TRINETRA: An Ensemble Model for Cybercrime Text Classification with Comparative Evaluation of Machine Learning Approaches

Sukrati Agrawal
Research Scholar
SAGE University, Indore
Department of Computer Science Engineering
0009-0008-9682-7702

Hare Ram Sah, PhD
Professor
SAGE University, Indore
Department of Computer Science Engineering
0000-0003-3640-0869

Rajesh Kumar Nagar, PhD
Associate Professor
SAGE University, Indore
Department of ECE, IET
0000-0002-7805-4426

ABSTRACT
Nowadays, cybercrime has been soaring, which has necessitated the use of both automatic and efficient methods of identifying and classifying possible cases. This paper, in which we utilize text from the complaints section of electronic magazines, news stories, incident reports, and evaluations of proceedings from regulatory establishments, provides a machine learning-based method for cybercrime classification. Vectorization using Term Frequency-Inverse Document Frequency (TF-IDF) is used to transform the dataset after it is preprocessed with Natural Language Processing (NLP) techniques. Machine Learning (ML) models like Random Forest (RF) , Gradient Boosting (GB), and an Ensemble Classifier, together with the proposed Trinetra framework, are selected for evaluation using standard performance metrics. In this study, the potential of the online version of the cybercrime detection system using automatic techniques was demonstrated by the amount of accuracy, such as the proposed approach Trinetra showed higher accuracy than the others.

General Terms
Machine Learning (ML), Natural Language Processing (NLP).

Keywords
Cybercrime, Cybercrime classification, Random Forest, Gradient Boosting, Ensemble Classifier, Trinerta.

1. INTRODUCTION
The rapid growth of the internet and the digitization of business operations have led to a rise in cybercrime [1]. It has become increasingly important to digital platforms, and it is being used to conduct criminal investigations. Efficient approaches are necessary for law enforcement authorities and cybersecurity corporations to precisely identify the sorts of cybercrime. Like conventional procedures, they largely rely on manual processes that lead to an increase in the count of pending cases as depicted in Figure 1.

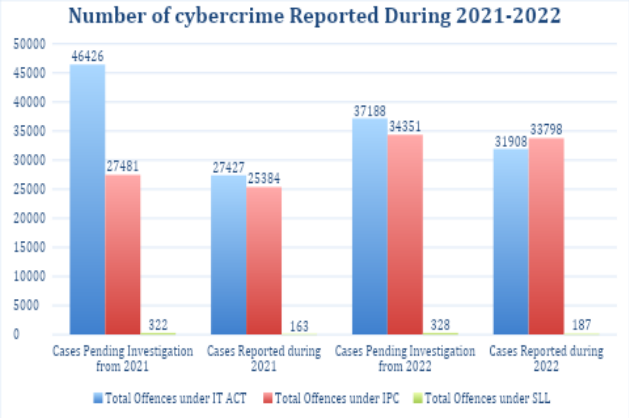

Fig. 1. Number of Cybercrime Pending from and Reported During 2021- 2022: Data Source[2]

This figure 1 clearly depicts that specific research is needed to improve current systems using machine learning-based algorithms for crime classification and reducing the delay time of solving cases. Conducting legal research is a laborious and intricate process that necessitates a thorough comprehension of legal terminology and ideas. To support attorneys and legal experts in this method, a ML and NLP can be used to create an AI-based legal aid system [3].

Cybercrime is an ever-growing challenge, with victims often experiencing frustration due to slow legal processes. Complaints exceeding a certain threshold (e.g., financial losses above Rs. 2 Lakh) are generally handled by specialized cybercrime cells, while smaller cases are dealt with at local police stations. There are several bottlenecks in the current system, including slow processing times and insufficient user awareness. Implementing an autonomous complaint-processing system could accelerate justice, effectively categorize complaints, and significantly streamline the initial processing stage.

The manual processing of cybercrime claims sometimes results in delays due to investigation limitations [4]. Given the rise in cybercrimes, law enforcement agencies need to employ AI-driven solutions to efficiently categorize, assess, and handle complaints. According to research, self-governing systems can manage enormous volumes of data while maintaining accuracy and reducing the need for human intervention [3]. Presently, state police handle internet complaints in a laborious manner which is time-consuming. To save time and effort on the present system and enable police to capture criminals more

quickly and the victim get justice timely, a new system seeks to automate the cybercrime categorization process [1].

