Shraddha J. Pawar, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Machine Learning for Crime Analysis and Prediction: Modern Technique and Issues

Research Scholar Shraddha J. Pawar[1], Baisa L. Gunjal[2]
Dept. of Computer Engineering
MET Bhujbal Knowledge city, Institute of Engineering[1]
Head, Information Technology Amrutvahini College of Engineering, Sangamner, Maharashtra[2]

Abstract- Crime forecasting has now acquired more signifi- cance in improving security and directing constabulary policies. This study will examine the current state of the art in crime pre- diction using Machine Learning techniques, specifically looking at classifiers such as linear regression, logistic regression, Neural Networks, Ensemble Methods, and Hybrid Frameworks. The paper presents the most important aspects of the approaches of analyzing the data, including data preprocessing, feature selection, evaluation of the models and algorithms, and applications. Part of the problems that are coming out in the development of this technology are touched upon briefly, such as data bias, privacy and scalability and solutions are given. The results indicate that there are certain trends as multimodal integration and spatial-temporal models, which have led to the change in the prediction accuracy.

Keywords— Crime forecasting, machine learning approach, public safety

I. INTRODUCTION

Criminal activities; continue to be a thorny issue that impacts on ;security, development, order as well as welfare, in society. With population density growing and social structures becoming constantly more nuanced, police forces worldwide are experiencing growing pressure to prevent crimes as well as react to them. In this regard crime forecasting has becomes fundamental area of research for estimating the future crime, to locate hot spots and to allocate resources in crime prevention efforts[1,2]. With the advent of machine learning techniques that im- prove the capacity to churn through massive and diverse data sets, probabilistic crime prediction has further gotten a signif- icant boost. Current conventional crime analysis procedures are, however, found not be very useful since they just use normal historical data and use normal models that do not change.However, machine learning algorithms allow for a mor targeted adaptation of the decision-making process to the wide range of patterns, correlations, and trends seen in criminal activity [3,4].

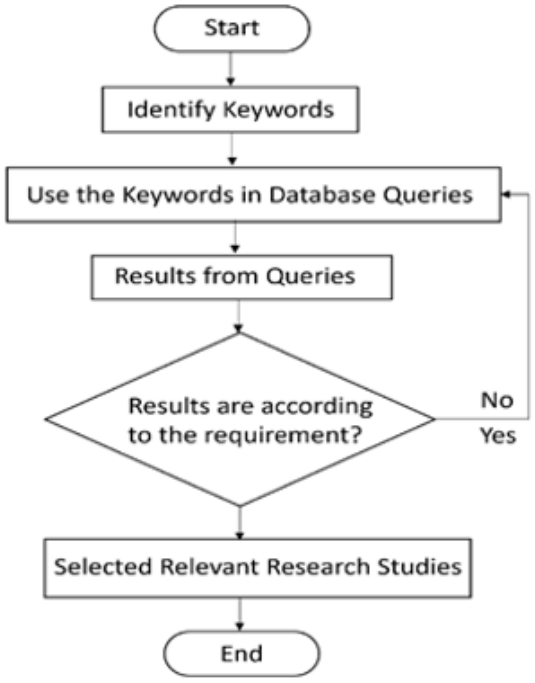

Fig. 1: Step involved for typical crime detection

Several approaches like decision tree, neural networks, ensemble, spatial-temporal model, etc., also classified good in the identification of crime areas, factors influencing crime occurrence, and making smart decisions for crime prevention strategies. The figure 1 below shows the number of

research papers published from year 2018 to year 2024 in the field of crime forecasting [5].

However, similar to any machine learning application in the area of crime analysis, crime forecasting is not immune to some of the following hurdles. Questions about the most effective Annual Event Strategies and the application of Big Data technologies raised some concerns including data bias, privacy, and ethicality of the predictive policing. Furthermore, the dynamic nature of crime and the increasing complexity of datasets necessitate continuous innovation in model develop- ment, feature engineering, and evaluation metric.

The purpose of this review article is to provide a thorough evaluation of the most current developments in crime fore- casting using machine learning techniques. It explores the advantages and disadvantages of various techniques, discusses applications and applications, and identifies the significant impediments that have to be overcome to increase use. In order to help achieve the objective of creating safer communities, this study aims to synthesize current research and provide future directions for the creation of reliable, ethical, and effective crime prediction systems.

