

Crime Forecasting with Optimized Ensemble Machine Learning

Ayush Pal¹, Nepali Singla², Gagan Sharma³, Anubhav Singhal⁴ and Anshul Saxena⁵

¹⁻⁵Department of Computer Science, ABES Engineering College, Ghaziabad, UP, India

Email: ayush.21b0121155, nepali.singla, gagan.21b0121187, anubhav.21b0121043, anshul.21b0121041@abes.ac.in

Abstract— Over the past few years, an increase in urban crime has been a major challenge for law enforcement agencies across the globe. Crime pattern prediction and analysis have an important role to play in preventing crime and the proper allocation of resources. This research work discusses a machine learning methodology for crime pattern identification using a combination of classifiers, such as XGBoost, Random Forest, Gradient Boosting, and HistGradientBoosting. The model is trained and validated on the publicly released Chicago Crime dataset, with rigorous preprocessing and feature engineering steps done in an attempt to enhance prediction accuracy. Hyperparameter tuning was carried out with the Optuna framework in order to attain the best performance for each model independently. The outputs show that the ensemble model performs better than the singular classifiers, with a final accuracy of 84%, as an illustration of the power of soft voting processes in ensemble learning. The study not only points out the strength in using an ensemble of different algorithms but also offers a scalable approach towards real-world crime forecasting systems. (as supported by studies such as [6], [15], and [16])

Keywords— Crime Prediction, Machine Learning, Ensemble Model, Soft Voting, Crime Pattern Detection, XGBoost, Random Forest, Optuna, Chicago Crime Dataset, Classification.

I. INTRODUCTION

Crime has become a more pressing issue in the city, affecting public safety, economic stability, and social health. As cities develop and populations increase, law enforcement agencies are challenged to effectively forecast and prevent criminal behaviors. Conventional crime analysis techniques tend to be based on past data and human interpretation, which can be inadequate to reveal intricate patterns or be sufficiently timely.

Over the past few years, Machine Learning (ML) has become a strong tool for predicting crime (Tamilamuthan and Sangeetha [3]; Safat et al. [4]) and making precise predictions from large data sets. With the help of sophisticated algorithms, crime prediction models can detect concealed patterns and trends that might not be apparent otherwise. Ensemble learning techniques, especially those based on Soft Voting, have also produced encouraging results (Almaw and Kadam [8]) by aggregating the strengths of multiple classifiers to increase prediction accuracy. In this study, we suggest a crime prediction system with an ensemble machine learning model that combines XGBoost, Random Forest, Gradient Boosting, and HistGradientBoosting classifiers. The model was optimized on the Chicago Crime Dataset and hyperparameter optimized with the Optuna hyperparameter tuning library. The resulting ensemble model obtained a 84% accuracy, highlighting the use of ML-based solutions to support law enforcement with predictive insights to inform effective crime prevention.