The potential of a cybercrime complaint system to categorize cases according to jurisdiction, offense type, and severity is an essential feature. Using AI and NLP to examine complaint narratives, existing frameworks have tried automatic classification [5]. By improving complaint registration and processing speed [4], ML-driven legal support systems can help close the gap between victims and law enforcement. This suggests that more sophisticated analytical tools utilizing ML techniques [6] are required to adequately handle the escalating problem of cybercrime [7].

By identifying trends and correctly categorizing crimes, ML-powered systems can handle cybercrime data. According to research, crime detection and response are greatly enhanced when cyber threat intelligence is combined with machine learning models [3].

The legal system often struggles with a backlog of cybercrime cases. ML-based systems can alleviate this by automating preliminary investigations and linking complaint data to relevant case laws [5].

2. RELATED WORK

Cybercrime encompasses a wide range of human driven or machine-driven unauthenticated activities conducted through various digital platforms, broadly including computer related crime, violation of confidentiality or privacy, cyber terrorism, sharing or transmission of obscene messages and cyber harassment.

As per the NCRB data [2] it has been found that there is a clear increasing trend in cybercrime incidents over the years. States like Uttar Pradesh, Maharashtra, Karnataka, Telangana, and Andhra Pradesh report the highest number of cases, indicating a significant concentration of cyber-related offenses in these regions. In contrast, smaller states and Union Territories, such as Sikkim, Arunachal Pradesh, and Nagaland, have comparatively lower numbers as shown in figure 2.

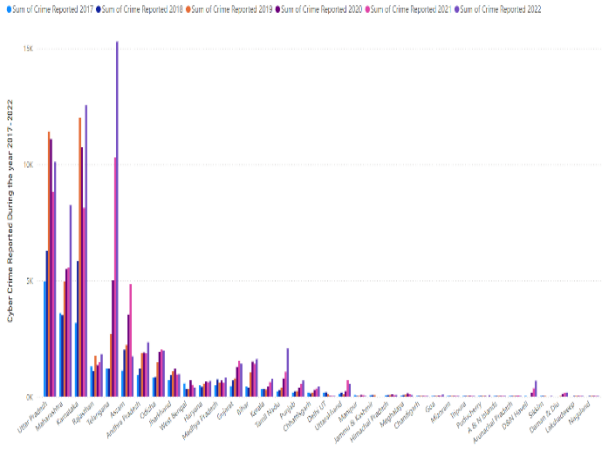

Fig. 2. Rising Cybercrime Trends (2017-2022) in India State-wise: Data Source [2]

The data suggests that cybercrime has been rising continuously, reflecting the growing demand of ML based systems to combat these cybercrimes.

According to research by Aviral Goyal et al. [8] on the motivations behind cybercrime, the most prevalent kind of cybercrime would be cheating by personation utilizing computer resources, and the hotspot will be Telangana followed by Karnataka, Maharashtra, and Uttar Pradesh.

G. Maheswari et al. [9] reported that the lack of crime reporting systems in developing countries like Pakistan that leads to lack of structured crime data.

According to Rupa Ch et al. [7], an increase in cybercrime is caused by a lack of computational techniques for anticipating cybercrime.

Text-based cybercrime classification involves analyzing textual data from emails, social media or any other digital sources to detect and categorize malicious intent. New cyber security measures are required to combat cybercrime. [10]. Several studies have explored ML and NLP-based approaches for detecting and categorizing cybercrime.

Ankit Bansal et al. [10] suggested a hybrid strategy based on intelligent dynamic malware analysis to provide cyber security measures utilizing decision tree categorization and conduct real-time action to prevent cybercrime.

Zaher Salah et al. [11] studied an ensemble approach to detect phishing emails resulting in F1-score of scores 0.90 using soft voting and weighted ensemble learning, 0.85 with different methods.

Shridevi Soma et al. [12] classified cybercrimes into cyberbullying and IP fraud analysis, using RF machine learning algorithms on the Twitter data.

