International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

Review paper on Intelligent Crime Rate Prediction using Supervised Machine Learning

Nikita Namdev[1], Ritu Prasad[2] Arjun Rajput[3], Saurabh Karsoliya[4]

Department of Computer Science and Engineering

Technocrats Institute of Technology (Excellence), Bhopal[1,2,3]

Technocrats Institute of Technology, Bhopal[4]

Abstract

Crime prediction has emerged as a significant research domain as cities increasingly rely on data-driven strategies to enhance public safety and optimize law enforcement resource allocation. Intelligent crime rate prediction using supervised machine learning has shown considerable promise due to its ability to learn from historical crime patterns and forecast future occurrences with high accuracy. This review paper provides a comprehensive analysis of supervised learning approaches used for crime rate and hotspot prediction, focusing on classical models such as Logistic Regression, Support Vector Machines, k-Nearest Neighbors, and Decision Trees, as well as advanced ensemble methods like Random Forests, Gradient Boosting, and XGBoost. The study also explores recent developments in deep supervised models including LSTM-based temporal predictors and hybrid spatiotemporal frameworks. Key aspects such as dataset characteristics, feature engineering, spatial–temporal aggregation, handling class imbalance, and model evaluation strategies are examined in detail. The review highlights that tree-based ensemble models generally outperform traditional algorithms due to their robustness on heterogeneous real-world crime datasets. However, challenges persist, including data quality issues, bias in police-reported crime records, limited generalizability across regions, and lack of interpretability in complex models. The paper concludes by identifying research gaps and recommending future directions such as fairness-aware modeling, explainable AI, improved spatiotemporal representation, and cross-city transfer learning to support more ethical and effective crime prediction systems.

Keywords: Crime Rate Prediction, Machine Learning, Intelligent

1. Introduction

Crime is a persistent challenge faced by societies across the world, affecting not only public safety but also economic stability, community well-being, and the overall quality of life. As urban populations grow and social environments become more complex, traditional crime prevention and response strategies are increasingly insufficient. Law enforcement agencies have historically relied on experience-driven intuition, manual analysis of crime records, and reactive policing methods. However, such approaches often lack predictive power and fail to effectively anticipate crime trends, especially in densely populated or rapidly evolving environments. In this context, the rise of data-driven technologies, particularly machine

learning (ML), has opened new opportunities for intelligent crime forecasting and proactive policing [1, 2].

Supervised machine learning has emerged as one of the most widely adopted techniques for crime rate prediction because it leverages labeled historical data to learn patterns and generate accurate forecasts. Crime data typically includes features such as geographical location, time of occurrence, crime type, environmental conditions, socioeconomic attributes, and seasonal patterns. By analyzing these variables, supervised learning models can predict the likelihood of future crimes at specific locations and time intervals. This predictive capability allows authorities to allocate resources efficiently, identify potential hotspots, and develop early intervention strategies that reduce the probability of crime occurrence [3].

A broad range of supervised learning algorithms has been applied to crime prediction tasks. Classical algorithms such as Logistic Regression, Support Vector Machines (SVM), Naïve Bayes, and k-Nearest Neighbors (k-NN) have served as foundational approaches for crime classification and hotspot detection. While these methods are computationally efficient and easy to interpret, they often struggle to capture complex nonlinear relationships inherent in real-world crime datasets. To overcome these limitations, researchers have increasingly adopted tree-based algorithms and ensemble techniques such as Decision Trees, Random Forests, Gradient Boosting, and XGBoost. These models offer improved accuracy, handle heterogeneous data effectively, and are capable of learning intricate feature interactions.

In addition to traditional algorithms, deep supervised learning models have gained traction in recent years, particularly for tasks involving temporal or spatiotemporal dynamics. Approaches such as Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNN), and hybrid models like CNN-LSTM have shown promising results in capturing long-term temporal dependencies and spatial variations. Some studies also incorporate graph-based neural networks to represent the spatial connectivity of urban regions, further enhancing predictive performance. These advanced architectures address the shortcomings of conventional ML models by leveraging the rich structure of crime patterns over space and time [4, 5].

Despite these advancements, several challenges persist in intelligent crime prediction. Crime data is often imbalanced, noisy, under-reported, or biased due to systemic factors in policing and reporting practices. Spatial crime patterns can shift over time, making long-term predictions difficult. Additionally, the use of machine learning in law enforcement raises ethical concerns such as algorithmic bias, transparency, data privacy, and the risk of reinforcing discriminatory practices. As a result, there is a growing demand for fair, explainable, and accountable ML systems that balance predictive accuracy with social responsibility [6].

