will be larger than or equal to the MAE. How effectively a model can compute a continuous value is evaluated by the RMSE statistics. Methods utilised in this study to foretell Bitcoin's performance on the stock market Yahoo! Finance is able to forecast the outcome for the following number of days exceeding $12,600 USD.

In their 2019 IEEE paper, "Share Rate Forecast using Artificial Intelligence Strategy," Jeevan B. et al. A rising number of individuals from academia and business have recently shown enthusiasm for stock exchange, making it the topic of much discussion. The main focus of this study is on a method for predicting stock prices on the National Stock Market using RNN and LSTM, taking into account factors like the current market price and anonymous occurrences. This study also mentions a recommendation system that is used in selecting the firm, along with variants developed using RNN and LSTM techniques.

[5] Naadun Sirimevan et al., "Predicting Stock Exchange Performance with AI Methods," 2020 IEEE, Rates in the securities market are crucial in the current economic situation. According to studies, people's decision-making process may be influenced by social media sites like Twitter and online news. In this research, the behavioural reaction to online news is considered in order to fill the gap and improve the accuracy of the forecast. Predictions for the next day, week, and two weeks were spot on.

METHODOLOGY

We have decided to provide a user interface that allows consumers to manually choose the stock information of the company whose market value is to be projected. After that, users may use the Generate Vector alternative to make a vector representation of the data products in that dataset. The supply information is trained once the vector is created. A person may anticipate the closing value by providing the initial value. The algorithm then feeds the data into the trained component, using it as a screening variable. The ending value will be predicted by the machine learning module using the kNN algorithm applied to the supplied data set and user input. The user is then shown the anticipated value. To report the algorithm's efficiency, a visualisation chart is employed.

How Is the k-Nearest Neighbours Formula Used?

One of the simplest AI formulae based on the Overseen Understanding approach is K-Nearest Neighbour.

The K-NN algorithm sorts the new case into the category that is most similar to the existing ones based on the assumption that the new instance is similar to existing cases.

In order to classify new data points, the K-NN algorithm searches through all the existing data and uses the similarity as a criterion. In other words, the K-NN method makes it easy to classify newly-found data into an existing collection category.

Though it is most often used for Category issues, the K-NN technique has Regression and Classification capabilities as well.

K-NN does not assume anything about the underlying data as it is a non-parametric formula.

It is also known as a careless student algorithm since it keeps the dataset and performs an action on it at category time rather than immediately learning from the training set.

During training, the KNN algorithm only stores the dataset and, when fresh data becomes available, sorts it into a group that is very similar to the new data.

RESULTS EXPLANATION

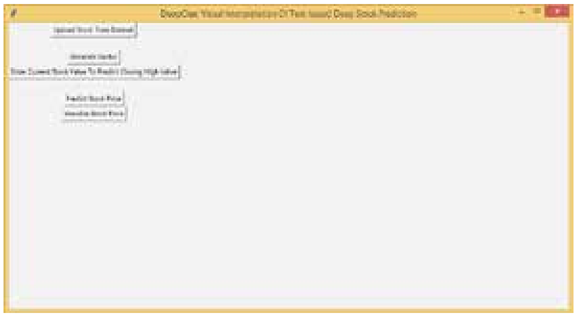

DeepClue: Visual Interpretation Of Text-based Deep Stock Prediction
Upload Stock Time Dataset
Generate Vector
Enter Current Stock Value To Predict Closing High Value
Predict Stock Price
Visualize Stock Price

Fig. 1. Admin page

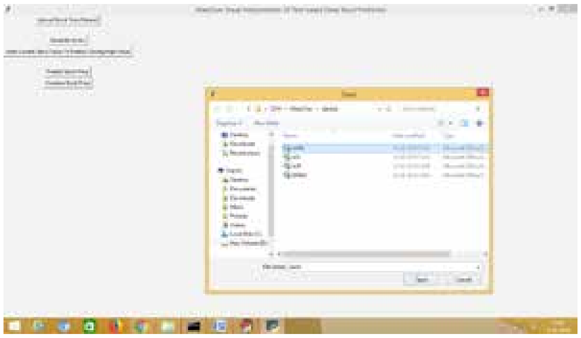


Fig. 2. Uploading Data Set

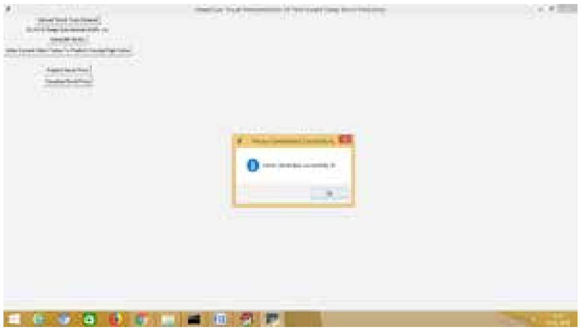


Fig. 3. Generate vector

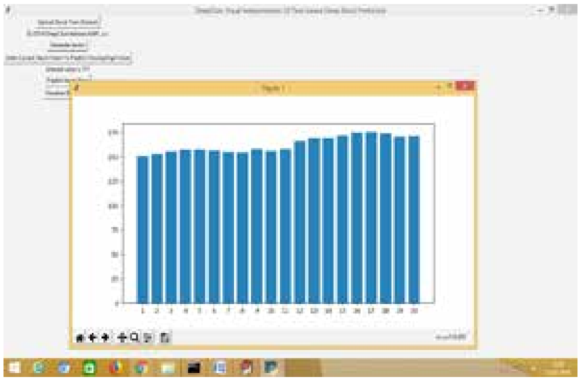


Fig. 4. Data Visualization

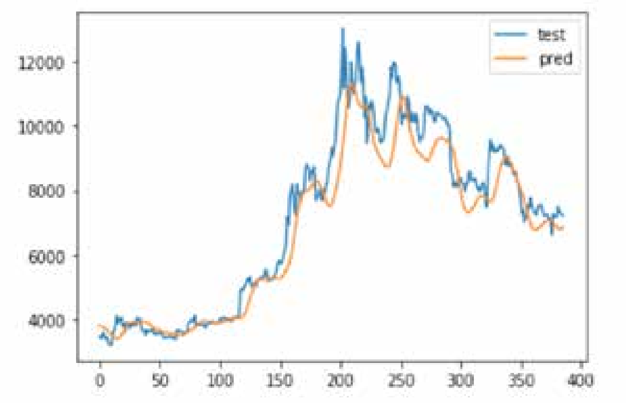


Fig. 5. Actual vs Predicted Visualization

CONCLUSION

Economic share information for markets and companies in the stock market moves at a snail's pace, making stock exchange forecasting an uphill battle. When it comes to accurate and efficient forecasting, nothing beats an expert system that employs AI techniques. The effective results were produced by the kNN-algorithm that was used in this undertaking. The findings were reasonable and applicable since the kNN formula was safe and had a low error ratio. In addition, the predicted results were rather near to the real rates, based on the knowledge about actual stock prices. Data mining techniques may aid decision makers at many levels when using kNN for data assessment, as shown by the reasonable results for predictions in particular and for using data mining approaches in real life. We conclude that kNN, in its current form, is a practical and practical tool for supply forecasting.

ACKNOWLEDGMENT

We thank CMR Technical Campus for supporting this paper titled "PREDICTING STOCK MARKETING TRENDS USING ML AND DL ALGORITHMS VIA CONTINUOUS AND BINARY DATA A COMPARATIVE ANALYSIS", which provided good facilities and support to accomplish our work. I sincerely thank our Chairman, Director, Deans, Head Of the Department, Department of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

1. Khedr, Ayman E., and Nagwa Yaseen. "Predicting stock market behavior using data mining technique and news sentiment analysis." International Journal of Intelligent Systems and Applications 9.7, pp. 22. (2017).

2. Chittineni, Suresh, et al. "A Comparative Study of CSO and PSO Trained Artificial Neural Network for Stock Market Prediction." International Conference on Computational Science, Engineering and Information Technology. Springer, Berlin, Heidelberg,pp. 186-195,2011.

3. M. Kumari and S. Godara, "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction", IJCST ISSN: 2229- 4333, vol. 2, no.2, (2011).

4. Khalid, Balar, and Naji Abdelwahab. "Big Data and Predictive Analytics: Application in Public Health Field.", International Journal of Computer Science and Information Technology & Security (IJCSITS), Vol6, No.5, 2016.

5. S.Archana and Dr. K.Elangovan, "Survey of Classification Techniques in Data Mining", International Journal of Computer Science and Mobile Applications, Vol. 2 Issue. 2, February 2014.

6. Nyce, Charles. "Predictive Analytics White Paper, sl: American Institute for Chartered Property Casualty Underwriters." Insurance Institute of America, p.1, (2007).

7. Shah, Dev, Haruna Isah, and Farhana Zulkernine. "Stock Market Analysis: A Review and Taxonomy

of Prediction Techniques." International Journal of Financial Studies, 7.2, pp. 26 (2019).

8. Alkhatib, Khalid, et al. "Stock price prediction using k-nearest neighbor (kNN) algorithm." International Journal of Business, Humanities and Technology 3.3, pp. 32-44, (2013).

9. Farshchian, Maryam, and Majid Vafaei Jahan. "Stock market prediction with hidden markov model." 2015 International Congress on Technology, Communication and Knowledge (ICTCK). IEEE, 2015.

10. BhaveshPatankar and Dr. Vijay Chavda, "A Comparative Study of Decision Tree, Naive Bayesian and k-nn Classifiers in Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, Issue 12, December 2014.

A Color-Coordinated, AI-Dried System for Identification, Trackage, and Prediction of Plant Diseases used by Farmers

Chandupatla Srinidhi, Vippakuntla Sharan Reddy, Kasula Rajesh Reddy
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

K Srujan Raju
Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ ksrujanraju@gmail.com

ABSTRACT

One of the biggest problems in farming is the prevalence of pests and diseases that attack plant leaves. A more rapid and precise method of predicting agricultural leaf diseases may lead to the creation of an early treatment strategy, hence mini missing economic losses. Thanks to recent cutting-edge advances in Deep Learning, researchers have been able to significantly enhance the efficacy and precision of systems that recognize and identify objects. Our goal in this study was to provide a deep learning-based method that could use photos of plant leaves to identify a wide variety of illnesses. Our objective is to identify and create the deep-learning approaches that are most suited to this particular job. To that end, we take into account the Single Shot Multibox Detector (SSD), the Region-based Fully Convolutional Network (R-FCN), and the Faster Region-based Convolutional Neural Network (Faster R-CNN), the latter two of which were used in this study. The suggested approach is capable of dealing with complicated situations from a plant area and may successfully identify various illnesses.

KEYWORDS: *SSD, R FCN, R CNN, Deep learning.*

INTRODUCTION

More and more, agriculture is about more than just feeding the world's expanding population. Plant energy's rising profile as a means of reducing human-caused climate change is a key consideration. Numerous plant diseases pose serious risks to human health, the environment, and the economy. Rapid and precise diagnosis of disease is of the utmost importance in this context [1].

Several methods exist for the detection of plant diseases. Some illnesses have delayed symptoms, and many individuals don't learn they have them until it's too late. In such situations, it is standard practice to use complex analytical methods, which may include the use of powerful microscopes. On other occasions, inside the invisible portion of the electromagnetic spectrum, the signals are hardly heard. One typical strategy in this kind of situation is to use remote sensing technology to examine multi- and hyperspectral picture captures. The employment of digital image processing technologies is a common way that this method achieves its aims [2].

The decline in plat leaf crop quality and quantity is mostly attributable to pests. The primary cause for the poor output of these commodities is the lack of scientific and technological expertise in pest disease prevention. Building a computer vision-based automated system to detect pest-induced illnesses in plat leaf plants is the main goal of this research. Automated disease diagnosis is the focus of this study, which employs a computer vision methodology and three distinct feature extraction techniques [1]. The 21-dimensional feature vector was constructed by using a grey level co-occurrence matrix (GLCM) and colour moments to extract textural characteristics from images of healthy and sick leaves. Using a genetic algorithm to select desirable traits and remove undesirable ones simplifies things. The end product is a feature vector with 14 dimensions. For this purpose, we make use of support vector machines (SVMs) and artificial neural networks (ANNs) [2]. Results using SVM were 92.5 percent accurate and ANN was 87.5 percent accurate after using the suggested strategy.

Currently, the only way to diagnose plant diseases is to physically examine diseased plants in great detail [3]. Particularly for big farms, the high expense is a result of the vast team of specialists required and the ongoing need to supervise them. Contrarily, many nations' farmers lack knowledge on where to get specialists and inadequate infrastructure. This is why seeking advice from consultants may be an expensive and laborious ordeal. This setting is ideal for the proposed method of crop monitoring across extensive areas. Another simple and inexpensive method for automated disease detection is to look for signs on the leaves of the plant. There may be comparable gains for machine vision systems, which employ visual signals to guide, inspect, and control robotic tasks autonomously [4].

Visual plant disease detection is not only labour-intensive and prone to error, but it also has limited practical use. Automatic detection approaches, on the other hand, can help you save time and energy without sacrificing accuracy [5], [6]. Common plant diseases include bacterial, viral, and fungal infections; early and late scorch; and brown and yellow patches. One way to find out how big the impacted region is and where the colours differ is to use image processing [7].

RELATED STUDY

Few solutions just address the detection issue since image processing approaches often provide not only illness diagnosis but also severity estimation. There are two main contexts where simple detection is used:

When attempting to identify an illness from among numerous probable pathologies, a partial classification might be beneficial instead of applying a full classification to all possible diseases. This entails classifying potential disease-causing locations as either causative or non-causative. "Neural networks" provides further information on the approach that Abdullah et al. (2007) suggested.

In real-time monitoring, the system keeps an eye out for the target disease in the crops and alerts you if it detects any. This setting calls for the research of Stories et al. (2010) as well as Sena Jr et al. (2003). Both concepts are elaborated upon further down.

Artificial neural networks

Abdullah et al. (2007) presented a technique to differentiate correspond from other leaf diseases affecting rubber trees. Segmentation is not used by the algorithm. Principal Component Analysis may also be used on a low-resolution (15×15 pixel) picture of the leaves and applied directly to the RGB pixel values, which is another alternative [10] . Finally, the output of a one-hidden-layer Multi-layer Perceptron (MLP) Neural Network, trained using the first two components, indicates whether the target illness is present or not in the sample.

Protected area

In 2003, Sena Jr. et al. suggested utilizing digital photographs to distinguish between healthy and autumn army worm-infested maize plants. They split the meat of the software into two sections: picture processing and analysis. Processing includes grayscale image processing, threshold application, and spurious artefact removal. The picture is examined in twelve distinct regions throughout the analysis process. The blocks are eliminated if their leaf area is less than 5% of their overall area. We maintain count of the number of linked objects for each block that is unaffected. Plants are deemed unhealthy if this number surpasses 10, according to empirical research [8].

METHODOLOGY

The majority of a country's GDP comes from agricultural products. A nation's agricultural production and economy take a hit when plant diseases strike. This study presents a system that can classify and identify diseases in plant leaves using deep learning techniques. We obtained the images from the Plant Village online resource. Because of their abundance in Iraq and around the world, we have chosen to concentrate on employing particular plant species in our research. These species include potatoes, peppers, and tomatoes. This collection contains 20636 images of plants and diseases. For the purpose of disease classification in plant leaves, our proposed method made use of a convolutional neural network (CNN). Twelve categories for plant illnesses (e.g., fungus, bacterium, etc.) and three healthy leaves were discovered by CNN. As a result, we attained outstanding accuracy throughout training and testing, with 98.29% accuracy during training and 98.029 percent accuracy during testing across all datasets.

RESULTS EXPLANATION

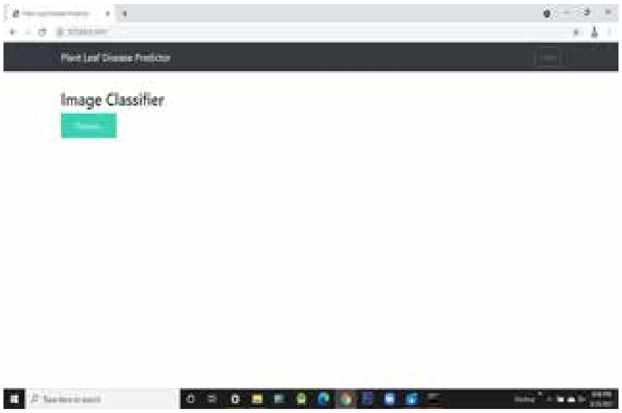


Fig. 1. OUTPUT results

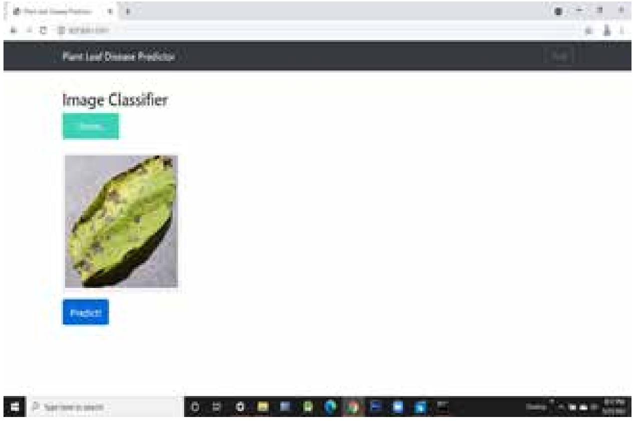
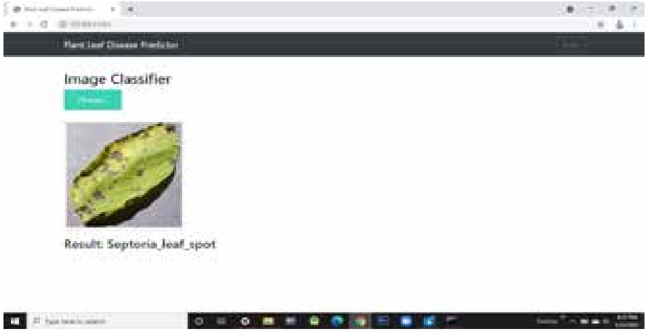


Fig. 2. INPUT image



Result: Septoria_leaf_spot

Fig. 3. OUTPUT image

Many different illnesses may affect plants. Temperature and humidity fluctuations, nutritional excess or deficit, lack of light, and the most prevalent diseases—infections produced by fungus, viruses, and bacteria—are just a few of the many things that may go wrong with plants. Depending on the illness, the leaves of affected plants may display a wide variety of shapes, colours, and textures. The aforementioned changes are hard to see because of their similar patterns, but if caught early and treated properly, the plant may save a tonne of money. Using world-class detectors like Single Shot Multibox Detector (SSD), Region-based Fully Convolutional Networks (R-FCN), and Faster R-CNN, our system can detect and classify plant leaf diseases. Not only does our system struggle with disease identification, but it also attempts to ascertain the infection state of the disease in the leaves and provides a solution—specifically, the names of appropriate organic fertilizers—for those maladies.

Plat leaf diseases may be detected automatically using the method described in this article. This system uses characteristics like colour moment, GLCM, and IN THE END AREA. Using a genetic approach to choose the retrieved characteristics leads to computational complexity and low dimensional.