Oluwatoyin Esther Akinbowale et al. [13] questionnaire 17 bank officials of South Africa to rank cybercrime impacts in the financial sector for better mitigation decisions.

I Amin et al. [14] explores the use of NLP to combat cybercrime in Bangladesh by enhancing threat detection, incident response, and proactive mitigation.

T Arjunan et al. [15] reviews the use of NLP techniques for detecting anomalies in unstructured data for cybersecurity.

F Ullah et al. et al. [16] addresses the challenge of cybercrime detection in vernacular languages, especially low-resource languages.

An overview of Natural Language Processing (NLP) methods for identifying irregularities and intrusions in cybersecurity utilizing unstructured data is given by S. Sharma et al. [17]. It covers important techniques that include categorization of document, topic modeling, named entity recognition, and analysis of sentiments.

D Srinivas et al. [18] uses ML and NLP techniques to employ RF, Naive Bayes, and SVM to identify social media crimes like threats, harassment, and cyberbullying to combat online abuse.

The very essence of transforming unstructured data into structured data has been emphasized for better cybercrime analysis and classification in many studies. Information technology now helps law enforcement in processing reported cases at greater speeds by utilizing the potential of ML and NLP, thereby minimizing delays in the delivery of justice. Swift and effective punishment also serves to deter potential offenders by showing them the ramifications of cybercrime as well as securing the conviction of those guilty of cybercrime, thus aiding in the resolution of outstanding cases.

In the era of digitalization, as more cybercrime is being reported every day, analyzing reported cybercrimes requires efficient text-based classification. Due to growing incidence of cyber risks in digital communication, text-based cybercrime

classification has recently become a highly investigated topic of concern. Such endeavors have seen researchers work with several techniques such as ML, NLP, and deep learning to effectively detect and classify cybercrimes.

3. PROPOSED METHODOLOGY

A framework for classifying cybercrime data is proposed in this study, which is shown in Figure 3, aimed at categorizing relevant data sources to identify emerging IT threats. The key steps in this process are as follows:

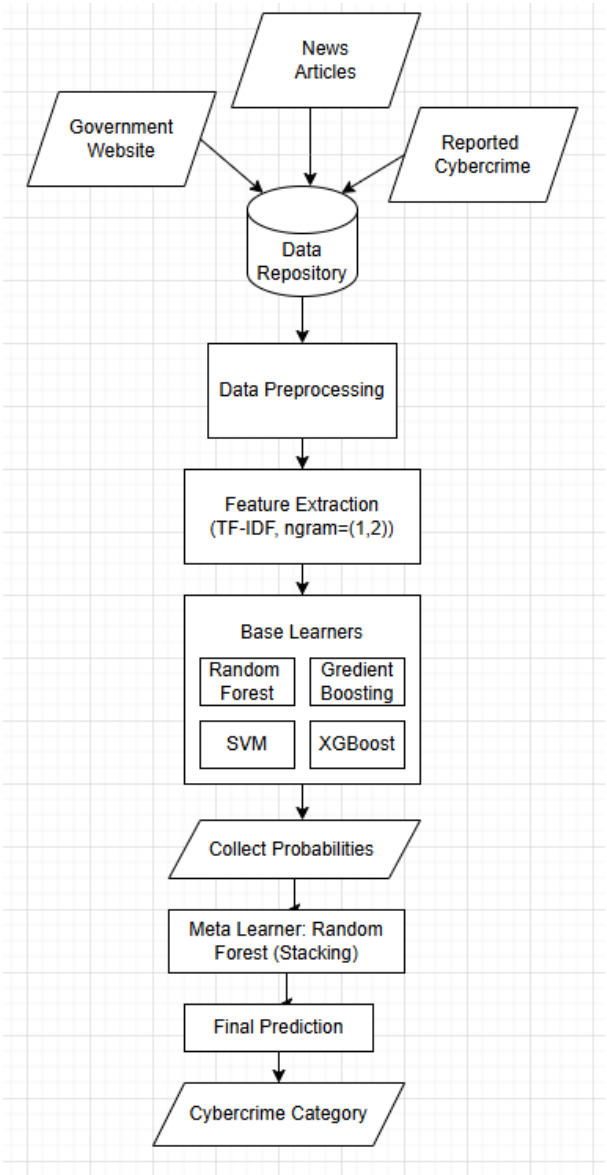

Fig. 3. Flow chart of proposed model

3.1 Dataset Collection

The dataset consists of cybercrime complaints extracted from various sources, including news articles, official reports, and user-submitted cases. The dataset is preprocessed to remove inconsistencies and standardize text inputs.