This review paper aims to provide a comprehensive analysis of supervised machine learning techniques used for crime rate prediction. It examines the types of datasets commonly utilized, major supervised algorithms, feature engineering strategies, evaluation metrics, and real-world deployment challenges. The paper also highlights research limitations and identifies emerging trends such as fairness-aware modeling, spatiotemporal deep learning,

and cross-city transfer learning. By consolidating current knowledge and outlining future research directions, this review contributes to the development of more intelligent, ethical, and effective crime prediction systems that can support data-driven decision-making and enhance public safety [7, 8].

2. Literature Review

Crime prediction has become a major focus area in smart city research, where data-driven approaches and machine learning models are increasingly used to analyze complex crime patterns and generate actionable insights. Over the last decade, numerous studies have explored the application of supervised learning, deep learning, time-series modeling, and spatial analytics to forecast crime rates and hotspots. This section reviews key contributions from recent research, highlighting the methodologies, datasets, algorithms, strengths, and limitations of existing systems.

Vanitha et al. (2025) [1] presented an intelligent crime risk prediction framework that combined behavioral analysis with advanced machine learning techniques. Their study emphasized the importance of identifying psychological and situational features from crime records, which significantly enhanced model accuracy. By integrating feature engineering, behavioral markers, and multiple classification models, the authors demonstrated improved prediction performance compared to traditional statistical methods. However, the framework's dependence on high-quality behavioral datasets limits its scalability to regions where such detailed data is unavailable.

Gupta et al. (2024) [2] proposed a systematic comparison of widely-used supervised algorithms—including Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines—for predicting crime categories and frequency. Using municipal crime data from Indian cities, the study showed that ensemble methods like Random Forest achieved superior accuracy due to their ability to capture nonlinear patterns and handle heterogeneous datasets. The authors also highlighted the challenges associated with missing values and imbalanced crime categories, recommending oversampling and normalization techniques to improve predictability. Although the models performed well, the study lacked a strong temporal analysis component, which is essential for forecasting crime trends over time.

Keerthika et al. (2024) [3] introduced a Recursive Neural Network (RNN)-based deep learning model for forecasting crime hotspots. Their method applied temporal crime sequences to capture long-term dependencies and identify areas with rapidly changing crime activities. The approach proved effective for hotspot detection, outperforming classical models that do not incorporate time-series patterns. However, the deep learning model required large volumes of sequential data and extensive training time, making it less suitable for regions with sparse or incomplete datasets.

Lilhare et al. (2024) [4] developed a crime hotspot mapping system integrated with FIR (First Information Report) data, using geospatial analytics and clustering methods such as DBSCAN and k-means. The study demonstrated the usefulness of spatial visualization tools in law enforcement workflows. While the focus was not strictly on predictive modeling, the

ability to convert raw FIR data into interpretable maps facilitated better understanding of crime concentration areas. Nevertheless, due to the absence of predictive algorithms, the system served more as a crime monitoring tool rather than a forecasting solution.

Mussiraliyeva and Baispay (2024) [5] explored machine learning methods for analyzing textual crime data, particularly police narratives and crime incident descriptions. Using algorithms such as Naïve Bayes, SVM, and TF-IDF feature extraction, the study highlighted the value of natural language processing (NLP) for crime classification and context-aware insights. Their research added a novel dimension to crime analytics by integrating unstructured text data, which is often ignored in traditional models. However, the performance heavily depended on the quality of text preprocessing and lacked integration with spatiotemporal predictors.

Sharaff et al. (2023) [6] presented a multi-algorithm crime rate prediction model using supervised learning techniques including k-Nearest Neighbors (kNN), Decision Trees, and Random Forests. Their comparisons showed that Decision Trees and Random Forests gave the highest accuracy due to their interpretability and low sensitivity to outliers. The study also emphasized the importance of feature selection—time of day, day of week, and crime type were found to be significant predictors. However, the dataset used was relatively small, limiting the generalizability of the outcomes.

Ghani et al. (2023) [7] took a different approach by applying ARIMA time-series forecasting models to develop predictive choropleth maps for crime in Visegrad Group countries. Their study demonstrated that traditional statistical models could still produce meaningful predictions when combined with spatial visualization. However, ARIMA models assume linearity and stationarity, which are rarely present in real-world crime data that is often noisy, seasonal, and influenced by irregular socio-economic factors. Despite these limitations, the study contributed valuable insights into the integration of time-series forecasting and GIS-based mapping.

