Journal of Informatics Education and Research
ISSN: 1526-4726
Vol 5 Issue 1 (2025)

Cybercrime Analysis and Online Fraud Prevention Using Advanced Machine Learning Models

[1]Shammi L,
Assistant Professor,
East Point College of Engineering and Technology, Bengaluru, India.
shammil.cse@eastpoint.ac.in
[2]Dr C Emilin Shyni,
Professor,
School of Computer Science and Engineering,
Presidency University, Bengaluru, India. emilinshyni2020@gmail.com

ABSTRACT:
With the rise of e-commerce on a global scale, numerous companies have been able to provide goods and services through cutting-edge, time-saving, and budget-friendly business strategies. A sizable portion of the global economy is supported by the ever-expanding digital economy. Due to the simplicity of committing these crimes, the attractive returns, and the slowness of prevention measures, cyber-crimes have also increased at an even quicker rate than this relatively rapid expansion. The necessity to efficiently manage the expansion of cybercrime is critical, since it poses an existential danger to online trade. Businesses should implement creative preventive measures because laws and police resources aren't keeping up with the dynamic nature of crimes. How big and small businesses alike are implementing cyber-crime prevention techniques is the main topic of this special issue.

Keywords—Cybercrime, E-Commerce, Machine Learning (ML), Decision Tree(DT), Random Forest (RF).

I. INTRODUCTION

The term security is related to the systematic and multi-faceted idea of risk. The expansion of the Internet has improved productivity in the workplace while opening up new opportunities for criminal activity. Internet and telecommunication fraud, which includes telephone, SMS, WeChat, QQ, and other telecommunications network platforms, is on the rise, in contrast to traditional kinds of fraud. This new type of crime doesn't involve physical touch and can quickly and widely affect large groups of people, posing new hazards to society and causing victims great harm. When authentic and fraudulent transactions intermingle, advanced fraud detection systems employ data mining and sophisticated analytics to identify suspicious patterns in the logs [1]. Viewers must comprehend extensive statistics and execute binary classifications to effectively differentiate between legitimate and fraudulent or abnormal transactions. The application of machine learning, particularly peer-to-peer learning methodologies, has been highly effective in addressing this issue. For many businesses, capitalising on societal technological progress has been the go-to marketing strategy for the last ten years [2]. This new digital trend makes it possible to zero in on specific user groups and provide relevant content or items to them in a more personalised and beneficial way. Unfortunately, along with unprecedented potential and benefits, technological growth has also brought about a plethora of problems. Its enormous popularity and profitability have made it a target for dishonest individuals and organisations who want to steal users' personal information, spread malware, or cut into ad revenue [3]. The phrase online fraud denotes any form of deception conducted through digital channels, including the internet [4]. Fraudsters continually devise innovative methods to exploit deficiencies in detection systems, exacerbating the prevalence of fraud across all sectors. Online fraud, encompassing phishing, identity theft, and various other modalities, is becoming progressively challenging to

mitigate. Despite the numerous technologies available for fraud detection, con artists continually devise innovative strategies to evade them [5]. The ever advancing methods of criminals are rapidly surpassing and revealing the inadequacies of conventional strategies such as rule-based detection or manual inspections. This illustrates the inadequacy of present strategies in addressing fraud, despite their efficacy in specific cases [6]. To address the existing research gap, it is imperative to develop more adaptable and advanced systems capable of detecting and predicting fraud across many contexts. While numerous fraud detection systems can identify fraudulent activities in real-time, few have explored the potential of employing machine learning to completely eradicate fraud. This research addresses a vital requirement by examining how machine learning may improve fraud detection systems and offer a more resilient defence via real-time fraud prediction [7]. This study's innovative and previously unexamined technique enhances protection for businesses and individuals against various forms of fraud [8]. The study highlights the potential of machine learning algorithms to help organisations, organisations, and customers prevent fraud instead of just responding to it after the fact. In response to the evolving nature of fraud and the limitations of existing approaches, this forward-thinking approach proposes a more adaptable and effective solution. The application of machine learning models to actual fraud prevention strategies is a noteworthy advancement. Organisations seeking to strengthen their defences against various forms of online fraud might find crucial insights in the paper, which details the incorporation of these models into existing frameworks. This research adds to the existing body of knowledge in the academic community and provides actionable steps to strengthen security and reduce financial losses.