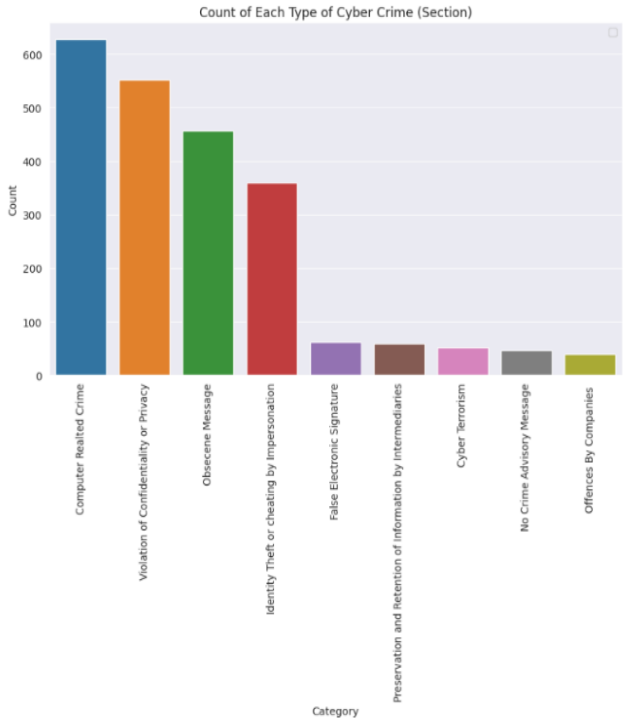

Fig. 4. Classification of Cybercrime Data

3.2 Data Preprocessing

Initially, all letters in the reported cybercrime are lowercase since it represents uniformity with no case difference. After this, removal of punctuations, numbers, and special signs from the text was done through regular expressions leaving behind only reasonable words for the analysis. The next step in this process involved the removal of stop words, mostly common words that do not add significant meaning with the help of the NLTK library. A process called tokenization was then applied that involves processing the text to separate out individual words or tokens, which could be useful for feature extraction. Finally, TF-IDF vectorization is applied to convert the textual data into numerical representations that reflect how important each word is concerning the entire dataset and therefore makes the classification effective.

3.3 Machine Learning Models

The study evaluates three classification models:

3.3.1 Random Forest (RF): A RF is an ensemble learning algorithm that creates a multitude of decision trees to counteract the reliability of the prediction with an overfitting effect. In finality, it averages the predictions from many trees to arrive at a single decision [7,19].

3.3.2 Gradient Boosting (GB): GB is a robust ensemble learning method that constructs trees sequentially so that every tree is able to rectify the mistakes of the previous tree. The approach optimizes the loss function by employing a gradient descent process which makes it appropriate for complicated datasets [19,20].

3.3.3 Ensemble Voting Classifier (EVC) : In this study, an ensemble classifier aggregates predictions from multiple models, including RF and GB. The final prediction uses a soft voting mechanism, thereby improving the predictions regarding precision and generalization [21,22].

3.3.4 Proposed Algorithm Trinetra: In proposed study Trinetra, employs a stacked ensemble architecture that integrates multiple heterogeneous classifiers, including RF, GB, Linear SVM, and XGBoost. Unlike conventional soft-voting ensembles, Trinetra generates meta-features from the probability outputs of all base learners and forwards them to a RF meta-classifier for final decision making. This stacked learning strategy enhances classification robustness by effectively capturing complementary patterns across models, thereby improving precision, reducing misclassification, and achieving stronger generalization across diverse cybercrime categories.

3.4 Performance Metrics

The accuracy, recall, precision, F1-score, and AUC of the three models have been assessed. With an emphasis on overall accuracy, significance of positive predictions, ability to detect all positive cases, balance between precision and recall, and the model's ability to discriminate across classes, these metrics offer a thorough evaluation of model performance. The assessment facilitates the comparison of each model's resilience and efficacy in classifying cybercrimes.