CONCLUSION

In order to achieve better accuracy with less processing time, the segmentation is done using the k-means clustering technique. The proposed method additionally evaluates convolutional neural network (CNN) and support vector machine (SVM) classifiers, revealing that CNN outperforms SVM in illness detection with a 96.7% accuracy rate compared to 92.5%. Feature extraction is followed by the selection of the 14 most essential qualities using a genetic algorithm. We classify the photos using two different kinds of classifiers, one for images with sick areas (such as brown patches or plat leaf burst) and one for images without. Uses certain characteristics to evaluate CNN and SVM classifier performance. Classification using CNN achieved a maximum detection accuracy of 96.75%, while SVM achieved an accuracy of 88.6%. A thirteen-dimensional GLCM, this feature stands for the image's grey level co-occurrence matrix. The feature execution time is 0.002287 seconds, and the accuracy after training with CNN is 95.5% and with SVM it is 87.25%. But when trained with CNN, the area feature—which represents the diseased part of the leaf morphology—obtains 96.5% accuracy, and when taught using SVM, it achieves 95.25 percent accuracy. For every returned attribute, CNN outperforms SVM in terms of detection accuracy. It is clear from the statistics that geographical characteristics also have a little role in disease detection.

REFERENCES

1. Jiang Lu, Jie Hu, Guannan Zhao, Fenghua Mei, Changshui Zhang, An in-field automatic wheat disease diagnosis system, Computers and Electronics in Agriculture 142 (2017) 369–379.
2. Andreas Kamilaris, Francesc X. Prenafeta-Boldu Deep learning in agriculture: A survey, Computers and Electronics in Agriculture 147 (2018) 70–90.
3. Konstantinos P. Ferentinos, Deep learning models for plant disease detection and diagnosis Computers and Electronics in Agriculture 145 (2018) 311–318.
4. Kulkarni Anand H, Ashwin Patil RK. Applying image processing technique to detect plant diseases. Int J Mod Eng Res 2012;2(5):3661–4.
5. Bashir Sabah, Sharma Navdeep. Remote area plant disease detection using image processing. IOSR J Electron Commun Eng 2012;2(6):31–4. ISSN: 2278-2834.
6. Rakesh Kaundal, Amar S Kapoor and Gajendra PS Raghava “Machine learning technique in disease forecasting: a case study on rice blast prediction,” BMC Bioinformatics, 2006.
7. Srdjan Sladojevic, Marko Arsenovic, Andras Anderla, Dubravko Culibrk, and Darko Stefanovic, Deep Neural Networks Based Recognition of Plant Diseases by Leaf Image Classification, Hindawi Publishing Corporation, Computational Intelligence and Neuroscience Volume 2016, Article ID 3289801, 11 pages http://dx.doi.org/10.1155/2016/328980
8. J. Howse, Open CV Computer Vision with Python, Packt Publishing, Birmingham, UK, 2013.
9. D. M. Hawkins, “The problem of over-fitting,” Journal of Chemical information and Computer Sciences, vol. 44, no. 1, pp. 1–12, 2004.
10. C. C. Stearns and K. Kannappan, “Method for 2-D affine transformation of images,” US Patent No. 5,475,803, 1995.

Software-Defined Network Traffic Classification for Service Quality

Kadari Sai Chandu, Korutla Mounica,
Mahadev Srinath Goud
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

Chenna V S Manoj Kumar
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ Manojkumar.cse@cmrtc.ac.in

ABSTRACT

The accurate categorization of traffic is crucial for a variety of network operations, including providing operators with valuable forecasts for long-term provisioning, accounting, quality of service, and security monitoring. The first stage in identifying and categorising the different kinds of apps that operate on a network is to categorise network traffic. Internet service providers and network operators could find this to be a helpful tool for managing their networks as a whole. Using machine learning algorithms, we classify traffic according to its intended use. It is feasible to do this by extracting the characteristics of the traffic. We can utilise this organised data to filter out unwanted traffic and let in just the user-requested data. With the information gleaned from the categorization, we can basically set priorities for the network traffic. For the sake of reducing network traffic and improving service quality, we are making an effort to limit functionality linked to OTT services. We want to block access from OTT services like Netflix, Amazon Video, and others. The quality of the requested apps' services could increase as a result.

KEYWORDS: *ML, DL, Traffic classification, Accuracy.*

INTRODUCTION

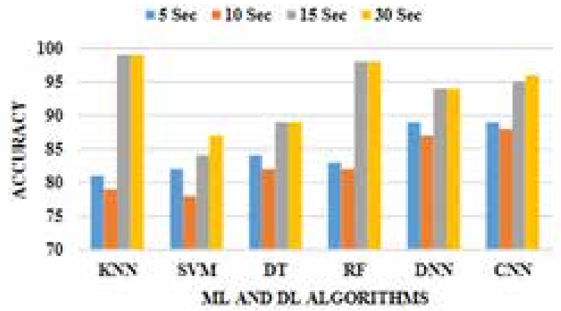
Information service providers and network operators are both intrigued by the many approaches used to classify network traffic. By classifying data flows and associating them with the applications that generate them, their management and comprehension may be improved [1]. There are a number of important uses for this data, including security, application behaviour analysis, network monitoring, and improving Quality of Service (QoS) [2]. The term "software-defined networking" (SDN) describes a specific sort of networking in which application programming interfaces (APIs) or controllers embedded in software interact with the underlying hardware to control and manage network traffic [3]. This design differs from conventional networks, which rely on dedicated hardware for traffic regulation (e.g., switches and routers). A software-defined network (SDN) may control physical hardware or build and oversee virtual networks, depending on the situation. By using network virtualization, companies may keep utilising software-defined networking (SDN) to manage data packet routing on a single server or to combine devices from several physical networks into one virtual one. The software-based nature of software-defined networking (SDN) makes its control plane much more adaptable than that of conventional networking. Admins may administer the network, change configuration settings, distribute resources, and increase network capacity from a centralised interface all without installing new hardware [4]. Classification is an important tactic for traffic treatment as it provides a foundation of information for determining the performance levels needed by applications. The two most common methods for classifying traffic are Deep Packet Inspection (DPI) and port-based categorization [5]. Several popular applications, like encrypted communication and dynamic ports, are making traditional systems obsolete [1]. Machine learning (ML) is a different way to classify traffic that uses statistical characteristics of network traffic to solve the fundamental problems with DPI and port-based classification for encrypted flows. When it comes to managing infrastructure and ensuring service quality, many applications depend on network traffic categorization. Improved reliability and accuracy in resource allocation is a direct result of thorough traffic categorization processes that provide efficient use of existing network resources [2].

RECOMMENDED FRAMEWORK

To address the issues with conventional networks, a new paradigm called software-defined networking (SDN) has emerged. According to SDN, several types of network traffic may benefit from improved QoS if the control and data planes were separated. We suggest an approach that combines application awareness with traffic engineering's Deep Packet Inspection (DPI), as SDN traffic engineering can sufficiently handle the need. Application and service flows may be distinguished by the system using port numbers and traffic categorization based on DPI. In order to enhance the quality of service, the system partitions the observed flows into several, independently prioritised queues at each port of the switch. In order to link a specific flow with the system's QoS priority (queue), a mapping table may be constructed by the network administrator. By developing an SDN controller app and data plane entities, we demonstrated the system's viability via design and implementation. For the studied application traffic, the experiment demonstrated an improvement in throughput and packet latency.

IMPLEMENTATION

In order to gather and process incoming data, the deep learning classifier is placed on the network's outer edge, as previously stated. In reactive mode, an OpenFlow network will usually examine its flow tables the moment a packet hits the network. If the match detection process returns no results, the switch will contact the controller to determine the incoming packet's next move. Apps that generate traffic and the resources they need are not well documented for the controller. In the first stage of the process, the original packet is copied and sent to the classifier. Upon arrival, the packages will be sorted into predefined categories. You may see the matched class result in the IP packet header's Type of Service (ToS) field. The next step is to send the data packet on to the control system. After the classifier sends a packet, the controller checks the header for a ToS value and compares it to its stored rules. Then, according to the configuration of the queue, it processes the packets by adding them to the queue. The controller instructs the switches on how to treat the originating packet, ensuring that all subsequent packets from it are processed in the same manner. After that, the routing tables on the switches will be updated accordingly. The OpenFlow switch is where all the configuration and setup has to be done, since OpenFlow does not come with its own queue settings. By using the OpenFlow switch, we can choose the packet scheduler type and assign priorities to flows. The packet scheduler ensures the required communication by controlling the attachment and detachment of packets for queues and allows the prioritisation of flows using various priority values for each queue. As a result, flows with higher priorities are implemented first. With the help of the HTB packet scheduler, you may postpone packets until they reach the required speeds for every queue and determine the range of available bandwidth. Allocating bandwidth to each queue and configuring the bandwidth capacity allows you to prevent bandwidth depletion and ensure that all application types have enough. To control the capacity of the network, packet schedulers enqueue flows to appropriate queues. After that, traffic shaping controls the volume and velocity of incoming flows to prevent congestion. With traffic shaping, network performance is enhanced, bandwidth availability is increased, and unexpected spikes in bandwidth demand are prevented.



The system's Quality of Service (QoS) module primarily consists of three parts: enquiring, packet scheduling, and traffic shaping. Enqueueing is a controller component that manages the mapping of flows to relevant queues and the management of messages included in OpenFlow flow tables. Both the Packet Scheduler and the Traffic Shaping are located within the switch; the former is responsible for controlling the packets in queues and the latter for regulating the bandwidth volume in queues.

For machine learning, we use the Scikit-learn package, and for deep learning, we use the PyTorch library. Variations in flow rates of 5, 10, 15, and 30 seconds were considered throughout the experiment. We have also tried many flow timeouts to demonstrate how the flow timeout affects the final outcomes.

CONCLUSION

Using deep learning models to classify network traffic according to time-based flow characteristics and apply quality-of-service, this research suggests a traffic engineering system for SDN that can better distribute bandwidth across various applications. In order to improve the service quality of the traffic flows, the system partitions it into many queues with different priorities. To improve the network's efficiency and avoid traffic jams, we use a strategy for general load balancing. Compared to machine learning models, CNN and DNN attained an accuracy of over 88% in the experiments conducted with 5s and 10s flow timeouts. Conversely, compared to CNN and DNN's 94% accuracy, KNN and RF attained above 98% accuracy for the 15s and 30s timeout. When compared to other ML approaches, DL's ability to scale up, better performance, and lack of feature engineering are its main advantages. By implementing load balancing, we can boost the throughput of each queue, apply traffic shaping to the detected flow, and ensure bandwidth for all application types. The results are derived from the system-wide analysis.

REFERENCES

1. H. Shi, H. Li, D. Zhang, C. Cheng, W. Wu, Efficient and robust feature extraction and selection for traffic classification, Comput. Netw. 119 (2017) 1–16.
2. M. Finsterbusch, C. Richter, E. Rocha, J.-A. Muller, K. Hanssgen, A survey of payload-based traffic classification approaches, IEEE Commun. Surv. Tut. 16 (2) (2014) 1135–1156 doi: https://doi.org/10.1016/j.media.2020.101813.
3. Meenaxi M Raikara, Meena S Mb, Mohammed Moin Mullac, Nagashree S Shetty, Meghana Karanandie,Data Traffic Classification in Software Defined Networks (SDN) using supervised-learning,Science Direct 171 (2020) 2750–2759.
4. Yoshinobu Yamada, Ryoichi Shinkuma, Takehiro Sato, and Eiji Oki Graduate School of Informatics, Kyoto University Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan, Feature-selection based data prioritization in mobile traffic prediction using machine learning , 978-1- 5386-4727-1/18/$31.00 ©2018 IEEE.
5. Amirhossein Moravejosharieh ,Kourosh Ahmadi ,Saghir Ahmad , A Fuzzy Logic Approach To Increase Quality of Service in Software Defined Networking, Communication Control and Networking (ICCC 2018).
6. Gagangeet Singh Aujla, Rajat Chaudhary, Neeraj Kumar, An Ensembled Scheme for QoS-aware Traffic Flow Management in Software Defined Networks, 978-1-5386- 3180-5/18/$31.00 ©2018 IEEE.
7. Thomas Favale, Francesca Soroa , Martino Trevisana, Idilio Drago b, Marco Mellia, Campus traffic and e-Learning during COVID-19 pandemic.Computer Networks 176 (2020) 107290.
8. Xiaoling Tao, Yang Peng, Feng Zhao, Changsong Yang, Baohua Qiang, Yufeng Wang, Zuobin Xiong, Gated recurrent unit-based parallel network traffic anomaly detection using subagging ensembles. Ad Hoc Networks 116 (2021) 102465.
9. Xiaoshi Fana, Yanan Wang, Mengyu Zhang. Network traffic forecasting model based on long-term intuitionistic fuzzy time series Information Sciences 506 (2020) 131–147.

The use of Deep Neural Networks and Static Facial Features in the Diagnosis of Autism in Children

Bellari Sripriya, Donthulawar Surya Teja,
Chinnagani Santhosh Kumar
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

N Chandana
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ chandana.cse@cmrtc.ac.in

ABSTRACT

As AI has been steadily improving, facial expression detection has lately gained a lot of traction. An important part of the engagement time is the emotional recognition. Verbal communication technology relies on non-verbal clues 2.03, while spoken signals really have a 1.033 part in actual engagement. The facial emotion recognition (FER) method can detect facial expressions. Not only can people's facial expressions reveal their deepest emotions and thoughts, but they also reveal their intellectual status and human viewpoint. The objective of this study is to identify common human emotions by combining gender classification with age estimate. Along with facial expressions, elegant emotions include happiness, grief, disappointment, task, surprise, and impartiality. This paper proposes You Look Just Once (YOLO) version 2 styles as a real-time facial emotion recognition device. These styles are similar to a squeeze net design. Introducing Yolo, the ultimate real-time object detector. This project aimed to discover faces in real-time. Anchor boxes make it possible to capture these photographs with pinpoint accuracy. The 2D shape is a compressed net that may be used to assess one's age and one's attractiveness. Using essences top-level capabilities for high-quality, accurate item identification, it improves usual performance in image recognition and device localization. In terms of end result, these strategies beat a bunch of alternative methods that use massive wonder layers and transit via reputation in the neural network.

KEYWORDS: *YOLO, 2d, FER, Locating devices.*

INTRODUCTION

The initial step in face recognition is sorting photographs according to whether they include faces (desires) or other unnecessary details (clutter) that need to be removed. Facial features are consistent throughout age groups, even if skin tone, texture, and colour might vary substantially [1]. The additional difficulty is heightened by factors such as partial occlusion, concealment, image trends and geometries, and worries about varied lighting. Any face in any dataset should be recognisable by a good face detector, regardless of the lighting conditions [2]. The face detection test may be trained to do a wide range of tasks. As an initial step, run a categorization venture. This will take a tiny portion of the picture and provide a binary yes/no result indicating if face shapes are present or not [3]. A prime example is the face localization project, which takes a photo of you upon entry and utilises the coordinates of any object or face in the photograph to generate a set of x, y, length, and elevation bounding boxes. Automated facial applications has the ability to revolutionise intelligent robots. Crawlers like this may find a home in simulation games set around telecom hubs. Elman often makes use of six well-known expressions: horror, shock, anger, sadness, and joy. By arranging different facial features, you may choose from a range of emotions. As an example of someone enjoying, consider the expression of a smile accompanied by tightly drawn eyelids and pursed lips. Facial expressions provide a wealth of information about the characters, including their emotions, connections, and goals [4]. Among the many fields that have profited substantially from automated face recognition are emotion analysis, picture retrieval, chat crawlers, and natural human-computer interaction [5]. Recognition of faces by use of an oriented gradient pie chart People see faces as a very natural and effective way to communicate themselves, which is why the use of convolutional neural network detection has been causing problems in the technical network. At long last, the computer has arrived at face finding. Feature preference, classifier building, and function expertise are the three stages that make up the expression recognition system training process. Function understanding comes in at number four, followed by classifier construction and production at number five. Variances among the face-up attributes are all that are gained after you reach the function information level. Next, the top-notch qualities show the face via function selection. Not only are they squandering the inter-class model, but they also aim to lessen the intra-direction variability of the expressions. The difficulty in reducing the intra magnificence version of expressions stems from the fact that the various individuals in the shot have quite varied genuine expressions. Several choices exist for face recognition systems, such as YOLO, SDD, RCNN, and Faster RCNN.

LITERATURE SURVEY

Convolutional Neural Networks for Oriented Gradient Pie Chart-Based Face Expression Recognition 2.1

As if it were all members of the literary collective Adnan Khan's legal team—Fayez Ali, Sahara Afar, Iran Ali, Sub hash Guiro— For a long time, researchers have been interested in facial expression recognition because of its potential usefulness in several fields, including feature extraction and reasoning. Using the FER2013 data set—which has seven categories: Shock, Fear, Angry, Neutral, Sad, Disgust, and Delighted—this study suggests a new approach to dealing with fame. But, being the extreme intra direction version, it is still doing its hardest. To ensure the accuracy of our methodology and data, we double-checked all of the learned and homemade components that use HOG. One model, FER with deep convolutional semantics using a CNN, and another, FER with oriented slopes pie chart using a HOGCNN, are proposed in the study. The accuracy of the FER-CNN model was 98% during training and 72% during testing, with no losses in either case.being 2.02, 2.02, and 02. expressly. The opportunity problem yielded FER-HOGCNN model accuracy assessments of97%,70%, and0%, respectively, from a pedagogical standpoint.in addition to 2.04. The results: The results show that the FER-HOGCNN architecture delivers a medium level of accuracy when compared to Easy FER-CNN. Due to the dataset's very high-quality, compressed, small-scale images, the HOG misses out on certain important features while training and filtering. Advantages and uses: The findings of this study will hopefully form the basis of future investigations, and the research check will help enhance the image processing capabilities of the FER System. In addition, it will help with feature extraction from pictures by merging LBP with the HOG operator utilising Deep Learning versions.

Extracting Social Forms using Active Clustering with Ensembles (2.2)

The authors of this work are A. Cement, J. R. Barr, K. W. Bowyer, and P. J. Flynn. A brief summary: Our method allows users to retrieve their social network architecture while they are engaged in a demanding and rapid-fire film. Not only are individuals not known, but they are also not paired with verified registrations. One common way to identify a character is by counting the number of films they have been in. A node represents each unique group within the social media community. When two nodes' clustered faces line up in a large number of frames, it's considered a link. To improve the accuracy of identification clusters derived from user input on mismatched faces, our solution employs a single active clustering algorithm. It all adds up to a laundry list of community frameworks that probably includes the social agency or agencies, as well as a list of parents with connections to a plethora of social groups. The results show that the community evaluation methods and the proposed clustering methodology are both effective.

A set of pie charts with sloped slopes for quick people identification

In order to create a human discovery machine that is both fast and accurate, we combine the cascade-of-rejecters method with the Hog subsystem. Everyone from S. Avidan and K.-T. Cheng to M.-C. Yen and Q.

Zhu wrote a word in this piece. The finished products of our system are hogs of varying lengths that can take over massive human tasks on their own will. From among an overwhelming amount of valuable blocks, we find the right set using Gadabouts for feature selection. Our system can do computations much more quickly since it uses a rejection waterfall in conjunction with the fundamental picture representation. Depending on the photo's thickness, this technique can improve 320 × 280 images at a pace of five to thirty frames per second, while maintaining an accuracy level similar to existing approaches.

Part2: Directional ternary samples for area facial identification.

Ryun, J. Kim, A. R. Rivera, and O. Chafe make up the author team. A brief summary: We present LDTP, a cutting-edge face descriptor for facial recognition, in this study. By using LDTP's directional data and ternary pattern, we can effectively capture emotion-related elements such the eyes, eyebrows, top nose, and lips. In simple areas, we can even erase the drawbacks of the detail-based approach. This increases the longevity of the element's factor patterns. Our method, which utilises a -degree grid to generate the face descriptor while sampling expression-associated data at different degrees, is surpassed by modern histogram-based completely accurate face summary algorithms that evenly pattern the codes and divide the face into regions. We use two different grids: one coarse for energetic codes (very expression-related) and one finer for regular codes (very non-expression-related). We are able to accomplish a more thorough précis of facial sports and determine the expression's fundamental skills via this multi-stage approach. Furthermore, the emotion-related parts of the face provide the interactive LDTP codes. We tested the generalizability of our strategy using both individual-based and neutral skip-validation frameworks. By using our methods with six datasets, we demonstrate that the prevalence of facial capabilities may be more precisely anticipated.