II. LITERATURE SURVEY

Our investigation has not identified any study that categorises financial cybercrime and its countermeasures. We strive to recognise and reference the extensive body of published research on financial cybercrime, including surveys that concentrate on specific topics. The source A comprehensive review of GNNs was carried out by [9]. They look at the applications and future of Graph Neural Networks (GNNs) and provide a taxonomy for them. Group deviation detection approaches were thoroughly examined in a survey by [10], which included both static and dynamic settings. By comparing results from studies using the Fourier and Wavelet transforms, [11] examines proactive fraud detection techniques. To avoid using past fraud cases as a basis, these research don't include fraudulent transactions. In the context of Artificial Intelligence (AI), [12] examined safe knowledge management and cybersecurity. This article examines the pros and cons of AI in the context of cybersecurity and information management. Knowledge management security measures, such as encryption and access control systems, are detailed in this document. The importance of incorporating AI approaches into security systems is emphasised by the authors. In order to guarantee the safety, authenticity, and availability of information resources, this integration aims to improve threat detection and response capabilities [13]. Utilising examples from their study, [14] addressed a comprehensive framework for cybercrime mining. This model delineated the relationship between diverse types of crime and data mining techniques such as association, prediction, visualisation, and entity extraction. This research aimed to investigate criminal organisations and their related networks. In this instance, they evaluated their approaches utilising the Coplink data. A plethora of social media-related cybercrimes were investigated by [15] Several solutions to the problem of cybercrime on social media were proposed by them. At the moment, people utilise social media for a variety of reasons, including but not limited to: spreading mass messages, organising virtual meetings, and making remote work easier [16].A wide variety of cybercrimes are defined and discussed in this book, along with methods for avoiding them. The only people who will benefit from this study are academics. It can't spot cybercrime or foresee when it will happen. The literature evaluation indicates that ML is an effective method for detecting and classifying cybercrimes. Certainly,

there exists potential for advancement in this domain as well. This work presents a machine learning technique for identifying and categorising cyber-attacks that exploit security flaws.

III. METHODOLOGY

After determining computer system criminal offense within endeavor, experts examine the styles that have been devoted over the last, and the brand new techniques probably to seem down the road. Experts also check the challenge in evaluating and discovering computer system criminal activity, procedures for trying to put on trial or even stop such criminal offenses, and the performance of these solutions. Just a little portion of computer system breather-ins are sensed, as well as, in addition, stats on

computer system unlawful act are usually not available. The most usual types of computer system criminal activity mentioned to InterGOV consist of youngster fraudulence, email, and porn spam. Buffers versus computer system illegal activity might happen in the restricting access to the relevant information to be safeguarded, using file encryption to ensure personal privacy and stability, and enlightening the social concerning safety concerns.

A. There are two main categories of cybercrime:

When an individual or group uses an IT-enabled service or device to engage in any action that is against to state law, such as spreading hate speech, Using the Dark Web for the purpose of dealing, conducting illegal business online, gambling, online betting, etc [17].
Unauthorised access to or modification of any organization's or company's data. Examples include hacking into websites and emails as well as data harvesting.

B. Possible Cybercrime Vulnerable Industries:

Cybercriminals frequently endanger public life by posing economic, social, and psychological risks through cyber threats [18]. Theft of bank funds and credit card limits are among the many things that fall under this category, as are the unlawful downloads of music files. Cybercriminals engage in monetary and non-monetary crimes for a variety of reasons, including but not limited to: social stalking, identity theft, job and marriage fraud, and other forms of fraud.

- General Public
- Business houses
- Financial Institutions
- IT companies

Thef Identity:

The fraudulent acquisition of an individual's personally identifiable information (PII) for the purpose of impersonating them is known as identity fraud or identity theft. Name, photo, phone number, bank account, credit/debit card number, tax information (PAN), driver's license, Aadhar card number, etc., are all examples of personal data.

The Detrimental Effects of Identity Theft Include:

- Create a social media account and an email account in your name.
- Unauthorised Access to Your Bank Account
- Establish a fraudulent utility account.
- Your name can be disclosed during a police arrest.
- Request for a Replacement SIM Card.
- Counterfeit Identification and Immigration Document Fabrication.

