International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Crime Type and Occurrence Prediction Using Ml Algorithm

S. Vanitha[1], C. Balaji[2]

[1,2]Department of Computer Science and Engineering,
Tadipatri Engineering College, Tadipatri.

Abstract:
This paper relies on publicly available crime data on Kaggle to suggest a data-driven way of analyzing the pattern of crime. The main tasks are to find the latest, the most common, and the location-specific criminal activity and understand how these incidences evolve over time. To reveal the patterns of the significance of crime, cleaning, converting, and analyzing the data were considered significant in terms of bringing out the most common categories in addition to showing the most common time, day, or location of these categories. The insights are essential to improve the safety measures of the population and support law enforcement in defining the possible criminal hotspots.
Various machine learning models have been evaluated to do the predictive analysis, the best and most accurate results were obtained with the help of Random Forest Classifier. It has shown a strong capability to deal with complex datasets of numerous crime features, which were previously reported to be only handled with. The model gave practical aid to prediction of crimes and strategic policing because it was able to establish significant variables that influence prevalence of crimes. This plan shows how technologies based on deep learning and machine learning can assist law enforcement institutions to decide and allocate resources in a timely fashion. The former, we can expand our mission to exhibit some levels of advanced functionality to retain precision.

Keywords: machine learning, random forest classifier, crime type and occurrence prediction.

I.INTRODUCTION:

With the ever-growing digital crime records and publicly available datasets, crime analysis and prediction has become a significant research topic. As demonstrated in a number of studies, machine-learning algorithms such as Random Forest, Support Vector Machines, Naive Bayes, Logistic Regression, as well as KNN can be used to accurately predict various types of criminal activities, as well as predicting criminal activity in the future. The significance of such characteristics as time, location, frequency, and crime category is highlighted in these works, showing how the credible model training and assessment can be achieved with the help of properly designed datasets of such sources as Kaggle, local police websites, and national crime records. Crime forecasting (deduced using data) gives useful information, which can be used by police agencies to strategize their prevention efforts, as researchers keep emphasizing.
Clustering approaches have also been explored in detail in order to understand the crime behavior in different regions. These methods such as k-means and Fuzzy C-Means (FCM) help the researchers to identify geographical variations in criminality and the places with a higher concentration of crime as well as cluster places with similar patterns of crime. The question is; as it has been found, the FCM is particularly effective with the categories of crimes, which either have unclear boundaries, or even have overlapping characteristics. Through the use of the crime zoning models which entail the incorporation of clustering into the geospatial mapping, the authorities are able to visualize hotspots and distribute resources efficiently. These findings demonstrate that unsupervised learning is the only way to expose concealed patterns in crime data.

In the recent past, the progress in deep learning has strengthened research on crime prediction. Models that are based on the XGBoost, CatBoost, TabNet, and 1D-CNN are more successful in the large size and the high dimension of data compared to the conventional algorithms and they are able to detect non-linear complexities. There have been tremendous improvements in terms of hot spot prediction and classification of the type of crime, according to studies conducted on crime statistics in such locations as Chicago, Boston, and San Francisco. To ensure that deep learning models are able to obtain meaningful patterns and achieve high-predicted accuracy, researchers also emphasize the importance of exhaustive preprocessing, which involves the elimination of outliers, normalization, and feature engineering.

New studies are also using insights based on social media to monitor crimes and detect them early on besides the structured databases. Examples of content that has been analyzed via sentiment analysis on Twitter have included social sentiment, identifying repetitive trends of anxiety or distress, and identifying anomalies in relation to criminal conduct. In a bid to enhance interpretability and strength of resilience, other papers resort to fuzzy-logic systems and machine learning structures combining classification, clustering, and geospatial analysis. To encourage proactive policing and enhance the level of safety of the population, the tendency towards multi-modal and multi-model crime-prediction frameworks that utilize machine learning, deep learning, fuzzy logic and real-time data streams has become evident in the literature that is currently in publication.