METHODOLOGY

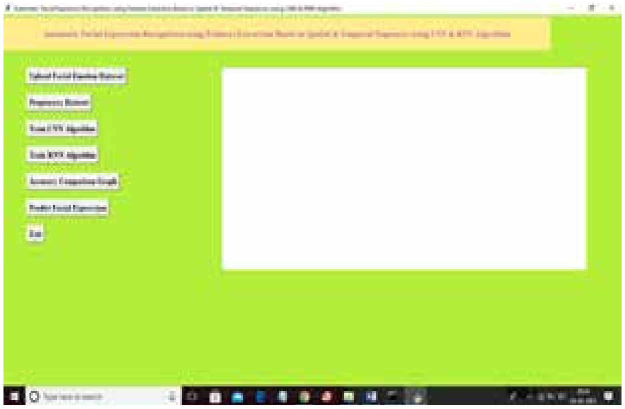
Existing System

In this day of constant communication, the magic of emotions does wonders. When it comes to interaction age, words are king.Just as non-verbal cues play a zero.33 in communication, so does the number 33 of spoken change. If you want to know how someone is feeling, you may utilise the facial emotion credibility (FER) method. An individual's facial expressions are a great way to convey not just their emotional state but also their inner thoughts and feelings, as well as their intellectual background and perspective.

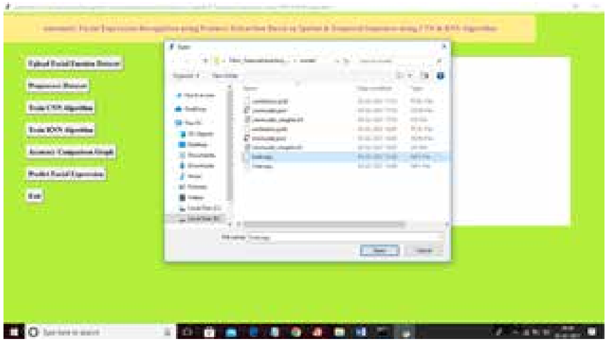
System Proposed

The purpose of this article is to discover the most important human emotions by combining gender attractiveness with age evaluation. Because emotions are thought of as the most important feeling, the facial feelings include joy, sadness, wrath, concern, and amazement. Here is a recommendation for a real-time facial expression online reputation builder that uses the You Look Pretty Now (YOLO) model 2 style and a squeeze net form. One such real-time item finding tool is the Yolo framework. Here it is used to become aware of and also to discover faces in real time. The employment of helpful assistance boxes allows for the precise recording of these damage photographs. Sex attractiveness and age evaluation are two other uses for the 2D version, which is a compressive net. It does a great job of identifying the right topics and extracting high-level talents that aid in achieving great common average performance in image categorization and gadget identification. When compared to other methods that use a high number of hidden layers to avoid validation in the brain region, both of these types provide much more accurate stop results.

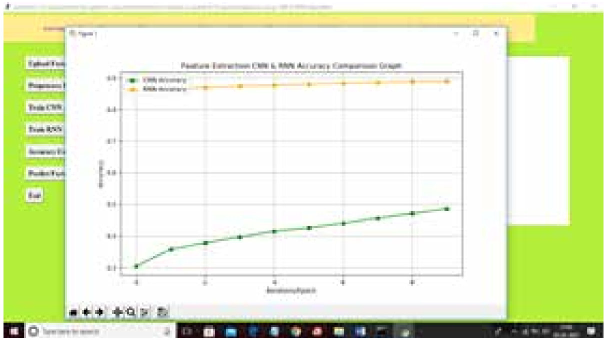
Access the 'run.bat' file in the title1 folder from the previous screen to bring up the one below.


Upload Facial Emotion Dataset
Preprocess Dataset
Train CNN Algorithm
Train RNN Algorithm
Accuracy Comparison Graph
Predict Facial Expression
Exit

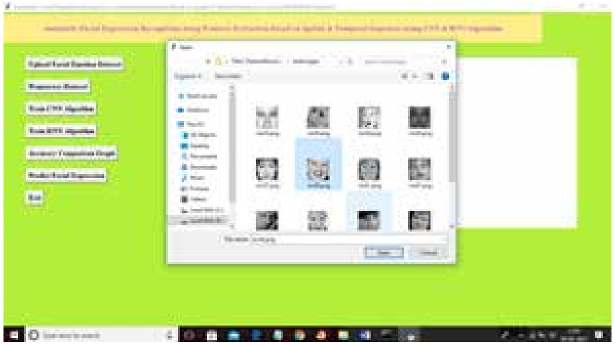
You may upload a dataset by clicking the "Upload Facial Emotion Dataset" button on the previous screen.



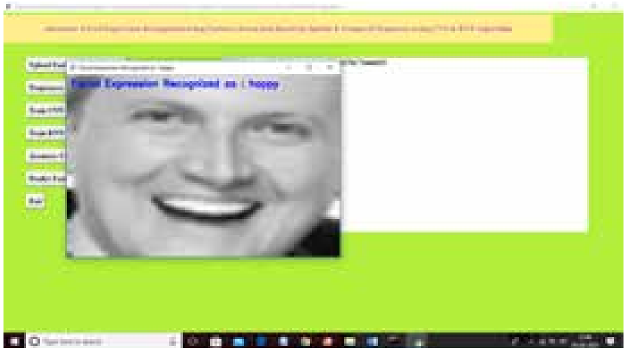
To load the dataset, choose the 'X.txt.npy' file that contains images of all emotion faces from the previous page, then click the 'Open' button. The dataset will then be shown on the following screen.



The number of iterations or epochs is shown on the x-axis and the accuracy is shown on the y-axis in the screen above. The graph shows the accuracy of the RNN as an orange line and the CNN as a green line. Both algorithms become better with each iteration, but RNN is obviously more effective, as shown in the graph. Click the "Predict Facial Expression" button to ask the system to estimate an emotion based on a fresh test photo.



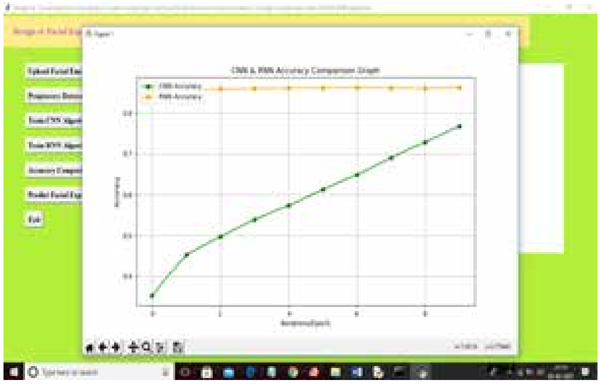
In the above screen select and upload the im38.png image and then click on the 'Open' button to get the below result



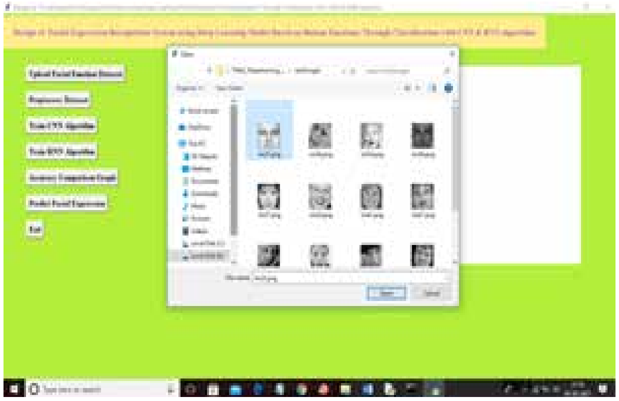
Facial Expression Recognized as : happy

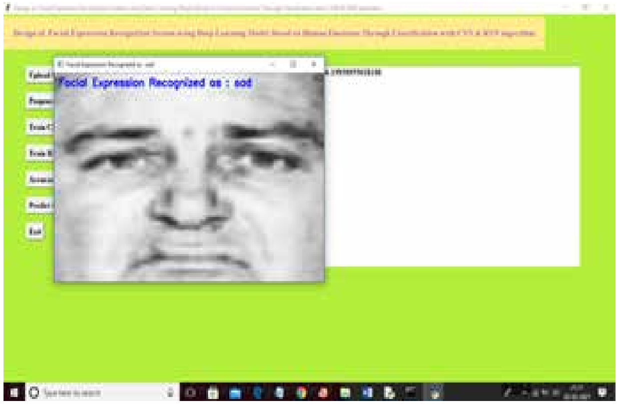
On the previous page, it identified "happy" and other emotions; on this screen, you may input any picture and it will predict a mood. This is the result of running TITLE 1.

Activate the Title 2 project by navigating to the 'Title2_Deeplearning_CNN_RNN' folder and executing the 'run.bat' file. The outcome is the screen that you can see below.

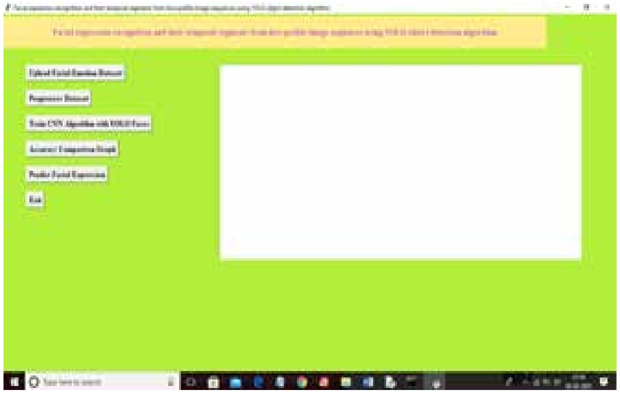


Next, submit the test photograph by clicking the "Predict Facial Expression" button. The programme will then attempt to forecast the user's emotional state based on that image.

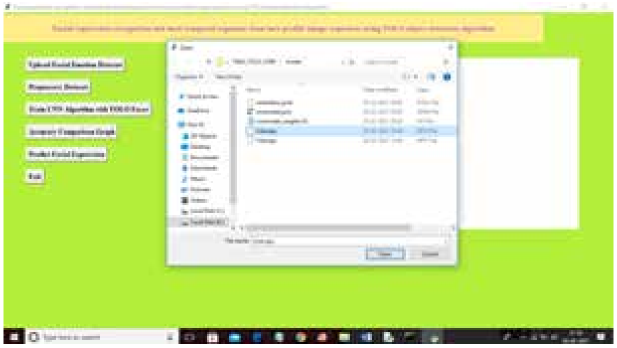

Upload Facial Emotion Dataset
Preprocess Dataset
Train CNN Algorithm
Train RNN Algorithm
Accuracy Comparison Graph
Predict Facial Expression
Exit

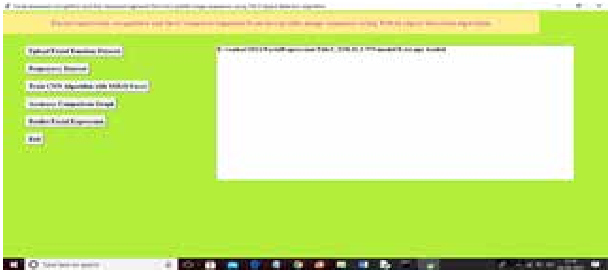

Facial Expression Recognized as : sad

'Sad' is the emotion identified in the screen above. To access the next screen, open the 'Title3_YOLO_CNN' folder and locate the 'run.bat' file. Then, execute title3.


Upload Facial Emotion Dataset
Preprocess Dataset
Train CNN Algorithm with YOLO Faces
Accuracy Comparison Graph
Predict Facial Expression
Exit

Press the "Upload Facial Emotion dataset" button on the previous page to bring up the next screen after loading the dataset.


Upload Facial Emotion Dataset
Preprocess Dataset
Train CNN Algorithm with YOLO Faces
Accuracy Comparison Graph
Predict Facial Expression
Exit


Upload Facial Emotion Dataset
Preprocess Dataset
Train CNN Algorithm with YOLO Faces
Accuracy Comparison Graph
Predict Facial Expression
Exit

The dataset is loaded in the above interface, and all buttons provide access to output and information about its correctness. In a similar vein, you may get the error rate by running all modules in the title4 project after uploading the dataset.

CONCLUSION

The prevalence of makers in people's everyday lives has been steadily rising over the last several years. Devices are used by many different markets these days. They hope that the more time they spend interacting with others, the smoother and more natural their interactions will become. This can only be accomplished if the devices in question have some kind of environmental perception capability. Particularly the goals that an individual has. Emotion recognition is a basic and difficult subject in computer technology since every utterance is a combination of feelings. We provide below an eco-friendly stay face popularity gadget that integrates Yolo model 2 and squeeze internet fashion, two deep neural network-based algorithms, to enhance the accuracy and reliability of facial functions reputation. The future variation may be accomplished after the sensations are acknowledged. The system may play music, tell a story about a fluffy puppy, or even contact a buddy if it detects that he or she is feeling bad. When AI is able to experience what the character is going through emotionally and react accordingly, it has progressed to the next level. This bridges the gap between the human and robotic worlds. In addition, we are considering including an interactive keyboard into the programme. When users interact with it, the software may determine their mood and change it to their favourite smiley.

REFERENCES

1. C. A. Corneanu, M. O. Simon,´ J. F. Cohn, and S. E. Guerrero, “Survey on rgb, 3d, thermal, and multimodal,” IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 8, pp. 1548–1568, 2018.
2. P. I. Wilson and J. Fernandez, “Facial feature detection using haar classifiers,” Journal of Computing Sciences in Colleges, vol. 21, no. 4, pp. 127–133, 2017.
3. Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, “Fast human detection using a cascade of histograms of oriented gradients,” in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06), vol. 2. IEEE, 2016.
4. Ryu, A. R. Rivera, J. Kim, and O. Chae, “Local directional ternary pattern for facial expression recognition,” IEEE Transactions on Image Processing, vol. 26, no. 12, pp. 6006–6018, 2017.
5. B. Yang, J. Cao, R. Ni, and Y. Zhang, “Facial expression recognition using weighted mix-ture deep neural network based on double-channel facial images,” IEEE Access, vol. 6, pp. 4630–4640, 2017.
6. S. Xie and H. Hu, “Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks,” IEEE Transactions on Multimedia, vol. 21, no. 1, pp. 211–220, 2016.
7. K. Zhang, Y. Huang, Y. Du, and L. Wang, “Facial expression recognition based on deep evolutional spatial-temporal networks,” IEEE Transactions on Image Processing, vol. 26, no. 9.
8. Y. Liu, Y. Cao, Y. Li, M. Liu, R. Song, Y. Wang, Z. Xu, and X. Ma, “Facial expression recognition with pca and lbp features extracting from active facial patches,” in 2016 IEEE International Conference on Realtime Computing and Robotics (RCAR). IEEE, 2016, pp. 368–373.
9. H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, “Joint fine-tuning in deep neural networks for facial expression recognition,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 2983–299.

The use of X-Rays in the Diagnosis of Bone Tumours

Karla Raja Sulakshana, Chandolu Mohana Lahari, Byrraju Peddi Raju
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

S V Suji Aparna
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ Aparna.cse@cmrtc.ac.in

ABSTRACT

Bone sarcoma, sometimes called bone cancer cells, is a rare kind of cancer that manifests as an abnormal proliferation of bone tissue with metastasis potential. Children and teenagers are often the targets of its effects. Contrary to every other form of cancer cell—breast, lung, prostate, belly, brain, etc. The causes of bone cancer cells remain unknown. Thus, a patient's chances of surviving a bone sarcoma might be improved with even a basic early diagnosis. It is possible to improve the accuracy of the process of identifying bone lumps by combining image processing methods with medical imaging modalities (such as X-ray, MRI, and CT imaging). In this study, we presented a novel approach to medical sarcoma diagnosis: the Generalised Gaussian Density analysis (GGD). Beginning with the processed bone MRI, size-specific sub-images are generated and evaluated using GGD. The next step is to extract a region of interest (ROI) from the original MRI that matches the sub-images that have the maximum value for the shape specification.

KEYWORDS: *MRI, GGD, CT, ROI, X-ray, Data set.*

INTRODUCTION

Bone cancer cells are those that multiply abnormally inside the bone. I don't know how important it is. The development of primary bone sarcomas occurs inside bone cells, as opposed to secondary sarcomas, which start in other tissues [1]. Pain, brittle bones, and high calcium levels are very typical symptoms of bone cancer. Better treatment options and a lower risk of impairment could result from earlier detection of bone cancer cells [2]. Unfortunately, radiologists often make the wrong diagnosis of bone cancer due to their difficulties in interpreting clinical pictures. Thanks to medical imaging analytic techniques made available by image processing methods, radiologists will have a better chance of properly diagnosing bone cancer. First, the article will define bones and explain how cancer cells start off in bone tissue. We then proceeded to illustrate several cell types often seen in bone cancer [3].

Photo segmentation reveals details that were previously hidden. To do this, we make use of certain regions of the picture that are located below the fold. When creating electronic vision applications, these technologies have a plethora of possible uses, including photo compression, object recognition, and limit line detection of the given item. Picture segmentation simplifies the initial image by identifying groups of pixels that have similar traits and then naming them. [1] In order for parts of a picture to be grouped or tagged, such parts must consist of sets of pixels that share some attribute. Cancer cells are uncontrolled, immortal, and capable of spreading to other parts of the body. A research conducted by the National Institute of Cancer Prevention and Research (NICPR) estimates that 2.5 million individuals in India are coping with cancer. More than 7 lakh new instances of cancer are reported every year, with 556,400 people losing their lives to the disease. The International Agency for Research on Cancer (IARC) projects that in 2030, the cancer rate would reach 21.7 million new cases and 13 million fatalities.

There are now 75 different forms of bone cancer, with osteosarcoma and Ewing tumours being among the most frequent. Accelerating the identification and beginning of therapy for cancer based on disease type and stage might be one strategy to reduce mortality rates. Radiologists employ x-rays, often called radiographs, to produce pictures of the inside of an organ, allowing them to make a noninvasive diagnosis. Magnetic vibration imaging provides a far more detailed view of the same phenomenon by using powerful magnets and radio waves. Both procedures instantly produce a grayscale picture. It is possible to detect benign (not containing cancer cells) or malignant (containing cancer cells) tumours in the bone using magnetic resonance imaging (MRI) or X-rays. Size and form are two of several distinguishing features that may be used to classify bone cancer cells. The suggested method uses a combination of picture segmentation and x-ray or MRI to eliminate cancer, a major health problem. Using x-ray or MRI data, we tried to examine several image segmentation methods in this paper to get a better understanding of the aberrant bone growth. In this brief article, we will compare and contrast many picture segmentation algorithms and find out which ones work best in what situations.

The development of solutions focused on individuals has created intense competition in the field of therapeutic image processing. Malignant bone growth may be hard to identify and, if addressed, might pose serious health risks to the patient. Consequently, pinpointing a brain cancer via image analysis requires pinpoint accuracy. While X-rays are necessary for obtaining any kind of ray-based image, more costly and comprehensive imaging techniques like computed tomography (CT) and magnetic resonance imaging (MRI) provide a better look into the human body. In order to make an appropriate diagnosis of the bone utilising a 3D electronic photo framework, numerous algorithms need to be performed, since both CT and M.R.I. employ 3-D pictures of the bone structures. Digital picture analysis may give the best treatment. Making a system for electronic photo acquisition and handling is the goal of this project. Give him the green light to quickly and accurately categorise things using the data supplied by the formula. In order to identify cancer, every method must go through classification, filtering, segmentation, morphological operations, and function elimination. The formation of primary bone cancer cells may occur inside the bone itself, regardless of where the second bone dissolves in the body.