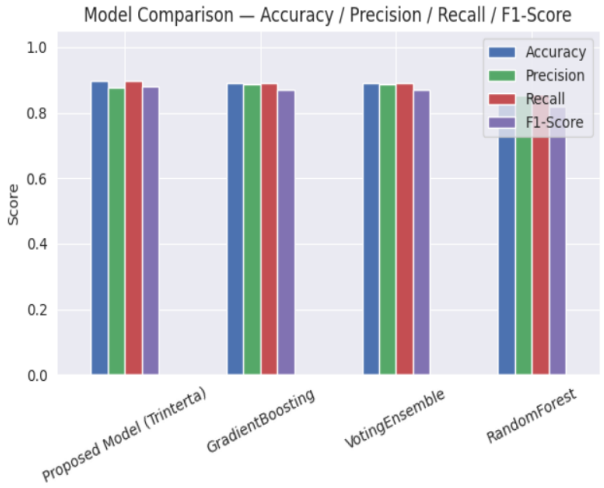

Fig. 5. Performance measure of different models used

4. Experimental Results & Discussion

Training accounts for 80% of the dataset, whereas testing accounts for 20% [20]. The findings are summarized in the following table:

4.1 Accuracy

It is the measure of ratio of instances predicted correctly to the total number of instances. In this research it has been found that Trinetra shows the best accuracy of 0.8968 calculated using mathematical expression:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

4.2 Precision

It shows the ratio of number of true positive predictions to all positive predictions. In this research it has been found that GB and EVC shows the best precision of 0.8883 calculated using mathematical expression:

$$\text{Precision} = (TP / (TP + FP)) \quad (2)$$

4.3 Recall

It is the ratio of true positive predictions to all actual positives. In this analysis it has been found that Ensemble Classifier shows the best recall of 0.8968 calculated using mathematical expression:

$$\text{Recall} = TP / (TP + FN) \quad (3)$$

4.4 F1- Score

It is the measure of harmonic mean of recall and precision. In this experimentation it has been found that Ensemble Classifier shows the best F1- Score of 0.8822 calculated using mathematical expression:

$$\text{F1- score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

4.5 AUC

Measures the ability of a model to distinguish between classes. In this study, it has been found that Trinerta shows the best AUC of 0.9778.

Table 1 shown below compares the performance of the Trinerta with various baseline classifiers. The table displays each model's performance in classifying cybercrimes using key evaluation metrics, such as Accuracy, Precision, Recall, F1-Score, and AUC.

Table 1: Comparing the Performance of Various Models

Model	Accuracy (in decimal)	Precision (in decimal)	Recall (in decimal)	F1-Score (in decimal)	AUC
RF	0.8581	0.8538	0.8581	0. 8191	0.9652
GB	0.8903	0.8883	0.8903	0.8694	0.9643
EVC	0.8903	0.8883	0.8903	0.8694	0.9727
Trinterta	**0.8968**	0.8789	**0.8968**	**0.8822**	**0.9778**

As per the table 1, it has been concluded that the best-performing model based on evaluation metrics is Trinerta.

5. DISCUSSION

The heatmap clearly shows that in every significant evaluation metric, the Proposed Model (Trinerta) outperforms the baseline models. When it comes to correctly identifying a range of cybercrime categories, Trinerta has the best recall (0.897) and accuracy (0.897). This increase is caused by its stacked ensemble approach, which combines RF, GB, SVM, and XGBoost to generate richer meta-features and more accurate predictions.

Despite having competitive precision and AUC values, both GB and the soft Voting Ensemble perform poorly in recall and F1-Score consistency. RF does the worst overall, especially when it comes to F1-Score (0.819), which indicates that it has trouble handling linguistically diverse and unbalanced complaint data.