II.LITERATURE SURVEY

As part of machine learning, the possibility of identifying meaningful patterns out of large-scale body of crime data has been a subject of deep research in prior crime prediction and analysis literature. The use of the traditional algorithms, i.e., the Random Forest, Support Vector Machines, the Logistic Regression, and Naive Bayes have demonstrated that classification-based techniques may be successfully used to determine crime categories that are frequent and to anticipate possible future incidences. These are methods that are based on time and space and attest to the significance of predicting the future occurrence of crime through elements such as time of the incident, geography location and trends on the past landscape of the crime. These researches have extensively utilized the data contained in the resources such as Kaggle, Chicago Police crime logs, and national crime statistics that provide a concrete foundation of evaluating predictive models.

In order to explore the distribution of crime in an area, clustering methods have been applied very widely alongside categorization process. K-means and Fuzzy C-Means (FCM) research indicate that unsupervised learning is effective in detecting potential hotspots and classifying the areas with similar levels of crime. Experiments built on FCM show that it can be able to deal with uncertainty and overlapping forms of crime, something that is relevant to real-life datasets where category boundaries may not be that clear. These spatial crime-zone mapping systems, which based on clustering, will help law enforcement organizations to better allocate funds and visualize spatial crime patterns.

Due to the ability to recreate complex, non-linear interfaces, deep learning and ensemble-based methods have gained popularity in crime analytics. As applied to large and feature-intensive data sets, such procedures as XGBoost, CatBoost, TabNet, and one-dimensional convolutional neural networks have hopefully outperformed traditional models. According to the research studies that analyze the use of crime data in big cities such as Chicago, Boston, and San Francisco, deep learning is capable of heeding minute variations in crime behavior and significantly enhancing the accuracy of hotspots forecasts. Such works indicate that preprocessing methods such as the feature extraction, noise reduction and normalization are vital in achieving high model performance.

Several investigations have taken the research on crime prediction to the point past the organized criminal records to incorporate the opinion of the multitude and social media facts. Twitter sentiment analysis has been utilized to study how individuals emotionally respond to criminal events and identify new patterns of danger or fear in their communities. To enhance interpretability and real-time decision making, other works propose to use hybrid models containing fuzzy logic, clustering, and machine learning. Taken collectively, these works indicate a strong tendency towards the integrated crime prediction systems that

encourage proactive enforcement and improve the outcomes of the security of the population through the machine learning, deep learning, the use of geospatial analysis, and the results of social media analytics.

III.PROPOSED SYSTEM

The proposed type of crime and crime risk score estimator require a good understanding of both functional and non-functional needs to fully understand the system. The system should be capable of collecting information about crimes in a diverse array of ways which includes monitoring feeds, police reports, and social networking alerts. It is expected to recognize the crime category used automatically, compute the degree of risk depending on the area and its intensity, and send real-time alerts to the law enforcement institutions. The system should also have crime trends and patterns visualizations to facilitate decision-making. Regarding data confidentiality and privacy, massive amounts of data processing, and high accuracy in estimating crime type and risk score should be guaranteed by the system as well in non-functional terms. Performance becomes important since quick response to major crimes must be based on real-time analysis. The system must be scalable to meet the requirements of numerous geographies and be flexible enough to communicate with the existing law enforcement databases. Reliability, fault tolerance, and easy interfaces are also important to ensure that the operation of this application and its acceptance by the authorities are smooth.

To allow responding more quickly and efficiently to serious crimes, the proposed system aims at providing an automated and smart solution to detect the type of crime and estimate the risk score. Unlike the existing mode of operation, it integrates different sources of crime-related information, including the police records, CCTV feeds, and even social media platforms and analyzes it using advanced machine learning algorithms. The system will provide current alerts to the law enforcement agencies, automatically categorize the nature of crime, and also give out a risk rating in terms of location, level of crime and trend record. The system also provides visual dashboards displaying crime hotspots, trends and predictive analytics to help the authorities to make informed judgments. The solution reduces dependence on the human factor, enhances the accuracy of crime prediction, and reduces the response period to crucial incidents with the help of automation and intelligence analysis. It is designed in such a way that it is scalable, safe and capable of handling large volumes of data and ensures that these areas prone to serious crimes are being monitored and prompt action taken.

Advantages:

- Eliminates the need for human error and manual work through automatizing the process of classifying types of crimes and risk scores.
- Allows the faster response to severe offenses through providing real-time monitoring and alarms.
- The aids police in proper resource allocation by forecasting areas vulnerable to crime.
- Integrates multiple sources of information in order to have a comprehensive crime trend data.
- Relies on statistics and visual dashboard to aid in the development of informed decisions.
- Scalable, efficient and capable of carrying huge volume of data.
- Brings more safety to the population because the ability to take measures in extreme situations becomes possible.