LITERATURE SURVEY

A technique for determining the typical strength and stage of cancer cells based on their increasing size was developed by Kishor Kumar Reddy C, Anisha P R, and Narasimha Prasad L V. Using an area-growing algorithm, Kishor Kumar Reddy C, Anisha P R, and Raju G V S [2] introduced a new method for assessing lump size and bone cancer stage. For the purpose of detecting various brain tumours, Dipali M. Joshi, Dr. N. K. Rana, and V. M. Misra developed the Neuro Fuzzy Classifier. For the Thermographic Image Evaluation Approach in Detecting Canine Bone Cancer, MaryamsadatAmini, Peng Liu, Scott E. Umbaugh, Dominic J. Marino, and Catherine A. Loughin proposed utilising the CVIP-FEPC programme [4]. [5] To improve the MRI image, Miss Hemangi S. Phalak and Mr. O. K. Firke proposed a preprocessing method. To potentially detect brain masses, this technique combines cellular automata side detection with modified texture based area growth. By using the k-mean clustering method, Madhuri Avula, Narasimha Prasad Lakkakula, and Murali Prasad Raja were able to detect bone cancer by adding the pixel strengths for the drawn-out tumour component to the mean intensity [6]. Muhammad Anshad PY and S.S. KUMAR investigated the benefits, drawbacks, and accuracy of current approaches to tumour identification using computer-aided medical diagnosis [9], while Rahul Kansal, Puneetgupta, Manjit Arora, Priyanka Mattoo, Arti Khurana, and Indu Bhasin reviewed an instance record to differentiate between osteosarcoma and Ewing's sarcoma [7].

Three approaches to biological picture segmentation based on fuzzy degeneration, decline, and the least square technique are described in detail by S. Vitulano, C. DiRuberto, and M. Nappi (1997) [10].

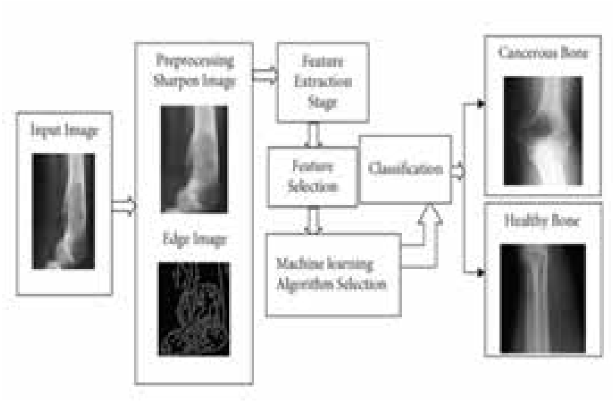
The Present Configuration

While magnetic resonance imaging (MRI) offers more contrast, x-ray and computed tomography (CT) imaging provide superior specificity and resolution. The broad use of hybrid imaging techniques is a result of their ability to combine the advantages and minimise the drawbacks of many imaging modalities. Various image processing studies have concentrated on the problem of bone tumour stage detection. has successfully detected bone cancer cells using a region-expanding technique.

By calculating the average intensity and growth size, he was also able to determine the cancer stage.

PROPOSED SYSTEM

This study trains a deep learning CNN to identify potential bone tumours using pictures of healthy and diseased bones. Without including any pre-processing processes, the suggested technique begins by dividing the MRI picture into blocks of a selected dimension. After that, we need to execute a GGD computer on every single bloc. Then, choose a ROI that matches the blocks where the form parameters are set to their maximum values.



METHODOLOGY

The thresholding method used by the department is straightforward and effective. This makes the low-quality, two-dimensional picture seem like it has three dimensions instead of just two. This also results in the same decision when using maximum values (max and minutes) to denote lumps. Interestingly, it can handle photos with really fine levels of detail with ease. As a result, we are able to get very precise pictures of bone tumours. As measured by the picture's greatest pixel value, the ideal value Support Backing for AI Applications Vector Maker (SVM), an AI tool, is built on the idea of a large side information order. Thanks to the grouping computations that were implemented on top of it, the device has great theoretical capabilities and remarkable conjecture execution.

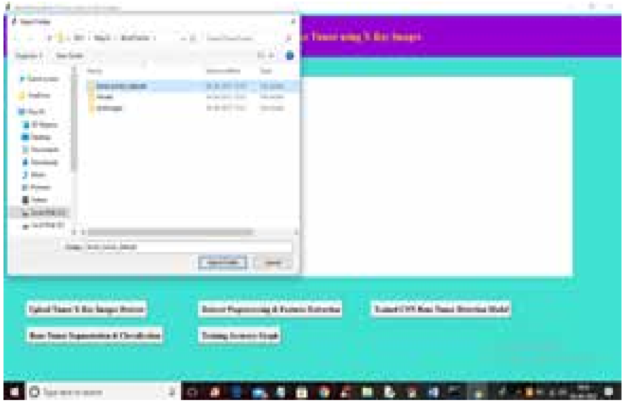
In this case, we will detect early indicators of bone cancer using artificial intelligence algorithms and the image division method. Our primary objective is to demonstrate the efficacy of magnetic resonance imaging (M.R.I.) in lump detection. Even though interference seems to be present in this specific photo, C.T. images yet maintain their quality. Because the specific location of the injured tissue has gotten so tough to discern, this also limits the quantity of room that may be used for surgery. The proposed approach relies heavily on determining a means of volume reduction prior to space partitioning. for the purpose of conducting reliable assessments of image handling systems. Presented below are several suggested frameworks and a basic stream diagram.

On the job:An important step in evaluating a picture is dividing it into smaller sections. Segmenting a picture means breaking it down into smaller, more manageable pieces, and then assigning each piece a certain function. Disease categorization, diagnosis, and imaging by X-ray, CT, MRI, etc. This picture segmentation technique has broad use in healthcare and allows for the creation of photographs. Images captured from several clinical organs, such as the liver, lungs, brain, heart, and brain, are included here. Examining any abnormal growth or disease is their goal. In order for clinicians to provide the most effective therapy, these image segmentation algorithms differentiate between normal cells and malignant cells, such as tumours. Dividing an image into its foreground and background components is the first stage in picture segmentation. This technique is very helpful in clinical research. All of these pictures are becoming smaller as a result of the quality setting.

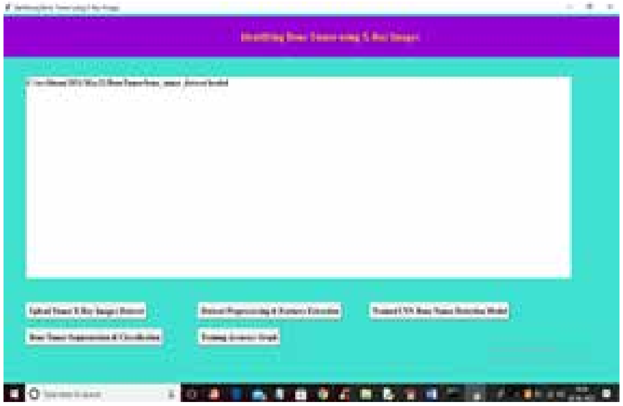
RESULTS EXPLANATION

Click "Run" twice to launch the project.run the attached bat file to get the output shown below

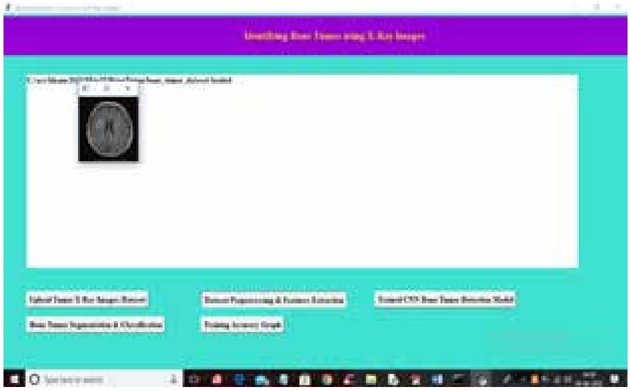
To get this result, go back to the first page and click the "Upload Tumour X-Ray Images Dataset" button.



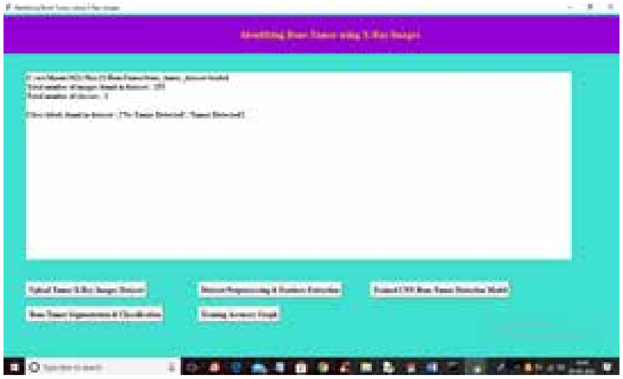
Once you've made your selection and uploaded the brain tumour dataset using the "Select Folder" option, you'll be able to see the results below.



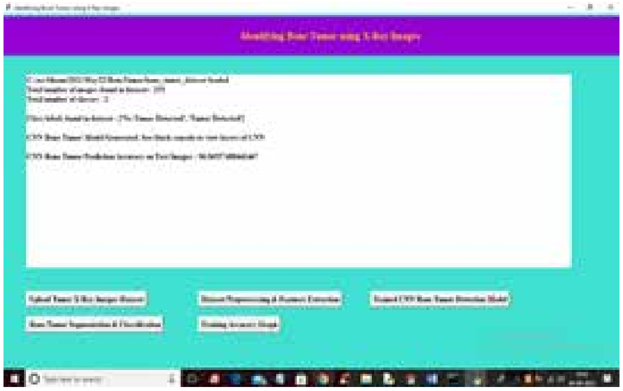
To read all the photos in the dataset, process them, and extract features to train using CNN, click the "Dataset Preprocessing & Features Extraction" button on the top screen.



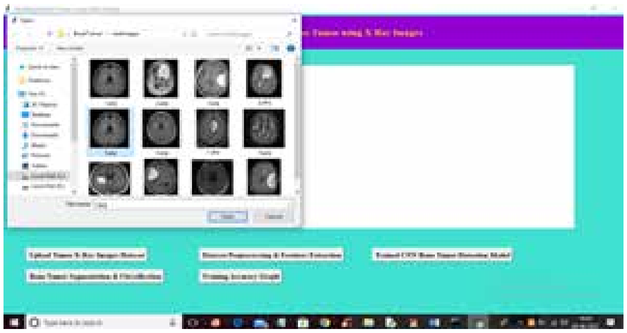
To make sure the photographs loaded correctly, I'm showing you a processed example image above; now you may close that image to see the results below.



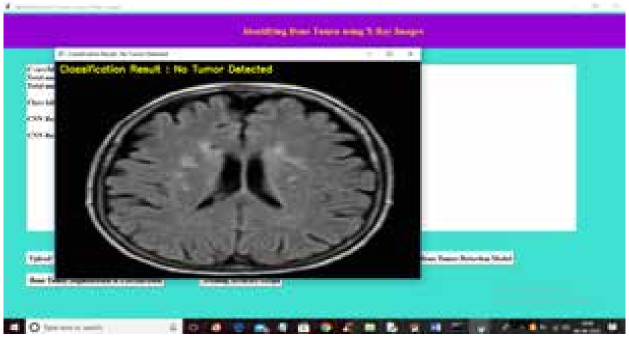
To train CNN using the extracted features shown above, click the "Trained CNN Bone Tumour Detection Model" button. The dataset comprises 253 pictures with and without tumour class labels. Then, you will obtain the result shown below.



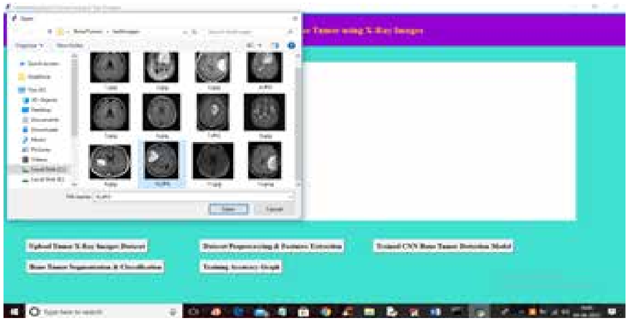
Following the completion of CNN training, which yielded an accuracy rate of 96%, we may submit a test picture by clicking the "Bone Tumour Segmentation & Classification" button, and the results will be shown below.

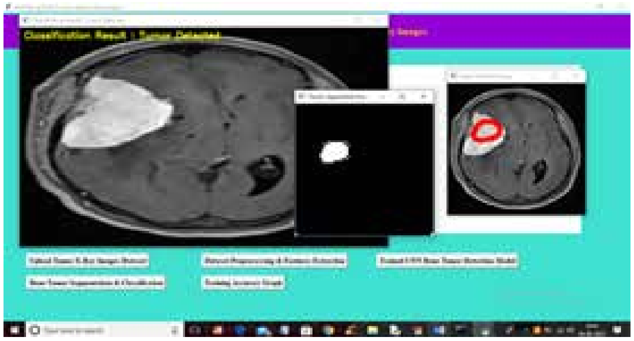


By going to the previous screen, choosing the 5.jpg file, and then clicking the "Open" button, you will get the following output:


Identifying Bone Tumor using X-Ray Images
Classification Result : No Tumor Detected

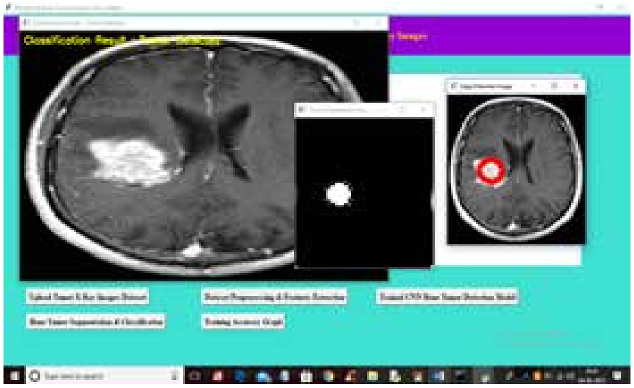
In above image 'No Tumor Detected' and now try another image.


Tumor using X-Ray Images
Upload Tumor X-Ray Images Dataset
Dataset Preprocessing & Features Extraction
Trained CNN Bone Tumor Detection Model
Bone Tumor Segmentation & Classification
Training Accuracy Graph

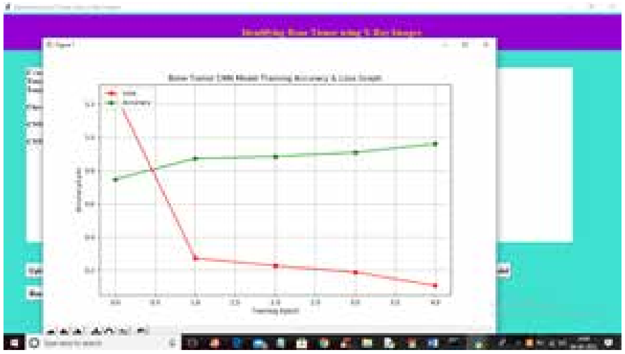

Classification Result : Tumor Detected
Upload Tumor X-Ray Images Dataset
Dataset Preprocessing & Features Extraction
Trained CNN Bone Tumor Detection Model
Bone Tumor Segmentation & Classification
Training Accuracy Graph

The three images shown above are the original, tumor-detected picture, the tumor-segmented image, and the tumor-edge-detected image; another image is shown below.

Tumor using X-Ray Images
Upload Tumor X-Ray Images Dataset
Dataset Preprocessing & Features Extraction
Trained CNN Bone Tumor Detection Model
Bone Tumor Segmentation & Classification
Training Accuracy Graph


Classification Result : Tumor Detected
Upload Tumor X-Ray Images Dataset
Dataset Preprocessing & Features Extraction
Trained CNN Bone Tumor Detection Model
Bone Tumor Segmentation & Classification
Training Accuracy Graph

Tumour identified using segmented out picture and tumour edge detected is shown in the above screen. To get the following graph, you may upload more photographs and run tests. Then, click on the "Training Accuracy Graph" button.


Identifying Bone Tumor using X-Ray Images
Bone Tumor CNN Model Training Accuracy & Loss Graph
Loss
Accuracy
Training Epoch

CONCLUSION

There is evidence that GGD analysis may successfully identify bone tumours from digital MRI scans. But without ground truth, we can't determine the bone cancer segmentation rate with any degree of certainty. Consequently, in order to conduct flawless assessments, a database of bone MRI scans must be built using precise and trustworthy expert opinion.

REFERENCES

1. RozyKumari, Narinder Sharma(2014, July). "A Study on the Different Image Segmentation Technique ". International Journal of Engineering and Innovative Technology (IJEIT) Volume 4, Issue 1, July 2014 ISSN: 2277-3754.

2. Kishor Kumar Reddy C, A. P. (2015). A Novel Approach for Detecting the Tumor Size and Bone Cancer. 2015 International Conference on Computational Intelligence and Communication Networks, (pp. 229-233).

3. Kishor Kumar Reddy C, A. P. A Novel Approach for Detecting the Bone Cancer and its Stage based. Recent Researches in Applied Computer Science, ISBN: 978-1-61804-307-8, 162-171.

4. Dipali M. Joshi, D. K. (2010). Classification of Brain Cancer Using Artificial Neural Network. 2nd International Conference on Electronic Computer Technology, (pp. 112-116).

5. Maryamsadat Amini, P. L. (2015). Thermographic Image Analysis Method in Detection of Canine Bone Cancer (Osteosarcoma). 5th International Congress on Image and Signal Processing, (pp. 485- 489).

6. Firke, M. H. (2016). Review Of Brain Tumor Detection Using MRI. International Journal for Research in Applied Science & Engineering, 4 (3), 479-484.

7. Madhuri Avula, N. P. (2014). Bone Cancer Detection from MRI Scan Imagery Using Mean Pixel Intensity. 8th Asia Modelling Symposium, (pp. 141-146).

8. Rahul kansal, P. g. (2014). Osteosarcoma or Ewing's sarcoma? Radiologist's Dilemma. Scholars Journal of Applied Medical Sciences (SJAMS), 2 (5), 1817-1820.

9. S.S.KUMAR, M. A. (2014). Recent Methods for the Detection of Tumor Using. International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), (pp. 1014-1019).

10. Vitulano, S., Di Roberto, C., &Nappi, M. (1997). Different methods to segment biomedical images. Pattern Recognition Letters, 18(11- 13), 1125-1131.

Earthquake Early Warning System: Using Machine Learning for Rapid and Reliable Source-Location Estimation

Bembadi Swathi, Basavantha Allen Chris, Dongre Laxmi
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

M Sireesha
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ sireesha.cse@cmrtc.ac.in

ABSTRACT

Our goal in creating this random forest (RF) model is to help earthquake early warning (EEW) systems make quick decisions when it comes to earthquake location. This method takes use of the P-wave arrival timings at the first five seismic stations to record an event and calculates the difference between each station's arrival time and a reference station. In order to determine the approximate position of the epicentre, the RF model categorises these differential P-wave arrival timings and station locations. The suggested system is trained and tested using a Japanese earthquake catalogue. The RF model does an excellent job at predicting where earthquakes will occur, with an MAE of 2.88 km. Also worth noting is that the suggested RF model may provide good results (MAE<5 km) with just 10% of the dataset and far fewer recording stations (i.e., three). This technique provides a novel and strong tool for fast and reliable source-location prediction in EEW. It is accurate, generalizable, and responds quickly.