II. LITERATURE SURVEY

Machine learning models have shown to be successful in crime prediction in several studies [6,7,8]. However, these models often fail where they do not predict those variables that are most important in the crime occurrence. More research on classification methods has been carried out to forecast various kinds of crime occurrences based on crime reports. They are useful for the regulating body for crime control in the society planning for the prevention strategies. Nevertheless, the issue that still remains unsolved is manual nature of such accumulation and classification in order to reveal types of a crime.

A research [9] proposed an effective new incremental clas-sifier which can make learning and prediction dynamic. An effective model for the dynamic prediction and categorization of crime reports was developed in this study using the Bi- objective Particle Swarm Optimization method. The crime data from sources were collected through online newspapers for the study, and the performance of the model was tested using blind police narrative reports. The approach applied was tested on four datasets and the significance was statistically confirmed. Another paper [10] lay its attention towards the prediction of crimes through the application of the XGBoost algorithm. The study used an improved decomposition and fusion method and included theft occurrence information from H city. Use of techniques like OVR-XGBoost and OVO-XGBoost allowed them to address the issue of having more features than samples, which is especially prevalent in theft datasets. The SMOTENN method was used to address the imbalance, and the accurate as OVR-XGBoost and OVO-XGBoost.

In [11], 17 variables were used to train crime prediction model by employing the XGBoost algorithm. The impact of each variable was assessed using the post hoc interpretability technique called Shapley Additive Explanation (SHAP). The study shows that contextual factors including non-resident proportion and the existence of young people between ages 25 and 44 were useful for predicting crimes. Such study also reveals that the probability of the commission of public crimes was higher in areas with a higher ambient population in this age bracket. The SHAP values confirmed the importance of each of the variables and even though some components showed spatial variation, they offered guidance on aspects that could be important for law enforcement. Another work [12] is devoted to the prediction of crimes that occur during or after psychiatric treatment. Researchers used information from Danish patient registries to create a pre- diction model after realizing the shortcomings of conventional danger assessment techn

Another research [13] introduced a graph-based ensemble classification method to enhance crime report predictions compared to traditional classifiers. This approach involved modeling crime reports graphically to identify a maximal independent subset of features, followed by classification using decision tree algorithms. Extensive experiments confirmed

the superior performance of this ensemble method across multiple datasets. To further our knowledge of the elements impacting crime detection, several research [14], [15] have concentrated on deciphering predictions connected to crime.

Kadar and Pletikosa [17] identified data sources like reg- istration records, crime rate, location-based social network, metro and taxi trip data and the like. They developed two types of crime prediction models: One to predict a crime rate for the future as far as one to five years and another for short-term predictions. Okutan et al. [18] used non-traditional metrics derived from other sources GDELT, OTX and Twitter for cybersecurity prediction of cyber-attacks. Their approach included mechanisms, which include predictive signal attribu- tion (PSA), aggregation of signals with large latencies (ASL) and SMOTE++ to address imbalanced data which allowed the accurate prediction of cyber threats

Kim et. al [19] presented a new approach where GMM and CNN were used for object detection. For baseline modeling, GMM was used for background subtraction and object extrac- tion and CNNs were utilized for object categorization in the ROI. This approach showed excellent performance, especially while recognizing small or distant moving objects.Yomfell et al. [20] investigated crime data employing the social disorganization theory, incorporated the aspects of pop- ulation, POIs, taxi traffic and the residents' tweets. Their study, which they conducted with the help of machine learning, spatial linear regression, and the Poisson GLMs identified the associations between crime and spatial variables. The information was useful for crime prevention and the working out of police measures. Seo et al. created a Partially Generative Neural Network (PGNN) to address missing value problems in predictive modeling and classification. To facili- tate the introduction and testing of models created in one area with parti

Mohd et al. focused on text characteristics including racial/ethnic background, age, income, family composition, education, population, geography, and unemployment rates in their feature selection techniques for crime data. This study showed that much work needs to be done in the determination of feature space relevant for crime prediction. Hardyns et al.

provided an overview of the Amsterdam Police Depart- ment's Crime Anticipation System (CAS), which is used to pinpoint areas prone to crime. Using cutting-edge methods like neural networks and strategic regression, proactive polycentric crime prevention and integrative policing focused on several data aspects.