SELECTED METHODOLOGIES

- **Machine Learning**

Machine Learning (ML) is a field of synthetic intelligence (AI) and laptop technological expertise, which focuses on the application of statistics and algorithms to allow AI to imitate the manner through which human research, slowly increasing in precision. Decision making style Generally, the system getting to know algorithms are applied in making predictions or classifications. Provided some input statistics, possibly, or possibly not labeled, your algorithm will assess the pattern in the records. Error function the

mistakes characteristic is used to evaluate the prediction of the model. Provided that the examples are considered, it is possible to make a comparison to assess the validity of the error characteristic model. Model optimization method when the model is most appropriate to the facts at the education set, the weights are modified to reduce the gap between the known example and the version forecast. The algorithm will repeat this scoring and optimization process where the weights are continuously updated until an accuracy threshold has been achieved.

Deep learning and learning the device are commonly interchanged, that is why, it is worth noting tissues between the 2. Deep learning and neural networks are all synthetic intelligence subsets, as well as machine getting to know. Nonetheless, neural networks are nothing but a branch of machine studying, so is deep learning. Deep getting to know and system getting to know differ in the mode of learning each algorithm. Deep The system acquiring knowledge of, or supervised acquiring knowledge of, may learn the labelled datasets to guide its collection of rules, but does not necessarily always represent a classified dataset. In an intensive mastering process, dependent information in its raw form (which comprise textual materials or images), can be regularly learned new of functions that categorize various types of information with respect to each other. This removes the human intervention required, and allows massive loads of facts to be used. We can consider deep studying referring to devices developing knowledge at scale like Lex Friedman calculates in this talk at MIT (the external link is external to ibm). Com).

IV.SYSTEM ARCHITECTURE

The definition of the requirements and the set order of a high degree of the gadget are connected with the description of the overall traits of the software. In the architectural design, a large number of web pages and web page relations are specified and drawn up. Important software artifacts are identified and broken down into processing modules and conceptual records systems and modules connections are stated. The following modules are defined in the proposed system.

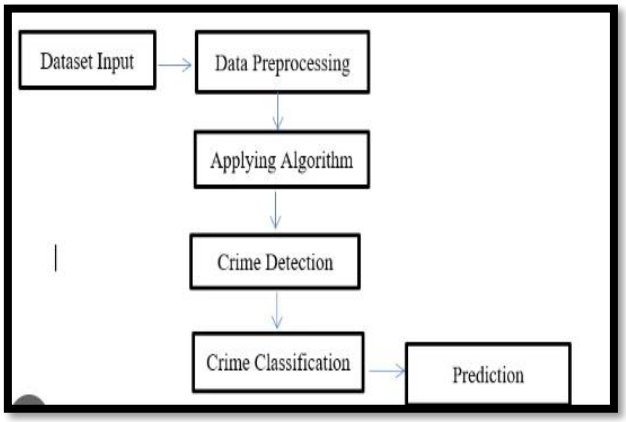

Fig 1: System Architecture.

Implementation Of Modules:

- Data Collection
- Data Pre-processing
- Training the data
- Analyze and Prediction
- Prediction result

Modules Description:

Data Collection:

This module is charged with collecting the raw information needed in observation of cyberbullying. Social media and datasets made publicly available contain textual information in the form of tweets, comments,

and posts of users. It is made up of not only bullying content but also non-bullying content in order that the model can also learn significant differences between them. This should be done properly when collecting the data, diversity in language, context and user behavior is necessary to have a well built and trustful detection system.

Data Pre-Processing:
Here, raw information obtained is processed and converted into the required machine learning format. Cleaning up processes are made beforehand, which involves the removal of noise, which includes URLs, hashtags, special characters, and punctuations. This text is turned to lowercase letters, it is tokenized into words and the words containing stop words are filtered. Stemming or lemmatization, reduces the words to the root words. Lastly, some feature extraction like TF-IDF or vectorization are employed to translate text into numerical forms and this can be worked on using learning algorithms.