KEYWORDS: *MAE, RF, EEW, Earthquake, P wave.*

INTRODUCTION

Many seismological applications depend on precise hypocenter localization of earthquakes, including tomography, source characterization, and hazard assessment. If we want to know when and where earthquakes started and where their hypocenters were, we need accurate seismic monitoring systems. On top of that, precise and timely characterization of active earthquakes is necessary for the development of tools to reduce seismic risks, such as earthquake early warning (EEW) systems [1]. Identifying the precise position of the hypocenter in real time is still challenging, even if classical methods are widely used in EEW system design. This is mostly due to the limited data that is accessible during the early phases of an earthquake [2]. Among other crucial aspects of EEW, one must improve hypocenter location predictions by using data from the first seismograph stations generated by the ground shaking and the few seconds after the arrival of the P-wave [3]. To solve the localization problem, seismograph stations that are triggered by ground shaking may be located and a time series of the waves that have been seen (arrival times) used. A recurrent neural network (RNN) is the optimal network design for controlling a network of seismic stations that are turned on in a sequential manner according to the paths that seismic waves follow as they travel. RNNs correctly extract data from a succession of inputs [4]. This method has been investigated for the purpose of improving real-time earthquake detection via the classification of source characteristics. Additional machine learning-based seismic monitoring methods have been proposed. Furthermore, more traditional machine learning methods like decision trees, support vector machines, and nearest neighbour algorithms have been compared to the earthquake detection problem [5].

EXISTING SYSTEM

If earthquake early warning systems are serious about mitigating seismic hazards, they must communicate the locations and magnitudes of upcoming earthquakes far in advance of the start of damaging S waves. It is possible that deep learning algorithms might be used to determine the earthquake's origins using whole seismic wave-forms rather than only seismic phase picks [6]. We developed a novel deep learning EEW system based on fully convolutional networks to detect earthquakes and estimate their source properties from continuous seismic waveform streams [7]. The system's capacity to adapt and improve its responses to incoming data allows it to precisely locate and determine the size of an earthquake even when only a limited number of stations register its occurrence. We put the system into action following the 2016 M 6.0 Central Apennines, Italy Earthquake and its subsequent aftershocks. Up to four seconds after the initial P phase, the precise location and magnitude of an earthquake may be determined with an accuracy of 8.5-4.7 km for the former and 0.3-0.27 for the latter.

PROPOSED SYSTEM

Using the positions of the stations and the periods of differential P-wave arrival, the system suggests an RF-based technique for earthquake localization (Figure 1). Only the timestamps of the Pwaves observed at the initial stations are used in the suggested method. The speed with which EEW notifications may be sent out is dependent on how quickly you can react to the arrival of earthquake early signs. Our approach takes the impact of the velocity structures into implicit consideration by including the source-station positions into the RF model. The suggested system employs a comprehensive Japanese seismic database for the purpose of evaluating the proposed technique. Our results provide fresh insight into the development of effective machine learning algorithms, since they demonstrate that the RF model can precisely pinpoint earthquake sites using sparse data.

WORKING METHODOLOGY

The technique suggests a radio frequency (RF) approach to earthquake localization using the stations' locations and the differential P-wave arrival timings. For this method to work, the initial stations' P-wave arrival timings are all that are needed. The speed with which EEW notifications may be sent out is dependent on how quickly you can react to the arrival of earthquake early signs. Including the source-station locations in the RFmodel implicitly includes the effect of the velocity structures in our technique. The suggested system employs a comprehensive Japanese seismic database for the purpose of evaluating the proposed technique. This RF model can localise DATA CHARACTERISTICS, as shown in our trials. A test case is only an executable object that other architectural modules may utilise; it does not interact with anything. To run a test with known parameters and anticipated outputs, you need a set of procedures, sometimes called test cases. Instead of software running automated test cases, humans carry them out by hand. Verify that all specified parameter values are covered by the test data while testing a system. Rather of attempting to test every possible value, we should choose a small subset of each equivalence class and work with them. While dealing with a set of values, you should always consider them to be an equivalence class. Test cases that verify error circumstances should contain procedures to examine the error messages and logs, in addition to the functional test cases. Verifying for error situations may still be done securely by conducting frequent functional test cases, even in the absence of stated test cases. If there are any problematic test data, please specify which ones.

Services Provider

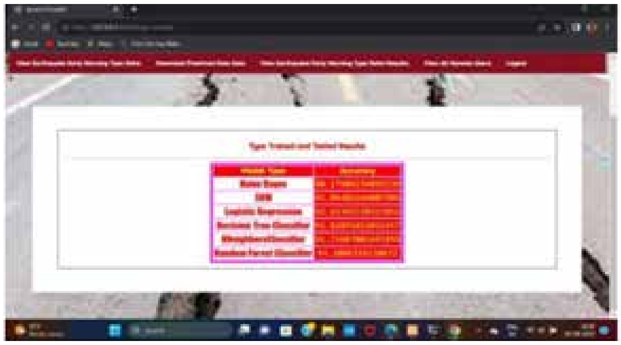
The Service Provider must log in using their credentials before they may access this module. Among other things, once he signs in, he will have access to the train and test data sets. Check out the Accuracy Bar Chart for Trained and Tested Data, Take a look at the Accuracy results for both Training and Testing, as well as the following: Some key elements to examine are the ratio of earthquake early warning kinds, the projected data sets for watching, the viewing outcomes, the ability to see from a distance, and the forecast of earthquake early warning types.

Always monitors users and the rights they have

This section gives an exhaustive list of all registered users, which is a source of great joy for the administrator. Administrators have access to all user information, including names, email addresses, and physical addresses. They may also use this area to provide access.

Maintain a distance

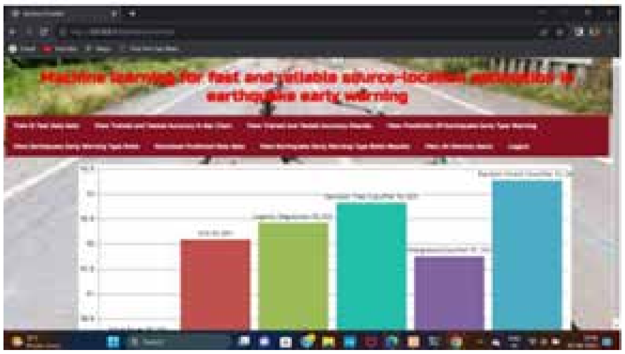
With a minimum of n active participants, this module is complete. Before the user can take any action, they must register. When a person signs up, their information is automatically put to our database. After he finishes registering, you will ask him for his authorised login credentials. Once the login process is complete, the user will have access to their profile, the ability to predict the kind of earthquake warning, and the ability to log in.



Type Trained and Tested Results

Model Type	Accuracy
Naive Bayes	[illegible]
SVM	[illegible]
Logistic Regression	[illegible]
Decision Tree Classifier	[illegible]
KNeighborsClassifier	[illegible]
Random Forest Classifier	[illegible]

Fig. 1. INPUT module



Machine learning for fast and reliable source-location estimation in earthquake early warning

Fig. 2. Output results

Machine learning for fast and reliable source-location estimation in earthquake early warning

Earthquake Early Warning Type Ratio Details

Warning Type	Ratio
Earthquake Warning	[illegible]
Explosion Warning	[illegible]

Fig. 3. Accuracy level indication

CONCLUSION

By analysing the timing disparities between the arrival of P-waves and the locations of seismic stations, we can precisely locate the epicentre of an earthquake as it is occurring. One possible solution to this regression problem is to use a random forest (RF), the output of which would be the difference in latitude and longitude between the earthquake and the seismic stations. Results from the case study of the seismic zone in Japan indicate that it is effective and ready for deployment. We extract all events with five or more P-wave arrival times from the nearby seismic stations. We then split the retrieved events into a training dataset and a testing dataset so that we could construct a machine learning model. The proposed method is also adaptable enough to be employed for real-time earthquake monitoring in more challenging places; it trains with as little as 10% of the dataset and three seismic stations, but it still gets interesting results. Due to insufficient catalogues and station dispersion, many networks are sparsely distributed, making it difficult to train an effective model using the random forest approach. However, by using several synthetic datasets, one may compensate for the absence of ray routes in a target location.

REFERENCES

1. General description about Seismic data detection (March 2021) - https://grillo.io/data/
2. Machine Learning Model (April 2021) - https://openeew.com/docs/machine-learning
3. Learnt about implementation of earthquake network (April 2021) - https://sismo.app/
4. Working Model (June 2021) - https://grillo.io/impact/#openeew
5. Case Study on ShakeAlert - An Earthquake Early Warning System for the West Coast of the United States (June 2021) – https://www.shakealert.org/implementation/shakealert -phase-1 U.S. Geological Survey. p. 4. doi:10.3133/fs20143083. ISSN 2327-6932.
6. Karlsson, I.; Papapetrou, P and Bostrom, H. (2016.) ‘Generalized Random Shapelet Forests.’ Data Mining and Knowledge Discovery 30:1053–1085
7. Kevin Fauvel.; Diego Melgar, Manish Parashar. (2018). “A Distributed Multi-Sensor Machine Learning Approach to Earthquake Early Warning”
8. Schafer, P., and Leser, U. (2017.) ‘Multivariate Time Series” Classification with WEASEL+MUSE.
9. Yoon, C.; O’Reilly, O.; Bergen, K. J.; and Beroza, G. C. (2015.) ‘Earthquake Detection Through Computationally Efficient Similarity.’ Science Advances.

A Cloud-based Crypto-Biometric Programme with Strong Security Measures

Kotagiri Rishi Kumar, Yellu Nithin Reddy, Gurram Varshith Goud
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

J Prasanna Babu
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ prasannababu.cse@cmrtc.ac.in

ABSTRACT

There is a wide variety of providers and cloud-based services accessible now that cloud computing has achieved maturity. However, a lot of focus is still on security concerns. Users are wary of adopting cloud computing because they worry about their security and privacy, even if the technology has many advantages. Biometric technologies are the backbone of many emerging methods of secure identification and personal verification; nevertheless, cloud computing has unique issues when it comes to storing biometric data, due to privacy rules and the need to trust cloud providers. This research suggests a cryptobiometric approach to cloud computing that safeguards users' personal biometric data and thereby removes these problems.

KEYWORDS: *Biometric, Cloud data, Secure data.*

INTRODUCTION

One new trend in application development and design is cloud computing, which is also a new method of conducting business. Success stories from many service providers, including Amazon, have shown that the method works for a variety of solutions across the many levels of the cloud paradigm (SaaS, PaaS, and IaaS). While cloud computing is still not without its limitations, it has matured to a certain point. Companies that move their applications, data, and infrastructure to the cloud enjoy several benefits, but they also have to relinquish some control over their data. Users do not own, operate, or control the machines that process the information. Here, a high level of trust is necessary as the user has no idea how the supplier handles data. Because the system's physical and logical components cannot be managed, substantial changes to security and privacy rules are required.

OVER VIEW: The goal of this project is to develop a secure cloud-based biometric solution. Cloud computing has exploded in popularity due, in part, to the ease and convenience of remote data storage and retrieval. The sharing and storage of sensitive information has, however, increased concerns about data security. The goal of this study is to find a solution by combining biometric authentication with encryption.

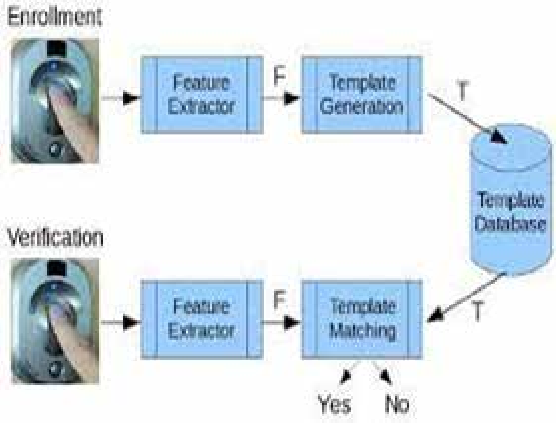
Data stored on the cloud will be protected by the system's usage of advanced cryptographic techniques, which ensure its authenticity, integrity, and confidentiality. These techniques will encrypt data before transmitting it and decode it once it is received, preventing any loss or modification of data. The system will use biometric identification techniques such as iris or fingerprint scanning to further guarantee that only approved users may access the cloud.

Biometrics and encryption work together to form the system's robust and multi-layered security architecture. It will prevent unauthorised individuals from accessing the infrastructure and ensure that only approved users with verified biometric credentials may access the data stored in the cloud. Traditional authentication methods are made obsolete by this technology due to problems like password cracking and sharing.

There is a well-known security mechanism, the data is stored in the user infrastructure, and its location is also known. However, customers have no clue about the physical location of their data when they use public cloud computing services. This makes it difficult to ensure adherence to national regulations. One example is the fact that European data protection regulations impose extra constraints on the processing and storage of data sent to the United States, making it possible to breach these regulations by keeping biometric templates in Amazon S3 services. Multiple ideas have addressed the issue of biometric template security. Of them, cancelable biometrics is one that shows promise [10]. In addition to ensuring that the stored biometric template cannot be used to retrieve the original biometric data (non-invertibility), it also allows for the issuance of a new biometric template in the case that an existing one is compromised (renewability).

SUGGESTED SYSTEM

Following the proposed framework, training a UBM becomes feasible if a large database including sample acquisitions is amassed. To make our system more secure and flexible, we employ the training procedure shown in Figure 3. New UBMs must be added for this to work. The computing resources required to train a new UBM are provided by virtual machines located in Amazon Elastic Compute Cloud (EC2). Using Amazon EC2's application programming interface (API), the management software automatically requests the required virtual machines. Distributed execution of the training programme and the new UBM on the machines speeds up computation. Biometric data processing is highly parallelizable, which opens the door to speedup.



IMPLEMENTATION

Modules

Firstly, we'll use the "Upload Fish Dataset" module to transfer the dataset to the application. Secondly, we'll read all of the photographs, process them using interpolation, CLAHE, and LAB. After that, we'll normalise the images and divide the dataset into a "train" and "test" set.

Third, put the model to work by running it through its paces using the processed train photos as input. The model will then be applied to the test images in order to determine metrics like prediction accuracy.

Step 4: Execute Logistic Regression: This step involves feeding the processed train photos into logistic regression in order to train a model. Then, using this model on the TEST images will allow you to measure metrics like prediction accuracy.

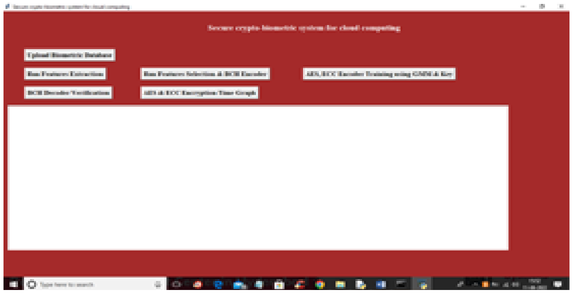
5)Run Naive Bayes: processed train photos will be fed to naïve bayes to build a model and this model will be used on TEST images to measure prediction accuracy and other metrics

6)Propose and Execute the Support Vector Machine Algorithm: The training photos will be fed into the SVM algorithm to create a model, which will then be tested on the test images to determine the prediction accuracy and other metrics.

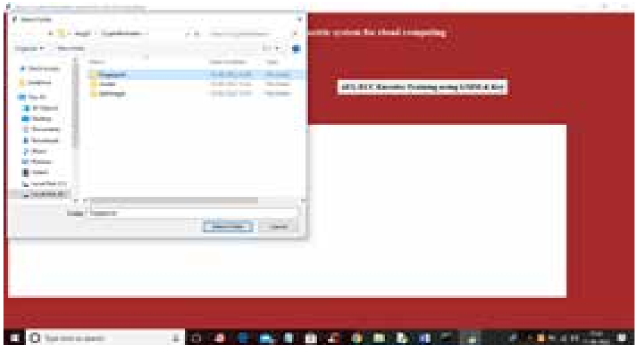
7).Comparison Graph: We will use this module to create graphs showing metrics like accuracy.

8)Anticipate Fish Condition: This module allows us to input a test picture and then use a support vector machine algorithm to determine whether the image comprises healthy or diseased fish.

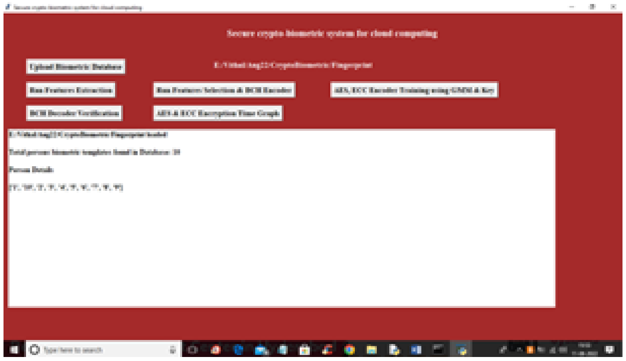
Get the following screen when you double-click the "run.bat" file to launch the project:


Secure crypto-biometric system for cloud computing
Upload Biometric Database
Run Features Extraction
Run Features Selection & BCH Encoder
AES, ECC Encoder Training using GMM & Key
BCH Decoder Verification
AES & ECC Encryption Time Graph

To import biometric data, go to the previous page and find the "Upload Biometric Database" button. Then, you should see the following results.


Secure crypto biometric system for cloud computing
AES, ECC Encoder Training using GMM & Key

Click the "Select Folder" button to load the database once you've uploaded the fingerprint biometric pictures dataset on the previous page. You can see the results down below.

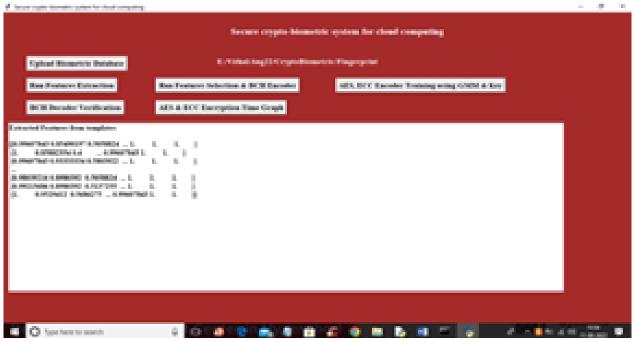

Secure crypto biometric system for cloud computing
Upload Biometric Database
Run Features Extraction
Run Features Selection & BCH Encoder
AES, ECC Encoder Training using GMM & Key
BCH Decoder Verification
AES & ECC Encryption Time Graph

E:/Vithal/Aug22/CryptoBiometric/Fingerprint loaded

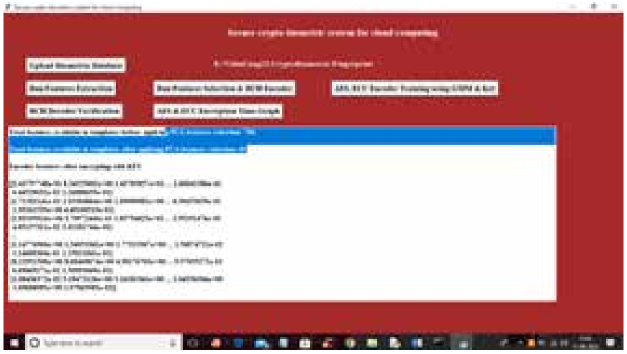
Total person biometric templates found in Database: 10

Person Details

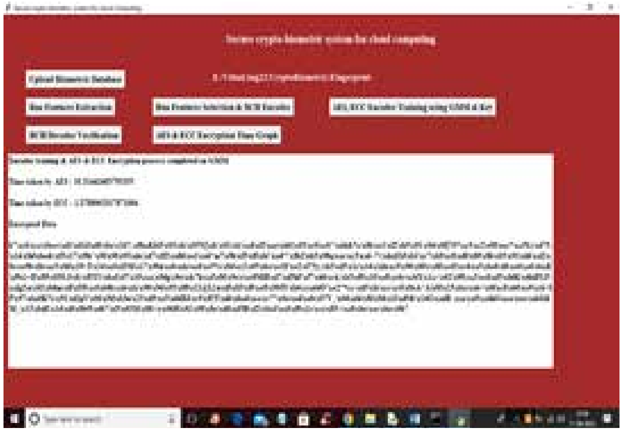
The database is loaded, and as you can see above, it includes biometric templates for ten separate individuals. To extract features from these templates, click the "Run Features Extraction" button, and the results will be shown below.