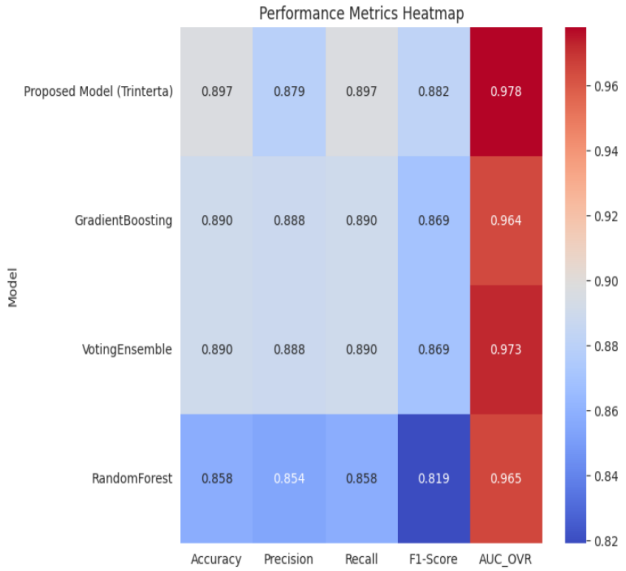

Fig. 6. Performance Metrics Heatmap

6. CONCLUSION AND FUTURE WORK

The rapid rise in reported cybercrime has created an urgent need to transition from traditional, manual text complaint processing approach to AI-driven automated approach. Conventional approaches often result in delays, workload overload, and inconsistent processing, which negatively impact timely justice delivery. By integrating machine learning and intelligent automation, complaint categorization and analysis can be performed swiftly, efficiently, and with minimal human intervention. This study demonstrates that machine learning models can effectively classify cybercrime complaints into relevant categories, supporting faster decision-making and improved case handling.

Among the various models evaluated, the ensemble classifier achieved the highest accuracy, highlighting the strength of hybrid learning techniques over standalone algorithms. This indicates that combining the strengths of multiple classifiers leads to enhanced predictive power and more reliable categorization performance.

For future research, there is significant scope for improvement in terms of both accuracy and legal applicability. Future studies may focus on advanced deep learning architectures, particularly transformer-based models, which have shown strong performance in natural language understanding. Additionally, incorporating legal rule-based knowledge, multilingual processing, and real-time system integration with law enforcement databases may further improve the system's practical relevance and usability. Building explainable AI frameworks will also be essential to ensure transparency and compliance with legal standards, ultimately enabling automated systems to support faster, fair, and accountable cybercrime response in the real world.

7. ACKNOWLEDGMENTS

I express my sincere gratitude to my guide, institution, and all those who supported and encouraged me throughout this research work. Their guidance and cooperation have been invaluable in the successful completion of this study.

8. REFERENCES

[1] Prabhu, A. V., Jefiya, M. J., Joseph, J. D., Sunny, T., & Abraham, C. M. (2023, January). Cyber Complaint Automation System. In 2023 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA) (pp. 1-5). IEEE.

[2] Open government data (OGD) platform India. (2022b, January 21). Gov.In. https://www.data.gov.in/resource

[3] Kandula, A. R., Tadiparthi, M., Yakkala, P., Pasupuleti, S., Pagolu, P., & Potharlanka, S. M. C. (2023, November). Design and Implementation of a Chatbot for Automated Legal Assistance using Natural Language Processing and Machine Learning. In 2023 Annual International Conference on Emerging Research Areas: International Conference on Intelligent Systems (AICERA/ICIS) (pp. 1-6). IEEE

[4] Jian, J., Chen, S., Luo, X., Lee, T., & Yu, X. (2020). Organized Cyber-Racketeering: Exploring the Role of Internet Technology in Organized Cybercrime Syndicates Using a Grounded Theory Approach. IEEE Transactions on Engineering Management, 69(6), 3726-3738.

[5] Diab, M. F. S. (2024). Criminal Liability for Artificial Intelligence and Autonomous Systems. American Journal of Society and Law, 3(1), 14-18.

[6] Jonas, D., Yusuf, N. A., & Zahra, A. R. A. (2023). Enhancing security frameworks with artificial intelligence in cybersecurity. International Transactions on Education Technology, 2(1), 83-91.

[7] Ch., R., Gadekallu, T.R., Abidi, M.H., & Al-Ahmari, A.M. (2020). Computational System to Classify Cyber Crime Offenses using Machine Learning. Sustainability.