For all these reasons, and more, we will go into detail about identity theft because it is a serious cybercrime nowadays.

C. Cybercrime in ML:

As technology advances, there is an increasing concern about the potential use of machine learning in cybercrime. A branch of artificial intelligence, machine learning allows computers to acquire new skills and knowledge without being explicitly programmed to do so; this capability has found numerous uses, one of which is cybersecurity. Hackers are enhancing their attack capabilities with machine learning. For instance, they employ algorithms to detect trends and anomalies in data, which they then use to launch phishing assaults, financial crimes, and advanced persistent threats [19]. The cybersecurity industry faces a challenge with the integration of machine learning into cybercrime. Algorithms can be designed to bypass traditional security measures, and since technology is constantly evolving, it is difficult to keep up with new techniques being developed.

To combat these threats, businesses must keep up with cybercrime trends and implement security solutions based on machine learning to identify and stop cyberattacks. A more proactive and advanced strategy for cybersecurity is required, as is the significance of ongoing research and development in this area, due to the growing use of machine learning in cybercrime.

D. Traning in the Model:

1) DECISION TREES:

Classification and regression issues are both amenable to decision trees (DTs). In the same way that an ow chart uses split points from the input features to break down big judgements into a collection of simpler decisions, DTs do the same thing. Decision nodes are the nodes in a network where decisions are made. The nodes that do not undergo any further splitting are referred to as the leaf nodes.
In order to make a forecast for a regression problem, we take the mean of all the items in the leaf node. The leaf nodes represent the expected classes in a classification issue.
If you want to know how DTs make predictions, a tree diagram is a great visual aid that simplifies the explanation. Unfortunately, it's very uncommon for a single DT to make inaccurate estimates and even over-tite.

2) RANDOM FOREST:

Multiple decision trees are aggregated to make predictions in random forest (RF). When building trees using multiple bootstrap samples (i.e., samples with replacement), the bagging approach is employed. For regression aggregation, we take the mean of the predictions made by each tree; for classification, we use the trees' majority votes. Ensemble ML, of which RF is a subset, involves evaluating and then integrating multiple ML models into a single model, the result of which is frequently better prediction performance than that of the individual models.

This method's impetus is analogous to soliciting the views of numerous specialists and then tallying their votes to get a final verdict. Also using multiple DTs, the main distinction in the gradient boosting approach (XGboost) is that it builds each tree sequentially while considering the faults committed by the prior trees. Both methods significantly lessen over-taking when contrasted with the basic DT model.

3) SUPPORT VECTOR MACHINE:

When used to regression problems, support vector machines (SVMs) are commonly known as support vector regression (SVRs), but their primary use is in classification problems. Using the optimal hyperplane that maximises the margin between each class, SVM separates them. It is

possible to make the inputs linearly separable by mapping them to high-dimensional feature spaces using kernels like linear, polynomial, or radial basis function (RBF). The long training period is one of the key drawbacks of SVM. Hence, SVM might not work for bigger datasets.

4) K-NEAREST NEIGHBOR:

Even though k-nearest neighbour (KNN) is versatile and may be applied to both classification and regression, it is primarily employed for classification problems. Known as a sort of lazy learning, KNN eliminates the need for a specific training phase [20]. Predicting the value of a new data point usually involves finding its k nearest neighbours using a distance metric, most commonly, the Euclidean distance. The class with the most nearby neighbours will then get it. This procedure can also be referred to as a 3-NN method, as seen in Figure 3, when k is set to 3. So, in this case, the newitemingreen is neighboured by two items from the orange class and one item from the blue class.So, the newitemingreen will be put in the orange class.

IV. RESULTS AND DISCUSSION

Cybersecurity protocols, fraud detection systems, reaction tactics, and potential avenues for further study are all covered in this paper. This study delves into the methodological restrictions, data limitations, and difficulties in generalising findings that cyber security research encounters. Proposed areas of future study include international collaboration, the psychological and social effects of cybercrime, and the policy and legal ramifications of this issue. In order to better combat cyber threats, this research seeks to enhance the quality and relevance of cyber security research by addressing these limitations and exploring future research areas.