Secure crypto biometric system for cloud computing
Upload Biometric Database
Run Features Extraction
Run Features Selection & BCH Encoder
AES, ECC Encoder Training using GMM & Key
BCH Decoder Verification
AES & ECC Encryption Time Graph

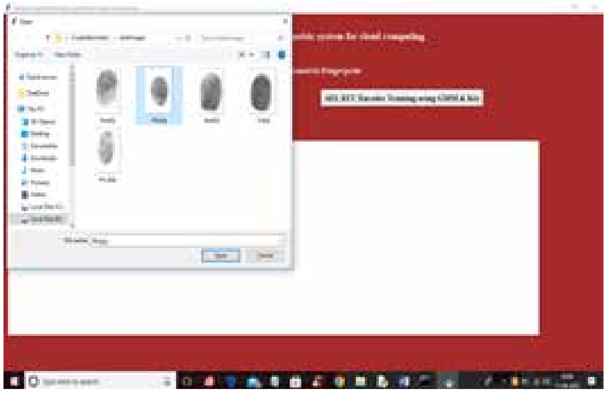
After the features have been retrieved, you may choose them by clicking the "Run Features Selection & BCH Encoder" button on the previous page.


Secure crypto biometric system for cloud computing
Upload Biometric Database
Run Features Extraction
Run Features Selection & BCH Encoder
AES, ECC Encoder Training using GMM & Key
BCH Decoder Verification
AES & ECC Encryption Time Graph

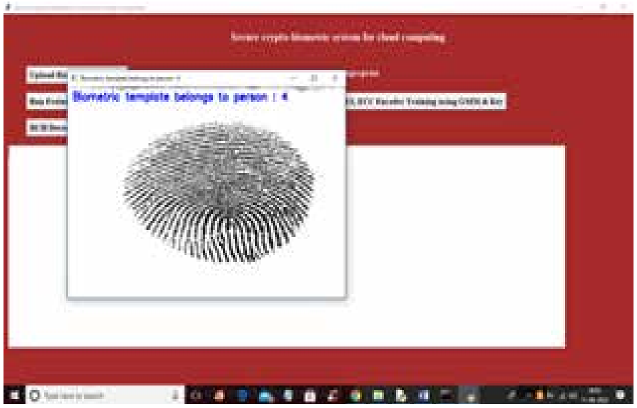
After sorting through 784 features using the PCA features selection algorithm, we're left with 60 that we should prioritise. To encode these features, we can use the "AES, ECC Encoder Training using GMM & Key" button. After that, we can train GMM, which will be encrypted using the ECC and AES algorithms. The result will be displayed below.


Secure crypto biometric system for cloud computing
Upload Biometric Database
Run Features Extraction
Run Features Selection & BCH Encoder
AES, ECC Encoder Training using GMM & Key
BCH Decoder Verification
AES & ECC Encryption Time Graph

Click the "BCH Decoder Verification" button to submit the template and retrieve the verification result. Then, we can observe that GMM is encrypted. AES took 10.31 seconds and the ECC tool 1.2 seconds.

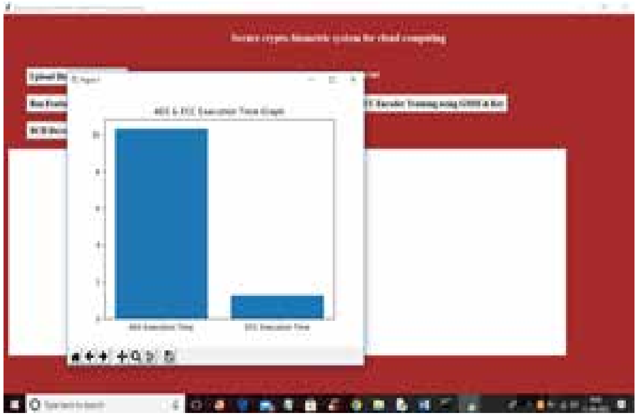

Secure crypto biometric system for cloud computing
AES, ECC Encoder Training using GMM & Key

To get the following result, choose and upload a finger template on the previous page, and then click the "Open" button.



Biometric template belongs to person : 4

Click the "AES & ECC Encryption Time Graph" button to get the graph below. In the above screen, you can see the template that belongs to person 4.


Secure crypto biometric system for cloud computing
AES & ECC Execution Time Graph
10
8
6
4
2
0
AES Execution Time
ECC Execution Time

For both algorithms, ECC had the faster execution time, as shown on the x-axis of the following graph, which shows the names of the encryption methods and the y-axis the execution time.

CONCLUSION

Finally, developing a secure crypto-biometric system for use in the cloud might be a solution to the growing problem of data privacy and security. This state-of-the-art technology offers a reliable means of safeguarding confidential information stored in the cloud by combining biometric identification with the strength of cryptographic algorithms. By using cryptographic techniques, the technology guarantees that data sent to and stored in the cloud is secure, rendering it inaccessible to unauthorised individuals. This establishes the framework for safeguarding user data across the whole cloud computing lifecycle. Another possible security solution is biometric authentication, which verifies a user's identity by analysing their unique physical or behavioural traits. Compared to authentication methods that rely on passwords, which might result in weak passwords or credential theft, this significantly reduces the likelihood that unauthorised persons would be able to access vital information. In addition to enhancing data security, the secure crypto-biometric system offers customers convenience and efficiency. By using biometric qualities like fingerprints or facial recognition, users are no longer need to recall complex passwords, leading to a more streamlined and user-friendly experience. Finally, a secure crypto-biometric system for cloud computing is the best way to safeguard sensitive data and fix big security problems. This will inspire trust in the cloud computing ecosystem among both organisations and consumers moving forward.

REFERANCES

1. A. A. M. Abd Hamid, N. and A. Izani. Extended cubic b-spline interpolation method applied to linear two-point boundary value problem. World Academy of Science, 62, 2010.

2. T. Acharya. Median computation-based integrated color interpolation and color space conversion methodology from 8-bit bayer pattern rgb color space to 24-bit cie xyz color space, 2002. US Patent 6,366,692.

3. A. F. Agarap. An architecture combining convolutional neural network (cnn) and support vector machine (svm) for image classification. arXiv preprint arXiv:1712.03541, 2017.

4. A. Ben-Hur and J. Weston. A user's guide to support vector machines. In Data mining techniques for the life sciences, pages 223– 239. Springer, 2010.

5. S. Bianco, F. Gasparini, A. Russo, and R. Schettini. A new method for rgb to xyz transformation based on pattern search optimization. IEEE Transactions on Consumer Electronics, 53(3):1020–1028, 2007.

6. E. Bisong. Google colaboratory. In Building Machine Learning and Deep Learning Models on Google Cloud Platform, pages 59–64. Springer, 2019.

7. A. P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern recognition, 30(7):1145– 1159, 1997.

8. S. A. Burney and H. Tariq. K-means cluster analysis for image segmentation. International Journal of Computer Applications, 96(4), 2014.

9. M. A. Chandra and S. Bedi. Survey on svm and their application in image classification. International Journal of Information Technology, pages 1–11, 2018.

10. L. de Oliveira Martins, G. B. Junior, A. C. Silva, A. C. de Paiva, and M. Gattass. Detection of masses in digital mammograms using kmeans and support vector machine. ELCVIA Electronic Letters on Computer Vision and Image Analysis, 8(2):39–50, 2009.

Design of an Online Tour Guide

Vanga Bhavana, Jeerla Pranay,
Daripelli Akshay Kumar
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

B K Bhagyashree
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ bhagyashree.cse@cmrtc.ac.in

ABSTRACT

Numerous tourists go to world-renown landmarks on a daily basis. Despite the fact that there are several undiscovered locations that deserve a visit, few people know about them because of the lack of public awareness. Leaders, as you are aware, are "people who show the way to others" those who, in every circumstance, take a stand that is crucial and guide others. We are working on a software that may lead visitors to the Maharashtra stronghold using mobile devices, similar to the, and decrease the need for workers. The proliferation of smartphones and other portable electronic gadgets has revolutionised the way people access and use information. One emerging trend in the travel business is the usage of virtual tourist guides, which are electronic travel guides accessible via mobile devices. There is a plethora of travel information available to users online. The challenge of creating a smartphone app that can read the names of landmarks and monuments from a live photograph is shown in this study. Fire, ambulance, and police departments are some of the emergency contacts that consumers may use in the event of an emergency. To enhance the application's interactivity, we have included a help desk where users may directly ask questions in the event that they need assistance.

KEYWORDS: *User, Admin, Travel, Tourism industry.*

INTRODUCTION

By fusing the ideas of virtual reality with those of traditional tourism, the term "virtual tourism" describes a hybrid concept. Essentially, virtual tourism allows for a tourist experience to be had without really leaving the house. There is a wide range of forms and levels of technology capacity in virtual tourism. [2]. Promoting e-tourism and exploring how the industry might embrace new technology are central to the "virtual travel guide" project's aims. The travel industry is constantly reevaluating its place in the world and necessitating management and marketing shifts as a result of the fast expansion of the e-tourism sector. Numerous sectors, such as the entertainment business, design, education, and tourism, are already making extensive use of virtual reality technology. On the one hand, by removing geographical and temporal barriers, this kind of community facilitates learning, the maintenance of relationships, and the discovery of like-minded persons with whom one may never have crossed paths otherwise. Many initiatives have been launched in the last ten years to address the widespread issue of information overload and incorporate technology into tourist services. While we have made several system discoveries, their reach is somewhat restricted. Travellers are still required to utilise numerous apps due to the fact that these systems operate with distinct services. The tourist industry is well-suited to the use of web computing since travellers want information at their fingertips at all times. This article details a tourist information system that makes it possible to have a centralised virtual travel guide with a variety of services. Travellers can get all the information and tools they need to plan their journey on the online services that constitute the basis of our proposed system.

LITERATURE SURVEY

When it comes to international trade, the tourist sector is among the most reliable. Approximately 700 million visitors visit each year, with that figure expected to almost double by 2020. It also employs close to 200 million people and contributes around 11% to GDP.

The development of information and communication technologies has made it possible for curious tourists to have access to interesting information online. Smartphones are quickly becoming the norm in this area, thanks to the 700 million active iOS and Android devices globally. There was a meteoric rise in 2012 as mobile data traffic made up 13% of total Internet traffic.

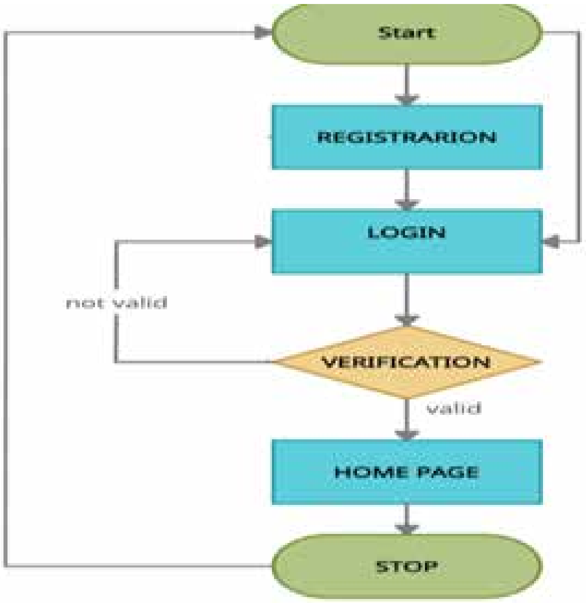
Third, for many countries, the tourist sector is already one of the most significant economic drivers. The methods used to lead tourists by tour guides might vary greatly. Most tourists rely on printed tour brochures that detail the history of the region as well as recommended itineraries.

PROPOSED SYSTEM

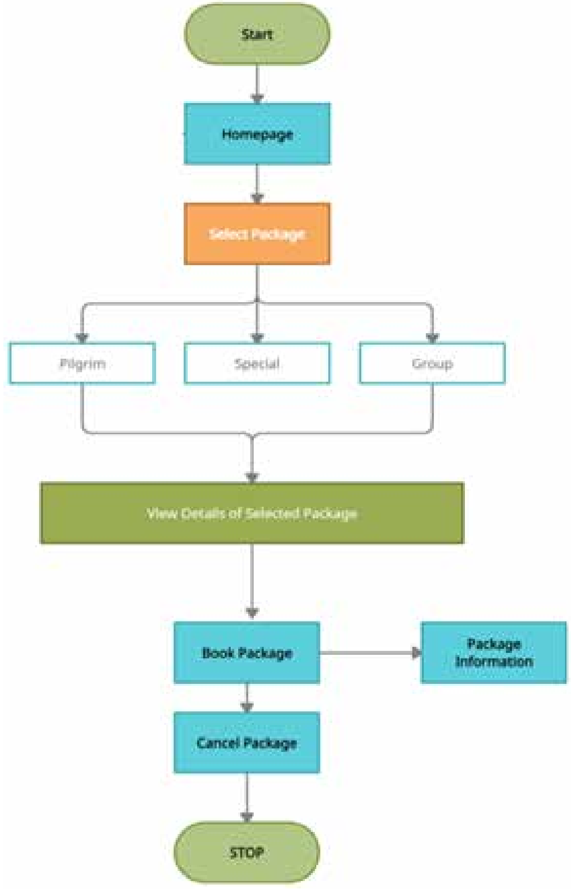
Virtual reality headsets and the technology that powers them have taken the world by storm in the last few years. Other businesses, including tourism, are beginning to see the light and are exploring the potential of the technology, especially in marketing, while the gaming industry has been leading the home adoption charge. First thing: Yes. The travel industry is both short- and long-term impacted by technological advancements. The travel industry has been lately impacted by new information and communication technology advances in many ways, from customer interest to the site as a whole. The tourist sector would do well to give more consideration to the many useful applications of virtual reality (VR). Virtual reality (VR) might revolutionise several aspects of the travel industry. These include planning and the board, advertising, distraction, training, availability, and legacy protection, among many others. The phrase "virtual community" (VC) describes an online group of people who have comparable interests and collaborate on projects using current ICT and predetermined standards and procedures. The fundamental needs of community members, together with the primary characteristics of virtual communities and virtual reality, form the basis of this study's theoretical framework for a virtual travel community. The viewpoints of different tourism companies on user contact are presented in a clear and concise way, along with the difficulties of dealing with online marketplaces and other similar scenarios.

METHODOLOGY

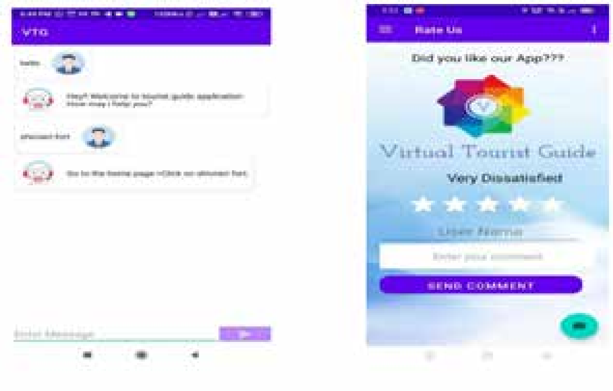
The first step is to create a system that streamlines the time-honored search-and-book method that most people use. A lot of people have to manually buy several plane and train tickets, but our technology will make that process much easier. With only a few simple actions, the user may do the task faster and more efficiently. Since the system will be online, the user will have access to the offered information whenever they need it. As part of user-centered design (UCD), designers iteratively consider the requirements and wants of users at every stage. Using a wide range of research and design methodologies, UCD design teams actively incorporate consumers all the way through the design process, resulting in products that are both accessible and very useful.[4]. Here, the client plays a key role. The system's user interface is designed to be interactive, so users may actively participate in finding the optimal destination. The system has been developed and launched as a website, which allows users to access it from anywhere at any time, according to the following criteria. After some initial difficulties with the website's implementation, we overcame them by creating a prototype and getting feedback from a number of individuals before settling on the optimal design.



More in-depth information on the locations that visitors visit is what the proposed system, "VIRTUAL TOURIST GUIDE," is all about. Paper tour guide systems aren't up to the task of intelligent representation and precise navigation due to their dated design and dependence on static, photocopied visuals. Here we set the goal of developing an app for smartphones that can identify landmarks and monuments and provide information about them as soon as the user takes a picture of them. Put simply, the app should enable the user to take a picture of the landmark or monument, and then use that photo to display and explain the landmark's story. We then follow the user's screen, as shown in the figure, and superimpose the 3D model on top of the real-world equivalent to make the augmentation evident. Our interactive visualisation spaces are an extra feature of the AR interface.



After they're defined, use cases may take many forms, including text and visual representations (like a use case diagram). Using use-case modelling, we may build a system considering the user's needs in detail. By outlining all system behaviour that is visible to the outside world, it is a powerful method for explaining system behaviour to the user in their own words. [5]. The intended flow of user data on the website may be better understood with the aid of the Login Module Diagram, a use case diagram. This figure will also be useful for showing the workflow's mechanism and sequence. The module depicts the user's expected interactions with the website graphically. This includes launching the site, selecting "Resign," logging in, and verifying the user's identity; if everything goes well, the user is sent to the homepage; otherwise, an error notice is shown.



hello

Hey! Welcome to tourist guide application How may I help you?

shivneri fort

Go to the home page >Click on shivneri fort

Did you like our App???

Very Dissatisfied

CONCLUSION

The primary goal of this project is to improve the Maharashtra Fortress's trip guide system via the use of location-based mobile apps. The data was useful for advancing the discipline in the future, and it was fascinating to acquire. Research papers are being used for a tremendous number of literature reviews. A further key assumption made during system development is that users should be acquainted with English and have a basic understanding of how to utilise Android mobile devices. Our long-term goal is to develop this app for all of our nation's forts, with the option to include other languages down the road.

REFERENCES

1. Wei, X., Weng, D., Liu, Y., & Wang, Y. (2016). A tour guiding system of historical relics based on augmented reality. 2016 IEEE Virtual Reality (VR). doi:10.1109/vr.2016.7504776

2. Grün, C., Werthner, H., Pröll, B., Retschitzegger, W., & Schwinger, W. (2008). Assisting Tourists on the

MoveAn Evaluation of Mobile Tourist Guides. 2008 7th International Conference on Mobile Business. doi:10.1109/icmb.2008.28

3. Burta, A., Szabo, R., & Gontean, A. (2020). Object Recognition Development for Android Mobile Devices with Text-to-Speech Function Created for Visually Impaired People. 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4). doi:10.1109/worlds450073.2020.9210381

4. Chaisoong, U., & Tirakoat, S. (2020). The Clustering of Questions Affect to Tourist's Decision Making for Chatbot Design. 2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON). doi:10.1109/ecti-con49241.2020.9158069

5. Deepthi Jordhana, P., & Soundararajan, K. (2014). Kernel methods and machine learning techniques for man-made object classification in SAR images. International Conference on Information Communication and Embedded Systems (ICICES2014). doi:10.1109/icices.2014.7034068

6. De Farias, I., Leitao, N., & Teixeira, M. M. (2017). Urbis: A touristic virtual guide. 2017 12th Iberian Conference on Information Systems and Technologies (CISTI). doi:10.23919/cisti.2017.7975918

7. Kenteris, M., Gavalas, D., & Economou, D. (2006). Developing Tourist Guide Applications for Mobile Devices using the J2ME Platform. 2006 Proceedings of the First Mobile Computing and Wireless Communication International Conference. doi:10.1109/mcwc.2006.4375218

8. Li, H., & Zhijian, L. (2010). The study and implementation of mobile GPS navigation system based on Google Maps. 2010 International Conference on Computer and Information Application. doi:10.1109/iccia .2010.6141544

9. Saranyaraj, D. (2013). The virtual guide for assisted tours using context aware system. 2013 International Conference on Signal Processing , Image Processing & Pattern Recognition. doi:10.1109/icsipr.2013.6497973

10. Sharma, S., & Agrawal, A. (2010). IMTS- an Interactive Multimodal Tourist-Guide System. 2010 International Conference on Signal and Image Processing. doi:10.1109/icsip.2010.5697475.