[8] Goyal, A., & Gnanasigamani, L. J. (2023, July). Cybercrime Analysis of India Using Machine Learning. In International Conference on Data Science and Applications (pp. 131-145). Singapore: Springer Nature Singapore.

[9] Maheswari, G., Felixa, S., Vidhya, A. J., Sambath, M., & Remya, K. (2024, April). Forecasting of Crime Hotspots and Analysis using Machine Learning. In 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS) (pp. 01-05). IEEE.

[10] Bansal, A., Athavale, V. A., Saluja, K., Gupta, S., & Kukreja, V. (2022, September). Computational System Based on Machine Learning with Hybrid Security Technique to Classify Crime Offenses. In International Conference on Emergent Converging Technologies and Biomedical Systems (pp. 237-248). Singapore: Springer Nature Singapore.

[11] Salah, Z., Abu Owida, H., Abu Elsoud, E., Alhenawi, E., Abuowaida, S., & Alshdaifat, N. (2024). An Effective Ensemble Approach for Preventing and Detecting Phishing Attacks in Textual Form. Future Internet, 16(11), 414.

[12] Soma, S., & Mehvin, F. (2022). A Machine Learning System to Classify Cybercrime. In Mobile Computing and Sustainable Informatics: Proceedings of ICMCSI 2022 (pp. 199-208). Singapore: Springer Nature Singapore.

[13] Akinbowale, O. E., Klingelhöfer, H. E., & Zerihun, M. F. (2022). Analytical hierarchy processes and Pareto analysis

for mitigating cybercrime in the financial sector. Journal of Financial Crime, 29(3), 984-1008

[14] Amin, I., Haque, E., Noor, M. N., Ahmmed, M. M., Puri, S., & Babu, M. A. (2024, April). From Detection to Disruption: Leveraging NLP Insights to Proactively Combat Cybercrime in Bangladesh. In International Conference on Innovations in Computational Intelligence and Computer Vision (pp. 81-95). Singapore: Springer Nature Singapore.

[15] Arjunan, T. (2024). Detecting Anomalies and Intrusions in Unstructured Cybersecurity Data Using Natural Language Processing. International Journal for Research in Applied Science and Engineering Technology, 12(9), 10-22214.

[16] Ullah, F., Faheem, A., Azam, U., Ayub, M. S., Kamiran, F., & Karim, A. (2024, May). Detecting Cybercrimes in Accordance with Pakistani Law: Dataset and Evaluation Using PLMs. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 4717-4728).

[17] Sharma, S., & Arjunan, T. (2023). Natural Language Processing for Detecting Anomalies and Intrusions in Unstructured Cybersecurity Data. International Journal of Information and Cybersecurity, 7(12), 1-24.

[18] Srinivas, D., Bansod, P. J., Singh, M., Takhar, S., Chouhan, K., & Fatma, G. (2023, December). Combating Cybercrimes: Leveraging Natural Language Processing for Detection in Social Media. In the International Conference on Mechanical and Energy Technologies (pp. 265-277). Singapore: Springer Nature Singapore.

[19] Akshaya, R., & Saravanan, C. (2024, November). A Novel Approach for Building Cyber Crime Prediction and Analysis Model using Random Forest. In 2024 8th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS) (pp. 1-6). IEEE.

[20] A. R, S. C and D. T. L, "A Novel Approach for Building Cyber Crime Prediction and Analysis Model using Random Forest," 2024 8th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS), Bengaluru, India, 2024, pp. 1-6, doi: 10.1109/CSITSS64042.2024.10816938.

[21] Yasar, S., Gazi, M. M. H., & Alam, K. S. (2023). Multilevel Voting Models in Cyber Aggression Detection for Bangla Texts. In Applied Informatics for Industry 4.0 (pp. 212-222). Chapman and Hall/CRC.

[22] Srinivasan, S., & Deepalakshmi, P. (2023). Enhancing security in the cyber-world by detecting the botnets using ensemble classification based machine learning. Measurement: Sensors, 25, 100624.Sannella, M. J. 1994 Constraint Satisfaction and Debugging for Interactive User Interfaces. Doctoral Thesis. UMI Order Number: UMI Order No. GAX95-09398., University of Washington.