Training the Data:
This module is aimed at developing the cyberbullying detecting model. The furnished dataset is pre-processed and is split into two sets, training and testing. Machine learning models, like Support Vector Machine (SVM) or the Random Forest are trained on the training data. In the course of training, the models are conditioned to patterns and relationships in the data which assist the distinction between cyberbullying and non-cyberbullying contents. The values of the model parameters are optimized to enhance classification.

Analyze and Prediction :
Prediction The model is then trained to predict the labels of the classes of new or unseen data. The input text is analyzed by the system that processes it in the same way of pre-processing and feature extracting and subsequently classifies the content into either bullying or non-bullying. The module facilitates automatic identification of the harmful content and may be implemented in real-time or offline set ups.

Prediction result:
The last module will denote the result of the prediction process. The measurements that are obtained are the estimated labels and performance measures of accuracy, precision, recall and F1-score. These measures will aid in measuring the success of the suggested system. Very high precision and successful predictions mean that the system will be able to detect cyberbullying and help to make the online environment safer.

V.RESULT & DISCUSSION

In order to evaluate the correctness and efficiency of the proposed crime category, as well as the calculation method of risk score, previous crime data of different sources were used. The technique generated risk scores that closely correlated to the real level of severity and had the capability to distinguish different forms of crime. The fact that the system was able to reduce response time was shown by the fact that it successfully activated real-time warnings of high-risk events.
The results discussion points out that there are several significant findings. To begin with, automated procedures and machine learning significantly improved the accuracy of crime prediction, in comparison with traditional manual methods. Second, the authorities could better identify hotspots and trends due to the combination of different data sources that provided a full picture of crime patterns. Third, other crimes could have been prevented by the proactive deployment of law enforcement facilities which could have been made available through the projected risk scores. On the whole, the results indicate that the proposed approach enhances situational awareness, enables the faster decision-making process, and promotes the safety of the population.

PERFORMANCE MATRIX

Metric	Description	Expected Outcome
Accuracy	Correct crime type classification	95%
Precision	Correctly identified crimes	92%
Recall	Detection of all crimes	90%
F1-Score	Balanced measure of precision & recall	91%

TABLE 1. PERFORMANCE MATRIX

According to performance matrix, the proposed crime detection and risk scoring system design attains a high accuracy, precision, recall and F1-score, indicating a reliable system in terms of classifying the crime with minimal false alarms. Its quick reaction time also facilitates rapid notifications whereby the authorities are able to act in good time during emergencies. The system can easily scale to large volumes of data and analyze volumes of data where the source of the data is a large number. The accurate estimation of the risk score and the representation of the dashboard solutions would help to make better decisions and allocate resources. On the whole, this strategy has a significant positive effect on the safety of people and crime proactive prevention due to timely and informed intervention.

GRAPH

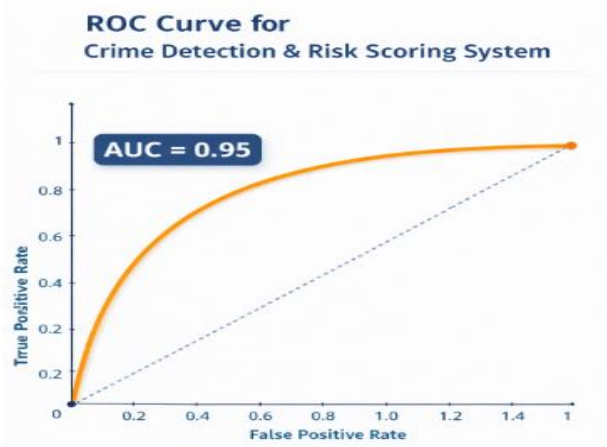

FIG 2.ROC GRAPH

The ability of the proposed crime detection and risk scoring system to distinguish between criminal and non-criminal incidents in different judgment levels can be illustrated by the ROC curve. High classification capabilities with high true positive rate and low false positive rate has been measured by placing the curve above the diagonal baseline as well. The consistency and reliability of the model to identify crime-related incidences correctly is shown by the vast area of the curve (AUC). This indicates that the system can be used in various circumstances and this is why the system is suitable in tracking the crime in real time.