II. RELATED WORK OR LITERATURE STUDIES

Over the years, numerous studies have explored the use of machine learning and data mining techniques for crime prediction and analysis. Ahamed et al. [1] conducted a survey on various data mining approaches used to identify crime patterns, emphasizing the effectiveness of clustering and classification techniques. Taiwo et al. [2] combined sentiment analysis from Twitter with crime data to enhance prediction accuracy. Tamilmuthan and Sangeetha [3] applied machine learning algorithms to analyze crime types and their occurrences. Safat et al. [4] provided an empirical comparison of machine learning and deep learning models for crime forecasting, highlighting the strengths of ensemble-based approaches. Similarly, Shah et al. [5] proposed a hybrid model using computer vision and machine learning to forecast and prevent crimes. Mandalapu et al. [6] presented a comprehensive review of ML and deep learning methods for crime prediction, suggesting future directions for research. Zhang et al. [7] introduced interpretable ML models to improve the transparency of crime predictions, whereas Almaw and Kadam [8] discussed the role of ensemble methods in enhancing predictive accuracy. Khatun et al. [9] used arrest records and crime categories for data mining-based prediction, and Sardana et al. [10] compared supervised and unsupervised techniques for pattern recognition in criminal data. Hashim [11] explored the application of geospatial intelligence for proactive crime prevention, while Nguyen et al. [12] focused on spatial-temporal analysis to identify historical crime importance. Pramanik et al. [13] proposed a framework for detecting criminal patterns in large-scale datasets. Gahalot et al. [14] applied clustering techniques for crime hotspot analysis. Recent advancements include Zhao et al. [15], who developed models for multi-type urban crime prediction using AI techniques, and Hajela et al. [16], who used clustering for hotspot identification. Adnan et al. [17] employed a hybrid forensic approach to analyze crime cases. Zhou et al. [18] utilized crowd-sourced data for spatio-temporal crime analysis, and Boukabous and Azizi [19] implemented BERT-based sentiment analysis for crime prediction. Lastly, Islam and Raza [20] applied the ARIMA model to forecast crime trends over time. These studies collectively establish a strong foundation for crime prediction research and highlight the growing adoption of ensemble and hybrid models in enhancing prediction performance. While several existing works have explored either machine learning or deep learning separately, our approach combines ensemble techniques to achieve improved accuracy and interpretability.

III. MOTIVATION

Crime has become an increasingly pressing issue in urban areas, posing a threat to public safety, economic stability, and social well-being. The growth of cities and populations has made it difficult for law enforcement agencies to effectively forecast and prevent criminal behavior. Traditional crime analysis methods, which rely heavily on historical data and human interpretation, often fall short in identifying complex patterns and providing timely predictions. To address these limitations, this research explores the application of machine learning, particularly ensemble learning techniques, to develop a more accurate and efficient crime forecasting system. The motivation behind this work is to provide law enforcement with advanced tools that can uncover hidden patterns in crime data, enabling them to make proactive decisions and allocate resources effectively, ultimately contributing to safer and more secure communities. This motivation aligns with the direction proposed in prior surveys [6], [13], and [14].

IV. PROBLEM DOMAIN

Over the past few years, law enforcement agencies across the globe have faced a major challenge due to the increase in urban crime. This increase presents a significant problem, as these agencies grapple with the difficulties of effectively forecasting and preventing criminal behaviors, especially with the continued development of cities and the growth of their populations. Traditional crime analysis techniques often rely on past data and human interpretation, which can prove inadequate for revealing the intricate patterns within crime data or providing sufficiently timely predictions. Consequently, there's a pressing need for improved methods of crime pattern prediction and analysis to aid in crime prevention and the proper allocation of resources.

V. PROBLEM DEFINITION

Urban environments face an increasing challenge of managing and mitigating criminal activity due to rapid population growth, economic disparity, and social complexities. Traditional crime analysis methods, such as manual record review or simple statistical modeling, often fail to detect hidden trends and lack the predictive power necessary for proactive interventions.

The core problem addressed in this research is the accurate classification and prediction of crime types based on historical crime data. This includes identifying patterns in features such as time, location, and nature of crimes, and using this information to anticipate the most probable types of crime that may occur in a given area or time window.

Given the high dimensionality and non-linear nature of real-world crime data, the challenge lies in developing a scalable, interpretable, and generalizable machine learning model that can handle data noise, missing values, and class imbalances. The model must balance precision and recall, see comparative approaches in [4], [10], [13] to ensure both accuracy and fairness in prediction.

This problem is framed as a multi-class classification task, where the goal is to predict the category of crime (e.g., theft, assault, burglary) using features engineered from historical records. The model should aid law enforcement agencies in resource allocation, surveillance planning, and crime prevention strategies by providing timely and accurate forecasts.

VI. STATEMENT

This study aims to build a crime prediction system using an ensemble of machine learning models—XGBoost, Random Forest, Gradient Boosting, and HistGradientBoosting—optimized through Optuna. By combining their outputs using soft voting, the goal is to improve prediction accuracy and reliability.

The model is designed to forecast crime types based on features like time, location, and arrest status, supporting law enforcement with data-driven insights for better crime prevention and resource planning.