In the study by Vural and Gok compared the decision trees and Naive Bayes classifiers in terms of their suitability for predicting crime. Decision trees used a top down hierarchi- cal model to organize attribute information; Na¨ıve Bayes used probabilistic reasoning to decide the chance of outcomes given current conditions. Kouziokas pointedly noted that the scaled gradient computation as the fast and less-costly learning approach, promising for enhancing models of a neural network for the crime hotspots identification based on spatial data. Russel has also called and measured validity, fairness, reliability and usability of predictive models for effectiveness, although the performance can be tested through the different techniques such as; ROC curves.

A detailed investigation of fear of crime in a small-town set- ting was presented , where both individual-level demo- graphic characteristics and community-level influences were examined. The study compared two explanatory frameworks: the vulnerability model, which links fear to demographic susceptibility, and the incivilities model, which attributes fear to neighborhood disorder, social disruption, and dissatisfaction with policing practices.

The relationship between economic conditions and crime patterns in India was analyzed using multiple machine learning techniques, including decision trees, random forests, linear regression, and neural networks. Economic indicators such as Gross District Domestic Product and unemployment rate were evaluated against crime categories like theft, burglary, and robbery.

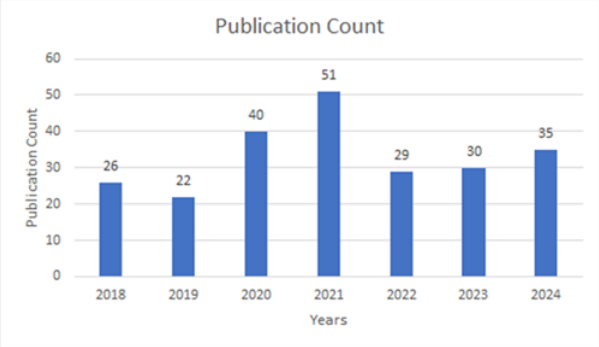

Fig 2: Research Publication Trend

Sr. No.	Year	Publication Count
1	2018	26
2	2019	22
3	2020	40
4	2021	51
5	2022	29
6	2023	30
7	2024	35

The results showed that the predictive accuracy of linear regression was the highest, which proved that there is a strong relationship between the level of unemployment and the rate of robbery. The heterogeneous crime prediction models were studied, which combined crime records, census statistics, location based social networks data and public transport activity. The models were divided into the long and short term predictions to include the crime patterns in the long term (one to five years) and in the short term (temporal). The study uses unconventional sources of data to forecast cyber-attacks because it uses signals related to global event databases, threat intelligence services, and social media feeds. The predictive signal imputation technique, lag signals aggregation, and SMOTE++ were used to address the issue of missing values and class imbalance, which led to better prediction of cyber-attacks.

an object detection framework based on probabilistic modeling and deep learning was presented, with Gaus- sian mixture models to do the background subtraction and object extraction followed by the convolutional neural networks to classify objects in regions of interest. The method proved to be good especially in the detection of small and far moving objects.

crime analysis based on the social disorganization theory was performed, which included demographic variables, point of interest, transportation flows, and social media activ- ity. The use of spatial regression models, generalized linear models, and machine learning techniques to determine significant relationships between crime occurrences and spatial charac- teristics were used to support informed policing strategies. To improve the accuracy of classification and prediction, presented a partially generative neural network architecture which is effective in generating missing data as the learning process progresses. A region-transfer learning, which allows models, trained in one location to be adjusted and tested in a different location, with the same features, was proposed in [20].

The problem of feature selection in crime prediction was studied in with a primary focus on the difficulties of feature selection, which are the influential variables, including income, age distribution, ed- ucation, population density, and unemployment. The research revealed the importance of considering features to achieve better predictive reliability. A framework of intelligence-led policing was proposed whereby tactical decisions were incorporated with predictive analytics using a crime anticipation system. The system used a combination of a logistic regression model with a neural network model to handle large datasets efficiently and identify crime hotspots to help proactive law enforcement.

In the traditional classifiers used in crime predic- tion were evaluated and found in decision tree models, which are hierarchical associations between the attributes, and Naive Bayes, which model the probabilities of events based on the assumption of condi- tional independence. It was highlighted in that neural network optimization scaled conjugate gradient algorithm was an efficient way to perform classification of hotspots basing on spatial data. Lastly, predictive model assessment was also addressed in , and it was pointed out that validity, fairness, reliability, and practical utility were important elements, evaluated through receiver operating characteristic analysis.