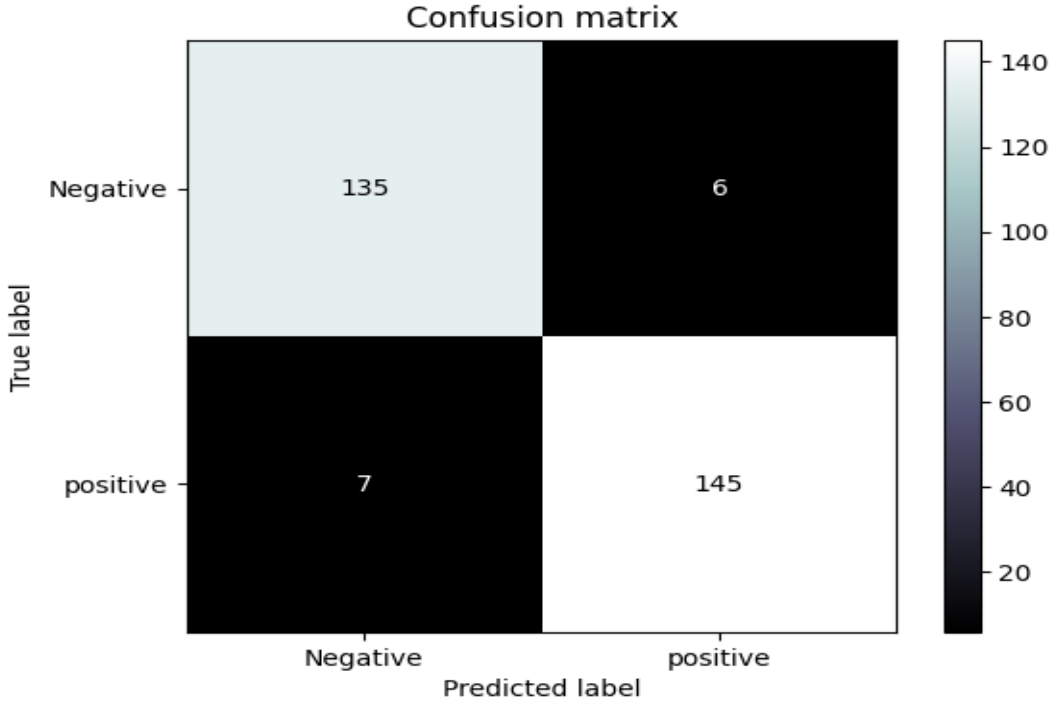

Fig. 1. Confusion Matrix for Proposed Model

There were few misclassifications (6 false positives and 7 false negatives), as seen in this matrix, indicating that the model worked adequately. The model's accuracy in classifying both types is well-balanced, as seen by the diagonal elements (135 and 145) that indicate valid predictions.

TABLE I. PROPOSED MODEL TABLE

Models	Accuracy	Precision	Recall	F1-Score	FP-Rate
DT	93.42	91.25	89.54	92.44	05.35
RF	95.32	93.67	91.62	91.55	04.18
SVM	**97.55**	**95.23**	**93.29**	**94.41**	**03.00**
KNN	90.30	89.71	87.42	87.34	05.00

Table 1 summarises the performance metrics for all four models. Examples of these models are decision trees, Random Forests, Support Vector Machines, and K-Nearest Neighbours. Compared to its rivals, SVM excels in most criteria, including Accuracy, Precision, Recall, F1-Score, and FP-Rate. The outcomes prove that support vector machines are highly effective in tasks requiring accuracy and predictability.