Rajesh et al. (2023) [8] developed a machine-learning-based crime identification system using data analytics and supervised classification algorithms. Their focus was on crime categorization rather than prediction. SVM and Random Forest models yielded the best performance. The study reinforced the usefulness of ML classification for criminal record analysis but did not incorporate temporal or geographical features, restricting its applicability to real-time prediction systems.

Zhou et al. (2023) [9] introduced a sophisticated unsupervised domain adaptation framework to predict crime risk across different cities. Since crime patterns vary significantly between regions, models trained in one city often perform poorly when transferred to another. Their domain adaptation method addressed this challenge by aligning feature distributions across cities, allowing more robust cross-city prediction. This work highlighted one of the major limitations in existing crime prediction research—the lack of generalizability—and offered a promising solution using deep transfer learning. However, the computational cost and complexity of the model were considerable.

Yin (2023) [10] conducted a comprehensive survey on machine learning-based crime prediction, reviewing statistical, supervised, unsupervised, and hybrid approaches. The survey highlighted key challenges such as data sparsity, noise, ethical concerns, class imbalance, and the lack of interpretable models. Yin emphasized the emergence of deep spatiotemporal models, the need for fairness-aware crime prediction systems, and the importance of integrating heterogeneous data from sensors, social media, and city infrastructure. This survey serves as a foundation for understanding the evolution and current state of crime prediction research.

Overall, the literature shows substantial growth in the adoption of supervised machine learning and deep learning for crime rate prediction. Ensemble models are widely favored for their high accuracy and robustness, while deep learning methods offer powerful spatiotemporal modeling capabilities. However, challenges such as data quality issues, lack of generalizability, ethical concerns, class imbalance, and dependence on region-specific datasets persist. Recent advancements like domain adaptation, text analytics, and hybrid architectures indicate promising directions for future research.

3. Machine Learning

Machine Learning is a subset of Artificial Intelligence concerned with “teaching” computers how to act without being explicitly programmed for every possible scenario. The central concept in Machine Learning is developing algorithms that can self-learn by training on a massive number of inputs. Machine learning algorithms are used in various applications, such as email filtering and computer vision, where it is difficult or infeasible to develop conventional algorithms to perform the needed tasks [11, 12]. Machine learning enables the analysis of vast amounts of information. While it usually delivers faster, more precise results to identify profitable prospects or dangerous risks, it may also require additional time and assets to train it appropriately. Merging machine learning with AI and perceptive technologies can make it even more effective in processing vast volumes of information. Machine learning is closely associated with computational statistics, which focuses on making predictions using computers. Machine learning approaches are conventionally divided into three broad categories, namely Supervised Learning, Unsupervised Learning & Semi-supervised Learning, depending on the nature of the "signal" or "feedback" available to the learning system.

Face anti-spoofing (FAS) has lately attracted increasing attention due to its vital role in securing face recognition systems from presentation attacks (PAs). As more and more realistic PAs with novel types spring up, traditional FAS methods based on handcrafted features become unreliable due to their limited representation capacity. With the emergence of large-scale academic datasets in the recent decade, machine learning based FAS achieve remarkable performance and dominate this area.

Supervised Learning

A model is trained through a process of learning in which predictions must be made and corrected if those predictions are wrong. The training process continues until a desired degree

of accuracy is reached on the training data. Input data is called training data and has a known spam / not-spam label or result at one time.

Unsupervised Learning

By deducting the structures present in the input data, a model is prepared. This may be for general rules to be extracted. It may be through a mathematical process that redundancy can be systematically reduced, or similar data can be organized. There is no labeling of input data, and there is no known result [13, 14].

Semi-Supervised Learning

Semi-supervised learning fell between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). There is a desired problem of prediction, but the model needs to learn the structures and make predictions to organize the data. Input data is a combination of instances that are marked and unlabeled.

4. Offences In Social Network

Social media has also created new concerns in relation to crime itself. Victimisation on social media platforms is not uncommon. However, it is not all bad ews. Social media has created new opportunities for criminal justice agencies to solve crimes, among other things [15, 16]. Thus, like many other advancements in communication technology, social media has a good, a bad and an ugly side when it comes to its relationship with criminal justice and the law. There are 5 Common

Types of Social Media Crime and these are listed below.