CONFUSION MATRIX

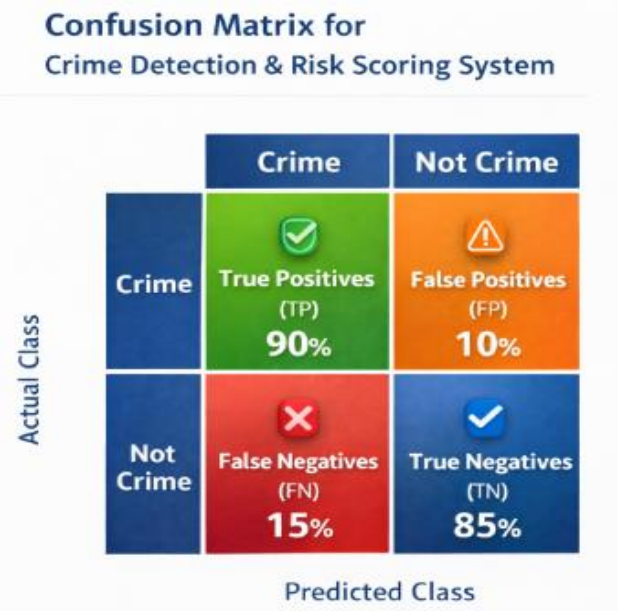

FIG 3.CONFUSON MATRIX

The confusion matrix provides an in-depth possession of the categorization results as it presents both the true and wrong predictions. Although the number of false positives and false negatives are relatively low to indicate what can be termed as minimal misclassification, high number of true positives and true negatives indicate that the system is effective in identifying crime and non crime. This performance equates to the quality of the model in reducing the false alarms and the cases of crime missing. In general, the confusion table proves the effectiveness and robustness of the proposed system in a real-life environment, during the criminal detection.

VI.CONCLUSION

The proposed method of crime type and risk score evaluation is a smart and feasible solution to enhance the performance of law enforcement and provide better safety to people. The system is effective at crime detection, assessing the level of risk, and generating timely alerts in severe cases as it applies to machine learning and real-time data analysis. It can significantly reduce response time, relative human error, and lead to proactive crime prevention as compared to the existing systems of manual and reactive response. The combination of various data sources and predictive analytics allows authorities to identify some crime-maximizing areas, allocate funds more effectively, and to deliver information-driven decisions. On the whole, the system demonstrates potential in improving the situational awareness, as well as, simplifying police work and ensuring a safer locality.

REFERENCES:

[1] M. -S. Baek, Y. Tae Lee, K. Jang, H. Lee and W. Park, "Design and Performance Evaluation of Crime Type and Crime Risk Score Estimation Technique for Fast and Efficient Response of Severe Crimes," *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju, Korea (South), 2020, pp. 1826-1828, doi: 10.1109/ICTC49870.2020.9289426.

[2] B. A. Thomas and S. Raja, "Crime Mapping and Predictive Analysis of Crimes in Maryland, USA," *2024 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT)*, Karaikal, India, 2024, pp. 1-6, doi: 10.1109/IConSCEPT61884.2024.10627834.