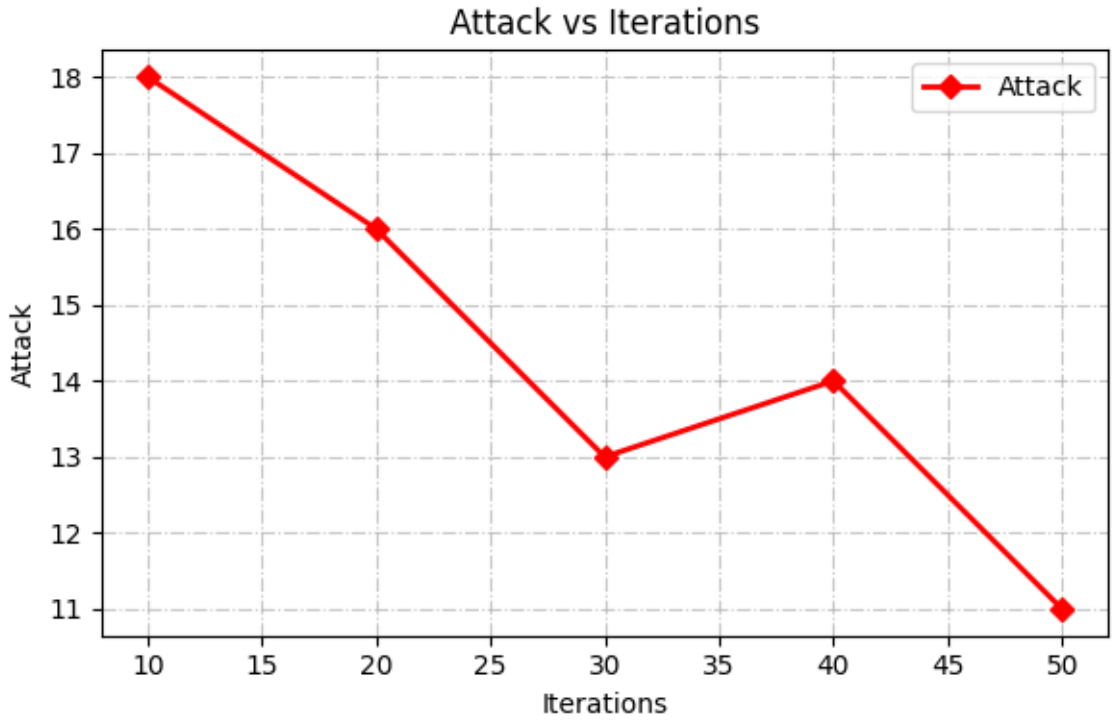

Fig. 2. Attack of the Proposed Model

Attack vs Iterations is a graphical representation of the correlation between iteration count and attack value. Attack numbers range from 11 to 18, while iterations range from 10 to 50 on the X-axis. Connected by a red line to emphasise the pattern, the red diamonds show particular attack values at different iterations. At 10 iterations, the attack value is 18, and by 30 iterations, it has decreased to 13. There is a small rise after 40 iterations, and then a last decrease to 11 at 50 iterations. This pattern indicates that the attack values generally get better or optimised with each iteration, with just minor variations along the way.