- Online Threats, Stalking,
- Hacking and Fraud
- Buying Illegal Things
- Posting Videos of Criminal Activity
- Vacation Robberies

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

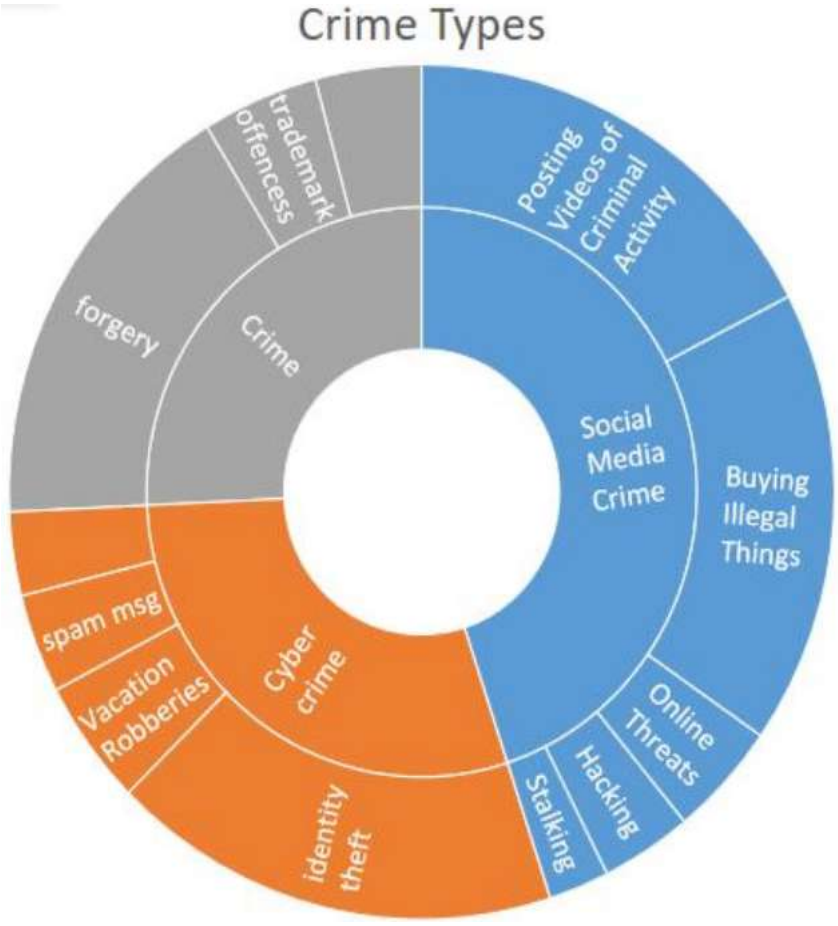

Figure 1: Crime Types

There are no single terms for defining this in the whole world [17, 18]. It is because the content changes as per the events along with changing phases of increasing technology. Remembering these things, current researchers explain cybercrime as "Using computers in the interlinked internet mode for carrying out illegal or unwanted activities". The real cyber terrorism and it would not cause any harm to lives of people as in the case of real time terrorism. Terrorism in an older way means death or loss of property and such kinds of incident will mark an impression of terrorist incidents in the people mind in real time world.

5. Conclusions

The rapid growth of urban populations and the increasing complexity of criminal activities have made traditional crime analysis approaches insufficient for effective law enforcement planning. This review highlights the transformative potential of supervised machine learning techniques in predicting crime rates, identifying hotspots, and supporting data-driven policing strategies. Across the literature, classical algorithms such as Logistic Regression, Support Vector Machines, Decision Trees, and k-Nearest Neighbors have demonstrated their effectiveness in foundational crime classification tasks. At the same time, advanced ensemble methods including Random Forest, Gradient Boosting, and XGBoost consistently outperform

simpler models due to their robustness, ability to capture nonlinear patterns, and suitability for heterogeneous real-world datasets.

The review emphasizes that while supervised machine learning offers powerful tools for intelligent crime prediction, building trustworthy, fair, and generalizable models remains a critical research priority. Future work must focus on integrating explainable AI, fairness-aware modeling, multi-source data fusion, and cross-region transfer learning to develop more accurate and ethically responsible crime prediction frameworks capable of supporting modern smart policing environments.