A System for the Real-Time Estimation of Contactless Vital Signals

Khaja Mubashiruddin, Puchakayala Bharath Simha Reddy, Gadde Sayi Khushhal
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

Chikati Madhava Rao
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ madhav.chikati@gmail.com

ABSTRACT

A growing body of research supports the use of facial video for non-contact vital sign monitoring, with a focus on HR and BR. Technological advancements in recent years have been remarkable, but they aren't perfect. For instance, the technology isn't portable, the estimation process is cumbersome, and there aren't enough hard datasets. This paper presents a new way to estimate HR and BR by combining a Convolutional Neural Network (CNN) with the Phase-based Video Motion Processing (PVMP) technique. The results of the experiments show that our technique is superior to the others. We also provide a new challenging dataset that loosens limits on things like light interference, facial expressions, and large movements. Our new Android app makes advantage of a convolutional neural network (CNN) to provide real-time HR and BR estimates, regardless of network connectivity.

KEYWORDS: *CNN, PVMP, HR, BR.*

INTRODUCTION

Numerous academics have focused on vital sign measurement in the last few decades. Health care for patients and illness prevention for the elderly are only two of the many ways in which the vital signs monitoring system helps humans. There have been several proposals for ways to detect HR and BR, since these vital signs—also known as respiratory rate (RR)—have been receiving more and more attention.

There are two basic categories for HR and BR measurement techniques: contact and non-contact. Because of their reliability and consistency, contact methods have long been the gold standard for sign measurement. A large number of contact techniques depend on photoplethysmography or electrocardiography (ECG). Putting sticky gel electrodes on a person's chest or limbs is the most common approach to generating an electrocardiogram (ECG) signal, which contains a wealth of valuable information about vital signs. The photoplethysmography (PPG) optical method may detect changes in the micro-vascular bed of tissue's Blood Volume Pulse (BVP) (i.e., changes in the observed light intensity). At the same time, it is possible to estimate the HR and BR from the variations in the BVP, which carry valuable information about the cardiovascular system. Since transmitted light is readily detectable, a PPG sensor may be conveniently positioned on the finger tip to efficiently monitor fluctuations in blood flow. So, PPG is only one of several contact-based methods that have gained popularity recently. The fact that users often have to wear gadgets that need skin-touch means that various contact techniques might potentially annoy and irritate users. So, to estimate HR and BR contactlessly in real-time, we suggest an iPPG-based approach.

Coronavirus Disease-19 (COVID-19) is a newly emerged virus that has recently caused a global outbreak. Transportation, education, and health care are just a few areas where this virus has drastically altered human lifestyles. Governments have also taken other critical measures, including quarantining cities, to halt the epidemic. Wearing a mask, avoiding close contact with others, and washing one's hands often are all recommended health measures. The importance of non-contact technologies for monitoring physiological signs is growing in light of the COVID-19 pandemic. The estimate job has also seen a plethora of contactless approaches based on various sensors, Hench. A large number of scientists are focusing on developing methods for the distant detection of physiological signals, such as heart and respiration rates. A PPG signal may be obtained from human faces using a regular digital camera and ambient light, as shown by Pavlidis et al. [1]. Imaging Photoplethysmography (iPPG) is a method for calculating the PPG signal from video pictures that is similar to contact PPG utilising a photodetector. Averaging the values of the Region of Interest (ROI) pixels is an integral part of the iPPG signal extraction process. Then, in either the temporal or frequency domains, vital indicators may be extracted from the iPPG signal. Researchers have suggested several unique or upgraded iPPG-based algorithms to achieve greater performance, thanks to the iPPG signal's low-cost and non-invasive collection method.

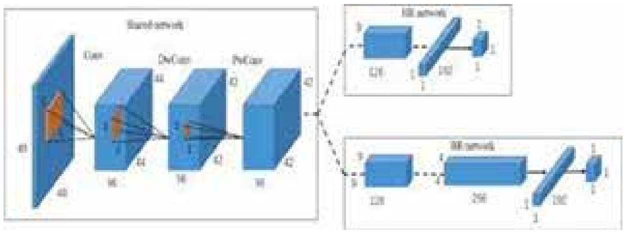
Furthermore, the fast advancement of Deep Learning (DL) in the last few years has resulted in several significant advances for computer vision jobs. By using a variety of deep learning approaches, several computer vision tasks were able to attain remarkable gains in accuracy. The use of deep learning techniques to estimate vital signs is therefore another novel approach that has shown promising results. In order to estimate the heart rate from a video sequence, Spetlik et al. [2] suggested a CNN model that consists of two stages. In order to predict HR, Qiu et al. [3] used a CNN model trained on feature maps taken from video sequences using the (Eulerian Video Magnification) EVM method [4]. However, rather of considering a number of factors, the aforementioned DL-based approaches just consider the heart rate. We improved upon the results of an EVM-based method by building a two-task CNN model with PVMP to estimate HR and BR, drawing inspiration from the work of Qiu et al. [3].

EXISTING SYSTEM

When evaluating HR and BR, both hands-on and passive approaches are acceptable. The reliability and consistency of contact methods made them the de facto norm for sign measurement for quite some time. Electrocardiography (ECG) and photoplethysmography provide the basis of several contact approaches. Putting adhesive gel electrodes on the patient's chest or limbs is the standard method for obtaining electrocardiogram (ECG) data. The patient's vital signs may be learned a great deal from these signals. Photoplethysmography (PPG) is an optical method that may be able to detect variations in blood volume pulse (BVP) inside the microvascular bed of tissue. Changes in the BVP may also be utilised to forecast the HR and BR, which are vital signs of the cardiovascular system. One possible use for a PPG sensor is to monitor variations in blood flow by placing it on the tip of a finger and detecting transmitted light. So, PPG is only one of numerous contact-based methods that have been trending upwards recently. Users may feel annoyed and even nauseous since they are constantly compelled to wear items that contact their skin. Based on this, we suggest an iPPG-based approach to estimate HR and BR in real-time without the need for interaction.

Recommended Framework

The database contains 206 movies with a total of 30 frames per second, and the video resolution is 1080 × 1960 pixels. There are a total of 602 minutes in our dataset. Three different types of vital signs—ECG, BR, and HR—are kept in the database. The eleven individuals that provided the signals were fourteen men and four women, with ages ranging from twenty-three to fifty. For the "exercise" and "stationary" scenarios, the dataset is further subdivided. The application's process is shown in Figure 13, and the study provides HR and BR values that are derived from exercise. Users have the option to switch between the front-facing and rear-facing cameras on their smartphones as soon as the app launches. The subsequent phase involves applying a classifier to the acquired picture in order to identify faces. The data buffer may store up to 48 photos, therefore it will choose one to transmit the ROI to when a face is recognised in the image. If you don't do this again, the data buffer will remain full. In addition, to maintain a rapid pace.



METHODOLOGY

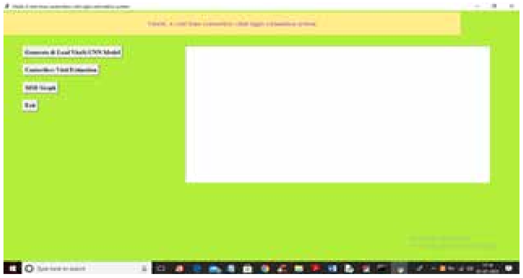
Modules

To begin, we can build a CNN model in VitaSi that can predict the user's breathing rate and heart rate. Then, we can input it into this module.

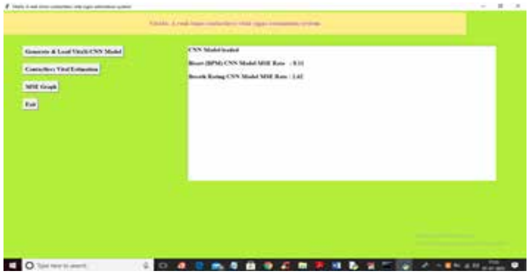
The second step is for the module to launch a camera, take ten frames, and then use a convolutional neural network (CNN) model and PPG signal extraction to estimate the respiratory and heart rates. This method is known as contactless vital assessment.

A third visual depicts the accuracy of the convolutional neural network (CNN) model's pulse and respiration rate predictions. It is called the mean square error (MSE) graph. The prediction model is enhanced by decreasing the MSE.

This is the screen that will appear once you start the project by double-clicking the "run.bat" file:

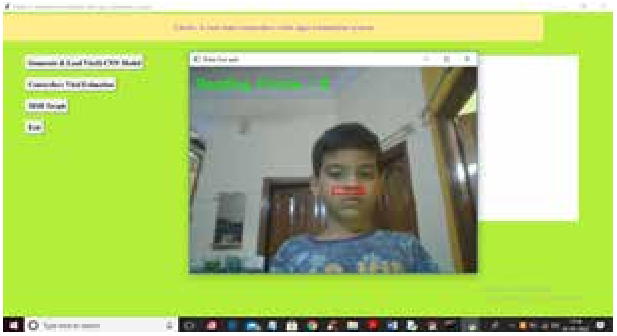

Generate & Load VitaSi CNN Model
Contactless Vital Estimation
MSE Graph
Exit

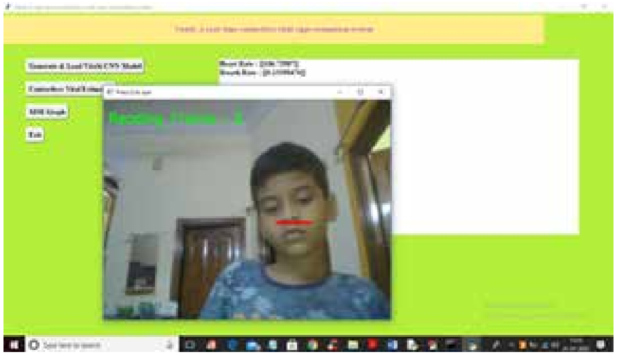
In above screen click on 'Generate & Load VitaSi CNN Model' button to generate and load CNN model and get below screen.


Generate & Load VitaSi CNN Model
Contactless Vital Estimation
MSE Graph
Exit

CNN Model loaded
Heart (BPM) CNN Model MSE Rate : 8.11
Breath Rating CNN Model MSE Rate : 2.62

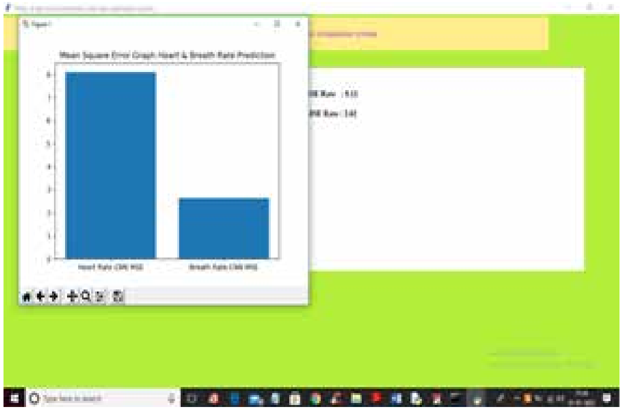
After loading the CNN model, which yielded an MSE of 8.11 for heart rate and 2.62 for breath rate, we may proceed to activate WEBCAM and make predictions for these vital signs by clicking the "Contactless Vital Estimation" button.


Generate & Load VitaSi CNN Model
Contactless Vital Estimation
MSE Graph
Exit
Reading Frame


Generate & Load VitaSi CNN Model
Contactless Vital Estimation
MSE Graph
Exit
Reading Frame

Heart Rate : [106.[illegible]]
Breath Rate : [0.13[illegible]]

It is shown and updated in the text area that the camera scans 10 frames, after which it forecasts the heart rate and respiration rate. A heart rate of 106 and a respiration rate of 0.13 are recorded here. To access the MSE graph, just choose the "MSE Graph" button.



The x-axis of the graph shows the prediction type, while the y-axis shows the mean squared error (MSE) value of the CNN prediction.

It should be noted that the programme will stop operating and produce an exception if the webcam is not visible.

CONCLUSION

This research presents a new framework for realistic video-based vital sign estimates, namely for heart rate and breathing rate. To simultaneously estimate HR and BR, the suggested technique employs a two-task convolution neural network. Using a 48-frame video clip, the colour shifts are magnified using the PVMP method. Features holding information on the HR and BR are then extracted. When analysing phase shifts of complex-valued steerables, PVMP is preferable to the EVM.

REFERENCES

1. Sörnmo, L.; Laguna, P. Bioelectrical Signal Processing in Cardiac and Neurological Applications; Academic Press: Cambridge, MA, USA, 2005; Volume 8.
2. Zhou, G.; Wang, Y.; Cui, L. Biomedical Sensor, Device and Measurement Systems. In Advances in Bioengineering; Serra, P.A., Ed.; InTech: London, UK, 2015; ISBN 978-953-51-2141-1.
3. Levick, J.R. An Introduction to Cardiovascular Physiology; Butterworths: London, UK; Boston, MA, USA, 1991; ISBN 978-0-7506-1028-5.
4. Lu, G.; Yang, F.; Taylor, J.A.; Stein, J.F. A Comparison of Photoplethysmography and ECG Recording to Analyse Heart Rate Variability in Healthy Subjects. J. Med. Eng. Technol. 2009, 33, 634–641. [CrossRef] [PubMed]
5. Evans, D.; Hodgkinson, B.; Berry, J. Vital Signs in Hospital Patients: A Systematic Review. Int. J. Nurs. Stud. 2001, 38, 643–650. [CrossRef]
6. Rohmetra, H.; Raghunath, N.; Narang, P.; Chamola, V.; Guizani, M.; Lakkaniga, N.R. AI-Enabled Remote Monitoring of Vital Signs for COVID-19: Methods, Prospects and Challenges. Computing 2021, 1–27. [CrossRef]
7. Wang, W.; Wang, X. Contactless Vital Signs Monitoring; Academic Press: Cambridge, MA, USA, 2022.
8. Hlastala, M.P.; Berger, A.J. Physiology of Respiration; Oxford University Press: Oxford, UK, 2001.
9. Sherwood, L. Human Physiology: From Cells to Systems, 9th ed.; Brooks/Cole, Cengage Learning: Boston, MA, USA, 2016; ISBN 978-1-285-86693-2.
10. Ruiz, M.; Garcia, J.; Fernández, B. Body Temperature and Its Importance as a Vital Constant. Rev. Enferm. 2009, 32, 44–52.
11. Hafen, B.B.; Sharma, S. Oxygen Saturation; StatPearls Publishing: Treasure Island, FL, USA, 2021.

Predictive and Monitoring System for Water Quality

Zeeshan Khan, Pagidimari Lakshya
B.Tech Student
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana

Sanjana S Nazare
Assistant Professor
Department of Computer Science and Engineering
CMR Technical Campus, Medchal
Hyderabad, Telangana
✉ Sanjana.cse@cmrtc.ac.in

ABSTRACT

Water is a resource that is very necessary for our daily lives. The vulnerability of source water to pollution has increased due to pollution and urbanisation. In order to protect human health and maintain the quality of water sources, a water quality monitoring system must be developed immediately. A wireless water quality monitoring system that continually monitors the quality of water held in above tanks is proposed in this research. The system is expected to be cost-effective. Some measures are considered key quality indicators for evaluating water quality. All of the information pertaining to these parameters is saved in a database on the cloud together with the time stamp. The monitored data is compared to standard, well-established standards in order to determine the water quality. The data is considered a time series since it has a timestamp. The different water quality characteristics are forecasted using a univariate non-seasonal ARIMA model. Water quality decrease might be anticipated based on the forecasted findings. Between the actual and predicted values, the model has mean square errors of 0.001 for pH, 0.076 for temperature, and 0.001 for turbidity.

KEYWORDS: *ARIMA, Real time monitoring, Turbidity.*

INTRODUCTION

As a general rule, houses and businesses often use above-ground tanks to store water for later use. Stagnant water may foster the growth of several harmful bacteria and illnesses. As it reacts with it, the acidity of the rainwater changes, rendering it unfit for human consumption and other use. The walls of the tank may get coated with harmful compounds as time goes on. When released into the open air, particulate matter has the potential to pollute it. As they sink to the bottom, these particles may alter the chemical makeup of the water. Rust, which develops in poorly maintained water collecting pipes, significantly reduces water quality. Water microbiological quality is a marker of illness. Cholera, typhoid, guinea worms, hepatitis, and schistosomiasis are only few of the infectious diseases that may be spread by water contamination. Illnesses like these might be the result of people not taking cleanliness seriously enough. The availability and quality of drinking water must be carefully considered in all matters pertaining to public health. A real-time system that monitors and updates data on water quality in real-time should be put in place to begin addressing these challenges. The data collected by the system accurately represents the water quality. With some realistic data analysis, a drop in water quality might be predicted. One way to do this is by use time-series forecasting. A technique used in statistics for forecasting future events from historical data is the Autoregressive Integrated Moving Average (ARIMA). A non-seasonal model is the way to go since it can withstand short-lived, regional trends in the data. These habits will degrade water quality over time.

LITERATURE REVIEW

The first stage in creating a system to monitor a factor's impact on water quality is to identify that factor. A number of physical and chemical characteristics, including temperature, acidity, hardness, pH, sulphate, chloride, dissolved oxygen, biological oxygen demand, and chemical oxygen demand, need to be understood in detail [1]. You should think about the budget when choosing the best measurement parameters. Implementing in hardware after software design and simulation was determined to be cost-effective in [2]. In contrast to [2], which only considers temperature, conductivity, and pH, [10] include both turbidity and temperature. Among its attractive features is the scalability it provides—the capacity to transport data wirelessly. The authors built a WSN in [3] using sensors, Xbee modules, and micro-controllers as its building blocks. When planning a system, power is an important consideration. Using the active and sleep states of each node, the wireless sensor network may achieve low power consumption, according to [4]. Power for the desperately required in [3] came from solar panels. Zigbee uses the IEEE 802.15.4 standards, and to avoid signal interference, a UHF transceiver operating at 920 MHz is used in [9]. The purpose of time series forecasting is to make predictions about future data points based on values collected in the past. Ultimately, you want to find a prediction function that works well with your data by minimising the Mean Squared Error (MSE) for every interval between your actual and forecasted values [5,6]. When it comes to time series forecasting, one popular linear model is Arima, which stands for Auto-Regressive Integrated Moving Average. In [7], the ARIMA model was used to assess the water quality of wetlands, and the total prediction error was less than 15%. Researchers achieved a relative error of 4-12% when comparing ARIMA-based coastal water quality forecasts to actual outcomes [8]. One may anticipate the changing features of drinking water by employing time series forecasting, which is based on a concept for water quality prediction. Developing a sensor system is essential for essential water quality measures. Careful consideration of power and cost trade-offs is required. The scalability and data accessibility of monitoring systems are powerful reworking components of system design.