Despite the fact that current literature shows the potential of perfor-mance in the framework of

ensemble and deep learning models, there are a number of gaps that are still noticeable. Most methods are based on their region specific data, which restricts their generalizability, whereas other methods propose accu- racy without adequate interpretability or bias analysis. Further, unequal assessment measures and discussion of ethical deployment are inconsistent, and further, there is the need to establish standardized transparent and context-sensitive crime prediction systems. This review takes a position by summarizing these limitations and describing research directions, which would deal with robustness, ex- plainability, and fairness.

III. DISCUSSION

Table 1: Comparative Analysis Of Existing Crime Prediction Approaches

Methodology	ML Algorithms	Dataset	Performance
Analysis of geo-graphical data for crime prediction across various ar- eas of New York City [8]	XGBoost, Random Forest (RF), Support Vector Machine (SVM)	New York City crime dataset	Aggregate Accuracy: 52%
A rules engine combined with a classifier optimized using Particle Swarm Optimization (PSO) to categorize criminal complaints [9]	Multiple classification algorithms	Crime-related articles from USA, India, and UAE (2007–2017)	Aggregate Accuracy: 79%
OVR-XGBoost and OVO- XGBoost models emphasizing dataset structuring for multi-class crime categorization [10]	XGBoost-based algorithms	H City, China (2019)	Aggregate Accuracy: 85%
Spatiotemporal crime prediction using XGBoost with SHAP-based explainability [11]	XGBoost	ZG City, Southeast China (2017–2020)	Accuracy: 89%, AUC: 0.586
Evaluation of multiple classifiers under random under-sampling and over-sampling strategies [12]	Logistic Regression, RF, XGBoost, LightGBM	Danish National Psychiatric Patient Register (474 crime records)	F1 Score: 76%
Graph-theory-based crime analysis with feature selection using decision trees and heuristics, followed by ensemble classification [13]	Multiple classification algorithms	Crime articles from USA, UAE, and India (2008–2016)	Aggregate F1 Score: 88%

Interpretable crime prediction using the Explainable and Accurate Decision Tree (EADT) technique [14]	Decision Tree (DT)	Pennsylvania State Prison Records	Aggregate Accuracy: 77.6%
Crime clustering and prediction using traditional data science workflows and classification modeling approaches [16]	K-Means clustering	National Crime Records Bureau (India)	Aggregate Accuracy: 78%

The literature survey shows the efforts made in crimes prediction by using machine learning recent techniques and the difference in the solutions based on the data type and goal. XGBoost, Random Forest, LightGBM have been also identified to help address other problems like imbalanced data, the feature selection issue, and the lack of interpretability. Various techniques like SMOTENN, SHAP, and graph-based modeling praised the exercise by boosting the reliability of predictions. However, the survey reveal some persist trends that are evident in the study which includes; the fluidity of crime patterns that may change over some time (McArdle & Erakovic, 2009), the need to localize studies so as to get a broader context about a region, and the problem with the traditional methods in handling missing or sparse data. To overcome these problems there are some relatively recent ideas including adaptive classifiers, region-transfer frameworks and ensemble models. However, for the dynamic crime pattern to predict and handle, future study could investigate more ma- chine learning algorithm and the methods to combine them for highe

IV. CONCLUSION

The use of machine learning for crime prediction is the main topic of this review, which also delves into the algorithm's growing impact in this area. The basic ideas of modelling have come from older models like the decision trees and logistic regression yet the newer techniques like neural net- works and techniques in ensemble have brought in aspects of enhancing the prediction factors in a number of urban situations. However, challenges like the presence of bias in dataset and privacy issues are still notable problems along with the problems like scalability. Further work needs to be devoted to system solutions for these issues to be discussed to meet the standards of openness and fairness. Using real- time data from social media and IoT devices adds an option to improve the predictive model's accuracy even more. To ensure the right use of machine learning systems while achieving maximum efficiency, researchers, police forces, and policy makers are going to have to work together tightly in the future.