References

[1] K. Vanitha, V. K, S. R and S. S, "An Intelligent Crime Risk Prediction Framework using Behavioral Analysis and Advanced Machine Learning," *2025 5th International Conference on Soft Computing for Security Applications (ICSCSA)*, Salem, India, 2025, pp. 1268-1273,

[2] J. K. Gupta *et al*., "Predictive Analysis of Crime Rates Using Machine Learning Algorithms," *2024 4th International Conference on Innovative Sustainable Computational Technologies (CISCT)*, Dehradun, India, 2024, pp. 1-6.

[3] V. Keerthika, A. Geetha and D. M. D. Raj, "Predictive Crime Analysis: Statistical Approach to Forecast Crime Hotspots Using Recursive Neural Network in Deep Learning," 2024 Second International Conference on Advances in Information Technology (ICAIT), Chikkamagaluru, Karnataka, India, 2024

[4] S. G. Lilhare, Y. Kumavat, G. Banait and A. Kurkelli, "Crime Hotspots Mapping and FIR Data Interface," 2024 International Conference on Innovations and Challenges in Emerging Technologies (ICICET), Nagpur, India, 2024.

[5] Mussiraliyeva, Shynar, and Gulshat Baispay. "Leveraging Machine Learning Methods for Crime Analysis in Textual Data." *International Journal of Advanced Computer Science & Applications* 15, no. 4 (2024)

[6] A. Sharaff, P. K. Kushwaha, S. P. Dwivedi, O. Krishna, S. Singh and D. Thakur, "Crime Rate Prediction Using Machine Learning," *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, Gautam Buddha Nagar, India, 2023, pp. 787-791

[7] U. Ghani, P. Toth and F. David, "Predictive Choropleth Maps Using ARIMA Time Series Forecasting for Crime Rates in Visegrad Group Countries", *Sustainability*, vol. 15, no. 10, 2023.

[8] S. M. Rajesh, I. Chiranmai and N. Jayapandian, "Machine Learning Based Crime Identification System using Data Analytics," 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA), Theni, India, 2023, pp. 951 – 956.

[9] B. Zhou, L. Chen, S. Zhao, S. Li, Z. Zheng and G. Pan, "Unsupervised Domain Adaptation for Crime Risk Prediction Across Cities," in *IEEE Transactions on Computational Social Systems*, vol. 10, no. 6, pp. 3217 - 3227, doi: 10.1109/TCSS.2022.3207987, Dec. 2023.

[10] Yin, J. (2023). Crime prediction methods based on machine learning: a survey. *Computers Materials & Continua*, 74 (2), 4601–4629.

[11] Bhardwaj, G. and Bawa, R. 2022). Assaying the statistics of crime against women in india using provenance and machine learning models. *International Journal of Advanced Computer Science and Applications*, 13 (7).

[12] A. H. Aiman Awangku Bolkiah, H. Hanin Hamzah, Z. Ibrahim, N. M. Diah, A. Mohd Sapawi and H. M. Hanum, "Crime Scene Prediction Using the Integration of K-Means Clustering and Support Vector Machine," *2022 IEEE 10th Conference on Systems, Process & Control (ICSPC)*, Malacca, Malaysia, 2022, pp. 242-246

[13] A. Thomas and N. V. Sobhana, "A survey on crime analysis and prediction," *Mater Today Proc*, vol. 58, pp. 310–315, Jan. 2022.

[14] H. Al-Ghushami, D. Syed, J. Sessa and A. Zainab, "Intelligent Automation of Crime Prediction using Data Mining," *2022 IEEE 31st International Symposium on Industrial Electronics (ISIE)*, Anchorage, AK, USA, 2022

[15] A. Mohammed et al, "Data Security And Protection: A Mechanism For Managing Data Theft and Cybercrime in Online Platforms Of Educational Institutions ", 2022 International Conference on Machine Learning Big Data Cloud and Parallel Computing (COM-IT-CON), pp. 758 - 761, 2022.

[16] F. Dakalbab, M. Abu Talib, O. Elmutasim, A. Bou Nassif, S. Abbas and Q. Nasir, "Artificial intelligence & crime prediction: a systematic literature review", *Social Sciences & Humanitie*, vol. 6, pp. 100342, 2022.

[17] Baek, Myung-Sun, Wonjoo Park, Jaehong Park, Kwang-Ho Jang, and Yong-Tae Lee. "Smart policing technique with crime type and risk score prediction based on machine learning for early awareness of risk situation." *IEEE Access* 9 (2021): 131906 - 131915.

[18] 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.