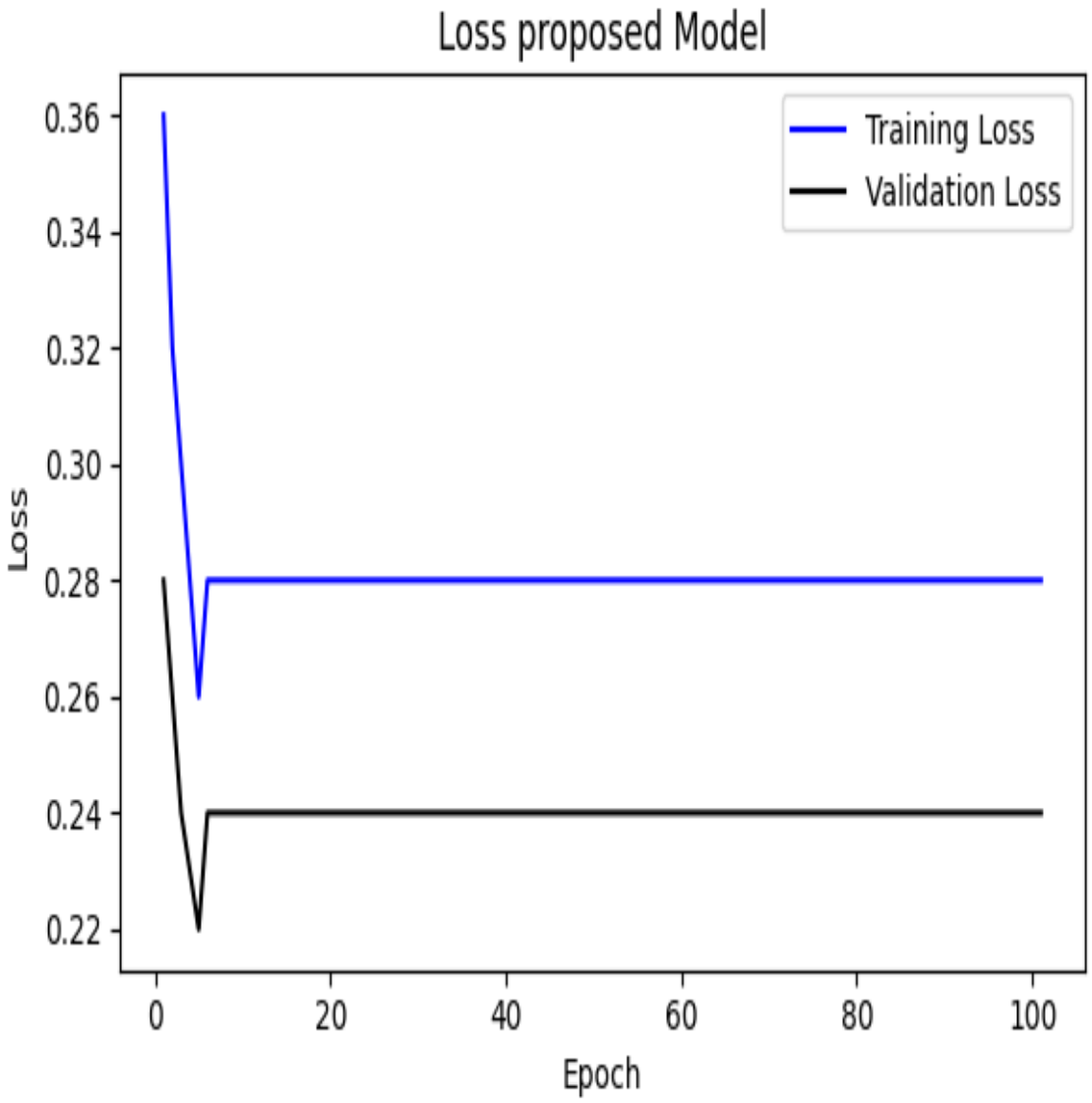

Fig. 3. Training and Validation Loss of the proposed model

Over the course of 100 iterations, the training and validation losses are shown in the chart under Loss Proposed Model. The epochs are shown on the X-axis, which goes from 0 to 100, and the loss values are shown on the Y-axis. The training loss, shown by the blue line, spikes at about 0.36 in the first epochs before falling sharply and then stabilising at about 0.28. Starting at around 0.22, the validation loss (shown by the black line) varies slightly in the early epochs until stabilising near 0.24. Both losses converge, as seen in the graph, indicating that the suggested model reaches a steady state following early fluctuations; this indicates good generalisation with minimal overfitting.

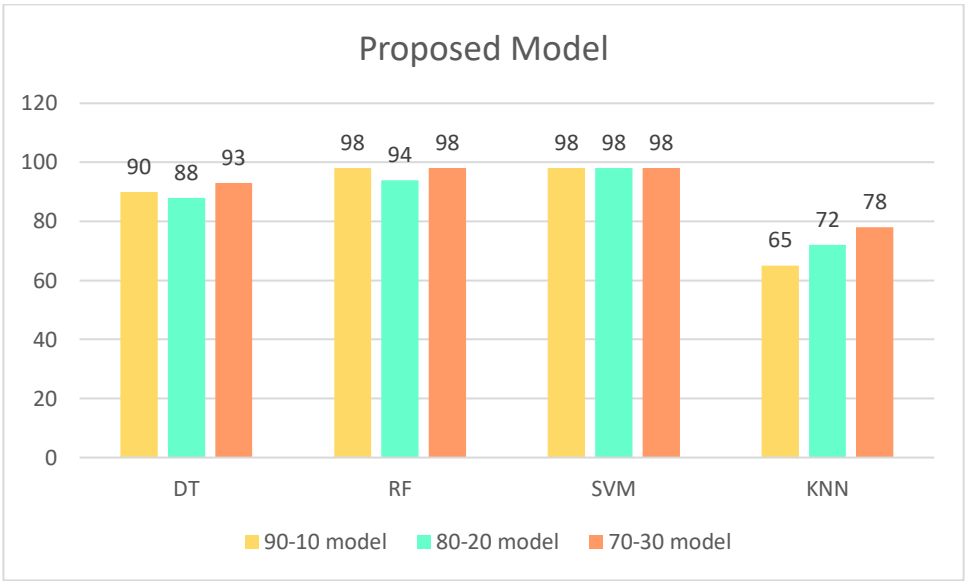

Fig. 4. Accuracy of the Proposed Model

Fig. 5.