REFERENCES

1. Mandalapu, Varun, et al., "Crime prediction using machine learning and deep learning: A systematic review and future directions," IEEE Access, vol. 11, pp. 60153–60170, 2023.
2. N. A. Sharma, A. B. M. S. Ali, and M. A. Kabir, "A review of sentiment analysis: tasks, applications, and deep learning techniques," International Journal of Data Science and Analytics, pp. 1–38, 2024.
3. N. A. Sharma, A. B. M. S. Ali, and M. A. Kabir, "A review of sentiment analysis: tasks, applications, and deep learning techniques," International Journal of Data Science and Analytics, pp. 1–38, 2024.
4. K. Jenga, C. Catal, and G. Kar, "Machine learning in crime prediction," Journal of Ambient Intelligence and Humanized Computing, vol. 14, no. 3, pp. 2887–2913, 2023
5. V. V. Mahale, N. P. Pareek, and V. U. Uttarwar, "Alleviation of DDoS at- tack using advance

technique," in Proc. Int. Conf. Innovative Mechanisms for Industry Applications (ICIMIA), 2017.
6. M. Sathiyanarayanan, A. K. Junejo, and O. Fadahunsi, "Visual analysis of predictive policing to improve crime investigation," in Proc. Int. Conf. Contemporary Computing and Informatics (ICI), Dec. 2019, pp. 197–203.
7. A. Araujo, N. Cacho, L. Bezerra, C. Vieira, and J. Borges, "Towards a crime hotspot detection framework for patrol planning," in Proc. IEEE HPCC/SmartCity/DSS, Jun. 2018, pp. 1256–1263.
8. A. A. Almuhanna, M. M. Alrehili, S. H. Alsubhi, and L. Syed, "Prediction of crime in neighbourhoods of New York City using spatial data analysis," in Proc. Int. Conf. Artificial Intelligence and Data Analytics (CAIDA), Apr. 2021, pp. 23–30.
9. P. Das, A. K. Das, J. Nayak, D. Pelusi, and W. Ding, "Incremental clas- sifier in crime prediction using bi-objective particle swarm optimization," Information Sciences, vol. 562, pp. 279–303, Jul. 2021.
10. Z. Yan, H. Chen, X. Dong, K. Zhou, and Z. Xu, "Research on prediction of multi-class theft crimes by an optimized decomposition and fusion method based on XGBoost," Expert Systems with Applications, vol. 207, Art. no. 117943, Nov. 2022.
11. X. Zhang, L. Liu, M. Lan, G. Song, L. Xiao, and J. Chen, "Interpretable machine learning models for crime prediction," Computers, Environment and Urban Systems, vol. 94, Art. no. 101789, Jun. 2022.
12. M. L. Trinhammer, A. C. H. Merrild, J. F. Lotz, and G. Makransky, "Predicting crime during or after psychiatric care: Evaluating machine learning for risk assessment using the Danish patient registries," Journal of Psychiatric Research, vol. 152, pp. 194–200, Aug. 2022.
13. A. K. Das and P. Das, "Graph based ensemble classification for crime report prediction," Applied Soft Computing, vol. 125, Art. no. 109215, Aug. 2022.
14. Y. Ma, K. Nakamura, E. Lee, and S. S. Bhattacharyya, "EADTC: An approach to interpretable and accurate crime prediction," in Proc. IEEE Int. Conf. Systems, Man, and Cybernetics (SMC), Oct. 2022, pp. 170–177.
15. M. Boukabous and M. Azizi, "Multimodal sentiment analysis using audio and text for crime detection," in Proc. Int. Conf. Innovative Research in Applied Science, Engineering and Technology (IRASET), Mar. 2022.
16. S. R. Bandekar and C. Vijayalakshmi, "Design and analysis of machine learning algorithms for the reduction of crime rates in India," Procedia Computer Science, vol. 172, pp. 122–127, Jan. 2020.
17. C. Kadar and I. Pletikosa, "Mining large-scale human mobility data for long-term crime prediction," EPJ Data Science, vol. 7, no. 1, Art. no. 26, 2018.
18. A. Okutan, G. Werner, S. J. Yang, and K. McConky, "Forecasting cyber- attacks with incomplete, imbalanced, and insignificant data," Cybersecu- rity, vol. 1, no. 1, Art. no. 15, 2018.
19. C. Kim, J. Lee, T. Han, and Y. M. Kim, "A hybrid framework combining background subtraction and deep neural networks for rapid person detection," Journal of Big Data, vol. 5, Art. no. 22, 2018.
20. L. Vomfella, W. K. Ha¨rdle, and S. Lessmann, "Improving crime count forecasts using Twitter and taxi data," Decision Suppot System