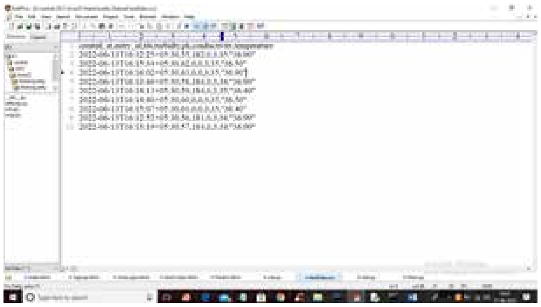
PRESENT CONDITIONS

For both commercial and domestic use, above-ground tanks are the go-to for water storage. Stagnant water may foster the growth of several harmful bacteria and illnesses. As it reacts with it, the acidity of the rainwater changes, rendering it unfit for human consumption and other use. The walls of the tank may get coated with harmful compounds as time goes on. When released into the open air, particulate matter has the potential to pollute it. As they sink to the bottom, these particles may alter the chemical makeup of the water. Rust, which develops in poorly maintained water collecting pipes, significantly reduces water quality.

Suggested System: Both the receiver's (Figure 8) and the transmitter's (Figure 9) flowcharts are shown. The system's power source may switch between solar panels and batteries on the fly to meet the power demands of the time. In response to a defined cutofffor the solar panel's power flux density, a control circuit switches between the two sources. A battery is available to mitigate power fluctuations that may happen during the day, even if solar panels provide the majority of the energy. If the transmitter goes dead, the database will display erroneous values since the user checks the quality every day. Inadequate power causes the micro-controller to become unstable, which in turn causes the values to be distorted. Figure 7 shows the schematic of the transmitter's power supply. The central processing unit (CPU) waits for the sensors to stabilise after turning on the transmitter components. Following core stabilisation, incoming data is analysed and encrypted using AES. Once the data has been encrypted, it is sent wireless via the RF module.

METHODOLOGY

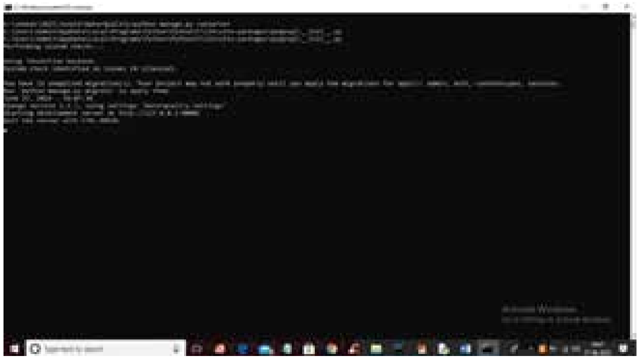
Monitoring and Forecasting System for Water Quality, Here, we're using a water dataset to make predictions and forecasts about water quality using algorithms like Random Forest, which outperforms LSTM. The next screen displays the values of the test data, which are applied to the trained model in order to forecast its quality.



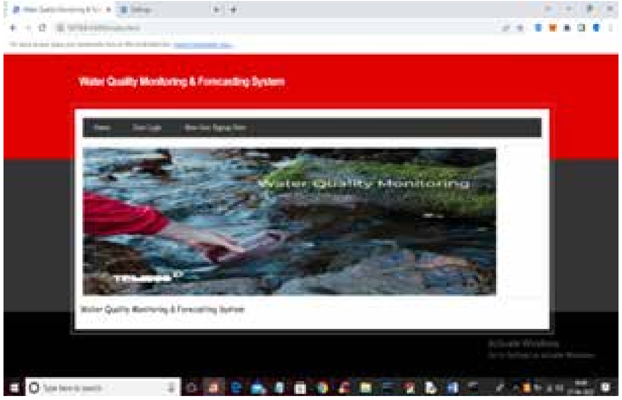
```
[illegible]
2022-06-13T06:[illegible]
2022-06-13T06:[illegible]
2022-06-13T06:[illegible]
2022-06-13T06:[illegible]
2022-06-13T06:[illegible]
2022-06-13T06:[illegible]
2022-06-13T06:[illegible]
2022-06-13T06:[illegible]
2022-06-13T06:[illegible]
```

Even if the test data screen up top doesn't include the GOOD and POOR labels for the water quality measurements, the algorithm-trained model will still use those labels to generate predictions.

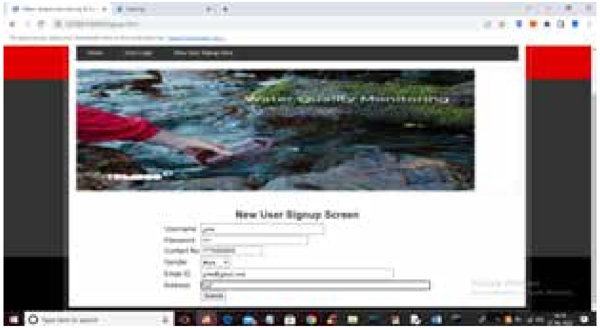
Some screenshots of the project in action are as follows: The steps to create a MySQL database are as follows: first, copy the contents of the "DB.txt" file. Then, to begin DJANGO server, double-click the "run.bat" file.



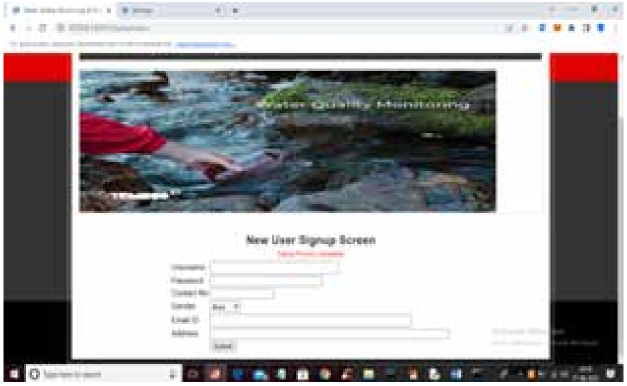
As you can see from the previous screen, the DJANGO server has started. Click the link below to visit the page: "http://127.0.0.1:8000/index.html."


Water Quality Monitoring & Forecasting System
Home
User Login
New User Signup Here
Water Quality Monitoring
Water Quality Monitoring & Forecasting System

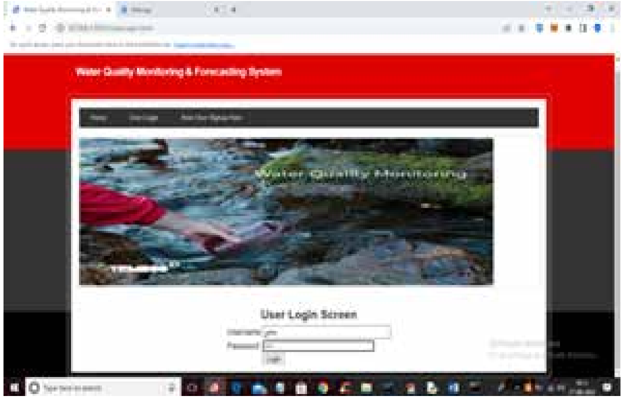
Click the "New User Signup Here" link on the previous page to access the screen below.



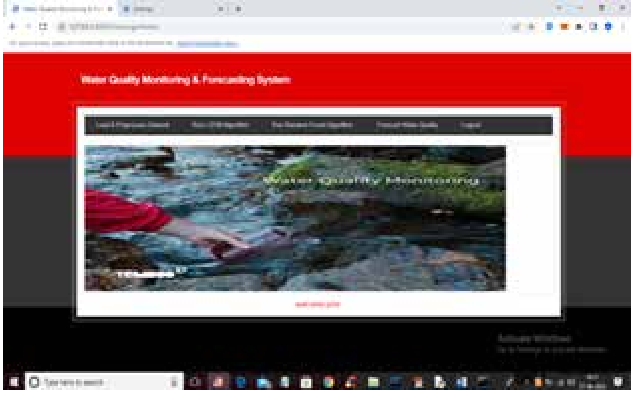
To get to the screen below, the user must first sign up on the page above and then push the button.



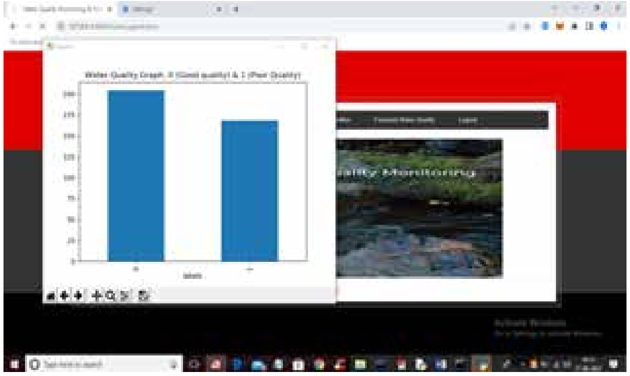
After you've finished signing up on the previous page, click the "User Login" link to go to the one below.



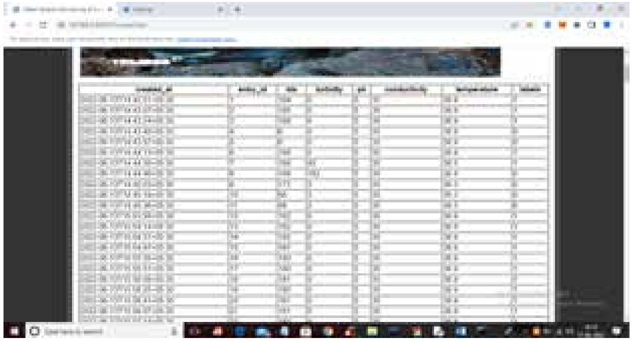
Once the user logs in, they will see the screen below.


Water Quality Monitoring & Forecasting System
Load & Preprocess Dataset
Water Quality Monitoring

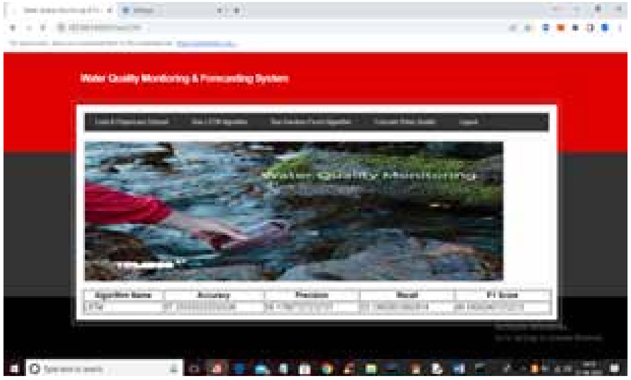
To import and prepare the dataset (i.e., fill in missing values with zeroes, divide it into a train and test set, etc.), see the "Load & Preprocess Dataset" button on the previous page. The results will be shown below.



The dataset is processed on the above screen. The x-axis displays the water quality as 0 or 1, with 0 indicating GOOD quality and 1 indicating POOR quality. The y-axis shows the number of records. Close the graph to see the processed screen below.

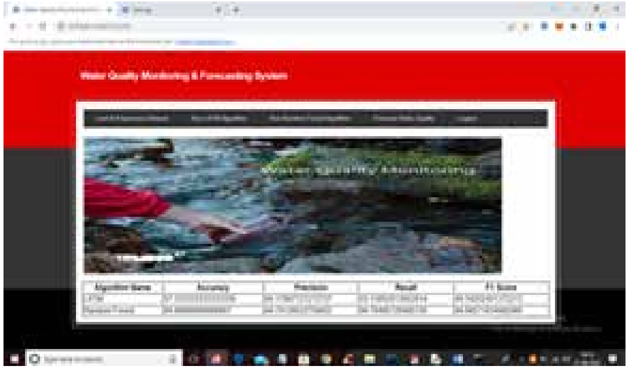


After the dataset has been processed and imported, you may train the LSTM algorithm by clicking the "Train LSTM Algorithm" link. The results will be shown below.



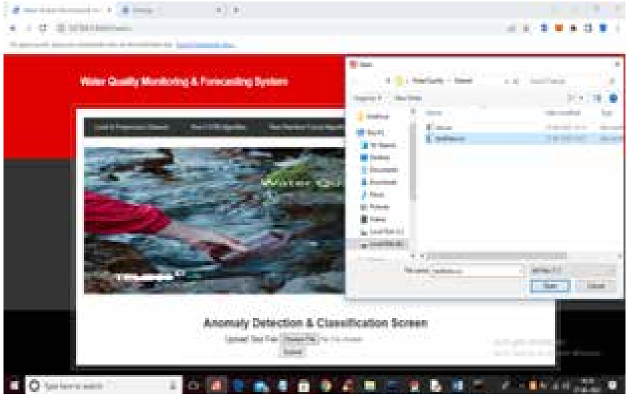
Algorithm Name	Accuracy	Precision	Recall	F1 Score
LSTM	[illegible]	[illegible]	[illegible]	[illegible]

You can see the results of training LSTM on the previous page; with LSTM, we achieved an accuracy of 57%. To train Random Forest, select the "Train Random Forest Algorithm" link.

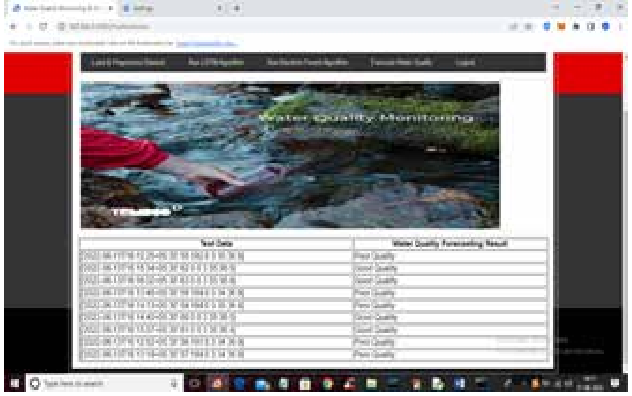


Algorithm Name	Accuracy	Precision	Recall	F1 Score
LSTM	[illegible]	[illegible]	[illegible]	[illegible]
Random Forest	[illegible]	[illegible]	[illegible]	[illegible]

Once LSTM was trained, we achieved an accuracy of 57%. To train Random Forest, click on the "Train Random Forest Algorithm" link. The result will be shown below.


Water Quality Monitoring & Forecasting System
Anomaly Detection & Classification Screen
Open
Cancel

To get the forecast result shown below, open the file named "testData.csv" on the previous screen, click the "Open" and "Submit" buttons.



Test Data	Water Quality Forecasting Result
[illegible]	Poor Quality
[illegible]	Good Quality
[illegible]	Good Quality
[illegible]	Poor Quality
[illegible]	Poor Quality
[illegible]	Good Quality
[illegible]	Good Quality
[illegible]	Poor Quality
[illegible]	Poor Quality

The top screen displays the tabular output, which includes the water test readings in the first column and the predicted results as "Good" or "Poor" in the second column.

CONCLUSION

A Bluetooth-enabled agricultural equipment with ploughing, seed-sowing, and muck-leveling capabilities has been considered for construction. Our top pick is a Bluetooth-enabled, battery-operated setup. Farmers may have a lot of fun with the robot for reasons beyond just running it. The Indian economy benefits from farmers' ability to multitask since it increases their revenue.

REFERENCES

1. Vishnu Prakash K, Sathish Kumar V, Venkatesh P, Chandran A, "Design and fabrication of multipurpose agricultural robot", International Journal of Advanced Science and Engineering Research, Volume: 1, Issue: 1, June 2016, ISSN: 2455 9288.
2. Ankit Singh, Abhishek Gupta, Akash Bhosale, Sumeet Poddar, "Agribot: An Agriculture Robot", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 1, January 2015 ISSN (Online): 22781021 ISSN (Print): 2319-5940.
3. Mr. Sagar R. Chavan, Prof. Rahul D. Shelke, Prof. Shrinivas R. Zanwar, "Enhanced agriculture robotic system", International journal of engineering sciences & research technology, ISSN: 2277-9655.
4. Nithin P V, Shivaprakash S, "Multipurpose agricultural robot", International Journal of Engineering Research, ISSN: 2319- 6890) (online), 2347-5013(print) Volume No.5 Issue: Special 6, pp: 1129 - 1254.
5. Ms. Aditi D. Kokate, Prof. Priyanka D. Yadav, "Multipurpose Agricultural Robot", International Advanced Research Journal in Science, Engineering and Technology National Conference on Emerging trends in Electronics & Telecommunication Engineering (NCETETE 2017), ISSN (Online) 2393-8021 ISSN (Print) 2394- 1588.
6. L. Manivannan, M. S. Priyadharshini, "Agricultural Robot", International Journal of Advanced Research in Electrical, Electronics and Instrumentation, Volume 5, Special Issue 1, March 2016, ISSN (Print) : 2320 – 3765, ISSN (Online): 2278 – 8875.
7. Mahesh. R. Pundkar, a seed-sowing machine a review, IJESS Volume3, Issue3 ISSN:2249.
8. Sankaranarayanan M, "Development OfAPush Type Seed Drill For Sowing Maize In Rwanda". InstitutE Supérieurd' Agricultureetd' Elevage, ISAE, Busogo, Post Box No.210, Musanze, Rwanda.
9. Rolando P, "Development of a manually operated disc-type corn seeder".
10. Ed Dager, "Proper Equipment for Small Farms" Kaitlin D'Agostino, Economics, SAS '13, Rabin, 12/03/09.
11. Parameshachari B D et. Al Optimized Neighbor Discovery in Internet of Things (IoT), 2017 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), PP 594-598, 978-1-5386- 2361-9/17/$31.00 ©2017 IEEE.

CMR TECHNICAL CAMPUS

Sponsored by CMR Technical Education Society

About CMR Technical Campus

CMR Technical Campus is an UGC Autonomous Engineering College established in the year 2009, under aegis of CMR Technical Education Society with NBA and NAAC accreditation with 'A' grade. The institute received an All India Rank of 201-250 band in National Institutional Ranking Framework (NIRF – 2022) consecutively for the second year and Excellent band in Atal Rankings of Institutions on Innovation Achievements (ARIIA-2021). The lush green Campus, situated in Kandlakoya, Medchal Road has been first choice option to both Urban and Rural students of Telangana. With well established Teaching & Learning processes, the students are imparted the best of the technological knowledge.

Chairman
Shri C.Gopal Reddy
B.E (ECE), Manipal Institute of Technology

It is rightly said by our Bharat Ratna, Dr. APJ Abdul Kalam "Dream is not that which you see while sleeping, it is something that does not let you sleep." Yes!!! Indeed, it was my dream, goal and ambition, to establish a Knowledge Center, a Temple of learning, i.e., "CMR Group of Institutions", which can be witnessed today on sprawling area of 70 acres on Medchal Road, Hyderabad. Currently under CMR Group of Institutions, which was established in the year 2002, have 04 Engineering Colleges and 01 Pharmacy College with over 16000 students studying in various streams of Engineering, Management and Pharmacy.The thirst for Excellence & Service has never subdued among us, thus we wish to make deeper inroads in the field of education by venturing into university, named "CMR Global University", which will be seeing the light of day soon.

CMR TECHNICAL CAMPUS

Sponsored by CMR Technical Education Society

Accredited by NBA & NAAC with 'A' Grade Approved by AICTE, Permanently
Affiliated to JNTUH Kandlakoya (V), Medchal Road, Hyderabad - 501 401

9100760559, 9100470559, 9247033441 info@cmrtc.ac.in

VOLUME 46 • SPECIAL ISSUE • NOVEMBER 2023

ISSN 0971-3034

PUBLISHED BY
INDIAN SOCIETY FOR TECHNICAL EDUCATION
Near Katwaria Sarai, Shaheed Jeet Singh Marg,
New Delhi - 110 016

Printed at: Compuprint, Flat C, Aristo, 9, Second Street, Gopalapuram, Chennai 600 086.
Phone : +91 44 2811 6768 • www.compuprint.in