This bar chart shows the results of a model evaluation utilising Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbours (KNN) on three data split configurations: 80/20, 70/30, and 90/10. The algorithms are shown on the x-axis, and performance measurements (likely accuracy or something similar) are shown as percentages on the y-axis.

V. Conclusion and Future Directions

The data explosion is being propelled by the ubiquitous Internet and the widespread adoption of computing and mobile devices. Fascinatingly, fraudsters are also taking advantage of these widely used technology to commit cybercrime, such as fraud, scams, and bogus internet reviews. Although current content filtering methods have been effective in reducing spam, they are ill-equipped to handle these emerging forms of cybercrime, as their perpetrators devise novel ways of sending text messages to evade detection. To identify certain types of text-based cybercrime, we combine natural language processing with a discourse on deception detection to create hybrid models. To train the hybrid models to detect more than one kind of cybercrime, we mix two datasets and three datasets, since each of the four datasets contains both truthful and misleading text messages representing different types of cybercrime. The models that combine ANN, NB, SVM, and k-Nearest Neighbour (kNN) for cybercrime detection are trained by this process. The models are subsequently tested using test sets that include cases that were not included in the training sets. This study compares the ANN classifier's results with those of NB, kNN, and SVM. Detecting cybercrime is a task that most models handle effectively in a generalised sense.

REFERENCES

[1] R. Udayakumar, A. Joshi, S. S. Boomiga, and R. Sugumar, “Deep Fraud Net: A Deep Learning Approach for Cyber Security and Financial Fraud Detection and Classification,” J. Internet Serv. Inf. Secur., vol. 13, no. 4, pp. 138–157, 2023, doi: 10.58346/JISIS.2023.I4.010.

[2] J. Nicholls, A. Kuppa, and N. A. Le-Khac, “Financial cybercrime: A comprehensive survey of deep learning approaches to tackle the evolving financial crime landscape,” IEEE Access, vol. 9, pp. 163965–163986, 2021, doi: 10.1109/ACCESS.2021.3134076.

[3] R. Ch, T. R. Gadekallu, M. H. Abidi, and A. Al-Ahmari, “Computational system to

classify Cyber Crime offenses using machine learning," Sustain., vol. 12, no. 10, 2020, doi: 10.3390/SU12104087.

[4] P. Ni and Q. Wang, "Internet and Telecommunication Fraud Prevention Analysis based on Deep Learning," Appl. Artif. Intell., vol. 36, no. 1, 2022, doi: 10.1080/08839514.2022.2137630.

[5] T. Porkodi, "An Automatic ATM Card Fraud Detection Using Advanced Security Model Based on AOA-CNN- XGBoost Approach," 2024 Int. Conf. Electron. Comput. Commun. Control Technol., pp. 1–7, 2024, doi: 10.1109/ICECCC61767.2024.10593851.

[6] S. Vii, G. D. Rede, P. Ramesh, R. Kumar A, A. Bharathi, and M. C. J. Anand, "Optimizing E-Commerce Fraud Detection with BiGRU and Capsule Network Architectures," in 2nd IEEE International Conference on Data Science and Network Security, ICDSNS 2024, 2024, pp. 1–6. doi: 10.1109/ICDSNS62112.2024.10691229.

[7] M. Penipuan Keuangan dan Kejahatan Dunia Maya and S. Literatur yang Sistematis Mardiana Ruslan, "Mitigating Financial Fraud and Cybercrime: A Systematic Literature Study," Account. Stud. Tax J., vol. 1, no. 4, pp. 258–273, 2024, [Online]. Available: https://journal.ppipbr.com/index.php/count/index

[8] S. Balachandar, S. Mann, A. Thangam, V. R. Dubey, V. Bindu, and N. Nishant, "Predicting Student Dropout Rates in Massive Open Online Courses Using an Attention-Based GCNN Model," 1st Int. Conf. Electron. Comput. Commun. Control Technol. ICECCC 2024, pp. 1–6, 2024, doi: 10.1109/ICECCC61767.2024.10593955.