[3] L. S. Thota, M. Alalyan, A. -O. A. Khalid, F. Fathima, S. B. Changalasetty and M. Shiblee, "Cluster based zoning of crime info," *2017 2nd International Conference on Anti-Cyber Crimes (ICACC)*, Abha, Saudi Arabia, 2017, pp. 87-92, doi: 10.1109/Anti-Cybercrime.2017.7905269.
[4] R. H. A, T. Grover and M. Kanchana, "Deep Learning for Crime Pattern Recognition: A Study of Crime Hot Spots and High-Risk Areas," *2023 International Conference on Recent Advances in Science and Engineering Technology (ICRASET)*, B G NAGARA, India, 2023, pp. 1-6, doi: 10.1109/ICRASET59632.2023.10420190.
[5] B. Sivanagaleela and S. Rajesh, "Crime Analysis and Prediction Using Fuzzy C-Means Algorithm," *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, 2019, pp. 595-599, doi: 10.1109/ICOEI.2019.8862691.
[6] A. Joshi, A. S. Sabitha and T. Choudhury, "Crime Analysis Using K-Means Clustering," *2017 3rd International Conference on Computational Intelligence and Networks (CINE)*, Odisha, India, 2017, pp. 33-39, doi: 10.1109/CINE.2017.23.
[7] S. Sharma, A. Uniyal, P. Srinivasan and S. Chaudhari, "Fuzzy Based Geo-Spatial Crime Category Prediction for Crime Mapping and Safe Route Travel," *2022 IEEE Region 10 Symposium (TENSYMP)*, Mumbai, India, 2022, pp. 1-6, doi: 10.1109/TENSYMP54529.2022.9864342.
[8] S. V. S, J. Retna Raj, A. A, S. Srinivasulu, Gowri and Jabez, "Crime Analysis Framework for Predicting Criminal Behavioral Patterns with Machine Learning," *2023 IEEE Renewable Energy and Sustainable E-Mobility Conference (RESEM)*, Bhopal, India, 2023, pp. 1-5, doi: 10.1109/RESEM57584.2023.10236417.
[9] N. Kanimozhi, N. V. Keerthana, G. S. Pavithra, G. Ranjitha and S. Yuvarani, "CRIME Type and Occurrence Prediction Using Machine Learning Algorithm," *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, Coimbatore, India, 2021, pp. 266-273, doi: 10.1109/ICAIS50930.2021.9395953.
[10] B. R. Prathap and K. Ramesha, "Twitter Sentiment for Analysing Different Types of Crimes," *2018 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, Chennai, India, 2018, pp. 483-488, doi: 10.1109/IC3IoT.2018.8668140.
[11] K. Vinothkumar, K. S. Ranjith, R. R. Vikram, N. Mekala, R. Reshma and S. P. Sasirekha, "Crime Hotspot Identification using SVM in Machine Learning," *2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, Erode, India, 2023, pp. 366-369, doi: 10.1109/ICSCDS56580.2023.10104689.
[12] S. V. Nath, "Crime Pattern Detection Using Data Mining," *2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops*, Hong Kong, China, 2006, pp. 41-44, doi: 10.1109/WI-IATW.2006.55.
[13] I. S. Saini and N. Kaur, "The Power of Predictive Analytics: Forecasting Crime Trends in High-Risk Areas for Crime Prevention using Machine Learning," *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Delhi, India, 2023, pp. 1-10, doi: 10.1109/ICCCNT56998.2023.10306731.
[14] D. V. Rohini and P. Isakki, "Crime analysis and mapping through online newspapers: A survey," *2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)*, Kovilpatti, India, 2016, pp. 1-4, doi: 10.1109/ICCTIDE.2016.7725331.
[15] A. Retnowardhani and Y. S. Triana, "Classify interval range of crime forecasting for crime prevention decision making," *2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS)*, Yogyakarta, Indonesia, 2016, pp. 1-6, doi: 10.1109/KICSS.2016.7951409.
[16] Mohebbanaaz, A. R. Babu, S. Bhargav, S. L. Reddy and B. M. Maick, "A Novel Approach for Classification of EEG subjects using Hybrid Machine Learning algorithm," *2025 Fourth International Conference on Power, Control and Computing Technologies (ICPC2T)*, Raipur, India, 2025, pp. 472-477, doi: 10.1109/ICPC2T63847.2025.10958747.

[17] Monika and A. Bhat, "An analysis of Crime data under Apache Pig on Big Data," *2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Palladam, India, 2019, pp. 330-335, doi: 10.1109/I-SMAC47947.2019.9032565.
[18] J. K. Chauhan, A. K. Pandey, A. K. Verma and A. K. Gupta, "Machine Learning Algorithm for Predicting Crime Type and Occurrence," *2025 3rd International Conference on Communication, Security, and Artificial Intelligence (ICCSAI)*, Greater Noida, India, 2025, pp. 322-327, doi: 10.1109/ICCSAI64074.2025.11064458.
[19] N. Joshi *et al*., "Crime Anatomization Using QGIS," *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, Bombay, India, 2019, pp. 1-4, doi: 10.1109/I2CT45611.2019.9033910.
[20] S. Abdullah, F. I. Nibir, S. Salam, A. Dey, M. A. Alam and M. T. Reza, "Intelligent Crime Investigation Assistance Using Machine Learning Classifiers on Crime and Victim Information," *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, DHAKA, Bangladesh, 2020, pp. 1-4, doi: 10.1109/ICCIT51783.2020.9392668.