[9] A. J. Chinchawade, M. Kalaiselvi, T. S. Kumar, Y. Singh, S. Srisathirapathy, and G. C. Babu, "Anomaly Detection of Smart City in IoT Cyberattacks based on Hybrid A - BiLSTM - CRF Method," 2023 Int. Conf. Sustain. Commun. Networks Appl., no. Icscna, pp. 430–435, 2023, doi: 10.1109/ICSCNA58489.2023.10370152.

[10] L. Shammi and C. E. Shyni, "A Novel Approach for Effective Detection and Prediction of Sophisticated Cyber Attacks using the Stacked Attention GRU and BiLSTM," 2024 Int. Conf. Electron. Comput. Commun. Control Technol., pp. 1–6, doi: 10.1109/ICECCC61767.2024.10593866.

[11] N. Penchalaiah, A. Kumar, A. Chaithrashree, A. Saivaraju, V. Bhoopathy, and A. K. Lamba, "Advanced Detection of Cyber-Physical Attacks in Manufacturing using LSTM-KNN," 2024 Int. Conf. Data Sci. Netw. Secur., pp. 1–7, 2024, doi: 10.1109/ICDSNS62112.2024.10690961.

[12] A. Kumar, T. Vyas, S. Ahmed, N. Girdharwal, E. Vijayakumar, and A. Thangavelu, "Security and Privacy Enabled Framework for Online Social Networks using Blockchain," 2023 4th Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2023 - Proc., pp. 641–647, 2023, doi: 10.1109/ICESC57686.2023.10193119.

[13] M. Jaiswal, "Cybercrime Categories and Prevention," Int. J. Creat. Res. Thoughts (IJCRT), ISSN2320-2882, vol. 7, no. 1, pp. 526–536, 2019, [Online]. Available: https://papers.ssrn.com/abstract=3946662

[14] R. Rajkumar, N. Kogila, S. Rajesh, and A. R. Begum, "Intelligent System for Fraud Detection in Online Banking using Improved Particle Swarm Optimization and Support Vector Machine," in 2023 8th International Conference on Communication and Electronics Systems (ICCES), Jun. 2023, no. Icces, pp. 644–649. doi: 10.1109/ICCES57224.2023.10192690.

[15] K. Prabhakar, M. S. Giridhar, A. Tatia, T. M. Joshi, S. Pal, and U. S. Aswal, "Comparative Evaluation of Fraud Detection in Online Payments Using CNN-BiGRU-A Approach," Int. Conf. Self Sustain. Artif. Intell. Syst. ICSSAS 2023 - Proc., no. Icssas, pp. 105–110, 2023, doi: 10.1109/ICSSAS57918.2023.10331745.

[16] N. Pol and S. Agarwal, "Online Transaction Fraud Detection : Exploring the Hybrid SSA-TCN-BiGRU Approach," 2024 2nd World Conf. Commun. & Comput., pp. 1–6,

2024, doi: 10.1109/WCONF61366.2024.10692254.

[17] A. V. Mbaziira and D. R. Murphy, “An empirical study on detecting deception and cybercrime using artificial neural networks,” ACM Int. Conf. Proceeding Ser., pp. 42–46, 2018, doi: 10.1145/3193077.3193080.

[18] C. S. Biswal and S. K. Pani, “Cyber-Crime Prevention Methodology,” Intell. Data Anal. Terror Threat Predict. Archit. Methodol. Tech. Appl., no. December, pp. 291–312, 2021, doi: 10.1002/9781119711629.ch14.

[19] E. Ramirez-Asis, R. Penadillo-Lirio, W. Acosta-Ponce, R. Norabuena-Figueroa, N. Ramírez-Asís, and P. S. Arbune, “Investigating the Intersection of Cybercrime and Machine Learning: Strategies for Prevention and Detection,” Int. Conf. Innov. Data Commun. Technol. Appl. ICIDCA 2023 - Proc., no. February, pp. 203–209, 2023, doi: 10.1109/ICIDCA56705.2023.10099631.

[20] S. Shahriar, A. R. Al-Ali, A. H. Osman, S. Dhou, and M. Nijim, “Machine learning approaches for EV charging behavior: A review,” IEEE Access, vol. 8, pp. 168980–168993, 2020, doi: 10.1109/ACCESS.2020.3023388.