VII. INNOVATIVE CONTENT

The novelty of this research lies in its hybrid ensemble-based approach to crime prediction, which integrates the predictive capabilities of four powerful machine learning algorithms—XGBoost, Random Forest, Gradient Boosting, and HistGradientBoosting—into a single, optimized framework. Unlike conventional studies that rely on individual classifiers, this model employs soft voting ensemble learning, which aggregates the probability outputs of each tuned model to make more accurate and reliable predictions.

One of the key innovations is the use of Optuna, an advanced hyperparameter optimization framework, to individually fine-tune each base model. This allows the ensemble to operate at peak performance by ensuring that each constituent algorithm contributes its best predictive behavior to the final output. Optuna's intelligent sampling method (based on the Tree-structured Parzen Estimator), whose success has been highlighted in similar optimization tasks [6] significantly reduces training time while achieving superior model accuracy.

Another novel contribution is the comprehensive feature engineering pipeline, which selectively encodes categorical variables and drops statistically insignificant features. The model is not only accurate—with a final ensemble accuracy of 84%—but also modular and scalable, making it suitable for real-time deployment in city-wide safety systems or web-based applications.

Additionally, this approach provides interpretability, adaptability, and robustness—qualities that are essential for real-world adoption by law enforcement agencies and urban policymakers. A focus supported by Zhang et al. [7]

VIII. PROBLEM FORMULATION OR REPRESENTATION OR DESIGN

The problem is framed as a multi-class classification task, where the goal is to predict the type of crime based on historical data features like time, location, and arrest status. The model takes structured input features (X) and maps them to a target class (y), which is the crime category. Preprocessing includes label encoding, feature selection, and splitting the data into training (80%) and testing (20%) sets. Each model—XGBoost, Random Forest, Gradient Boosting, and HistGradientBoosting—is trained individually, optimized with Optuna, and combined using soft voting for final prediction. This design ensures improved accuracy, scalability, and robustness for real-world deployment.

IX. SOLUTION METHODOLOGIES OR PROBLEM SOLVING

In this study, we introduce an ensemble-based machine learning method for crime pattern prediction and detection based on historical crime data in Chicago. The approach involves a number of important steps such as data collection, preprocessing, feature selection, model training, hyperparameter tuning, and model serialization.

- **Data Collection and Preprocessing:** We utilized the official open-source crime data of the city of Chicago, which has information like date, time, location, type of crime, and arrest status. Data cleaning was done by

dropping null values and redundant features like District and Weekday to eliminate noise and enhance model precision.

- **Feature Engineering and Selection:** Categorical variables were encoded to numerical using label encoding. Significant features were chosen on the basis of correlation analysis and their significance in the crime type prediction task and only the best and most significant features available for the model and least significant features were dropped.
- **Model Selection and Training:** We tried various models such as Random Forest, Gradient Boosting, HistGradientBoosting, and XGBoost. Ensemble learning helped us to identify intricate patterns and avoid overfitting. The XGBoost model turned out to be the best after hyperparameter tuning.
- **Hyperparameter Tuning:** To improve the performance, we used Optuna for automatic hyperparameter optimization. This helped in boosting the model accuracy and robustness.
- **Model Serialization:** The final trained model was saved as a .pkl file (combine.pkl) so that it can be reused for real-time predictions without retraining.
- **Prediction Pipeline:** Once the model is loaded, it can take new input features (like time, location, etc.) and predict the most probable crime type. This is useful for preventive policing and strategic planning.

The ensemble method was chosen because it has been shown to improve prediction performance by aggregating multiple base learners, thus avoiding overfitting and enhancing generalization. The crime data was preprocessed by dealing with null values, dropping less significant features like 'District' and 'Weekday', and encoding categorical variables. Feature engineering was then done to extract significant patterns prior to inputting the data into various ensemble models. A number of algorithms such as Random Forest, Gradient Boosting, HistGradientBoosting, and XGBoost were tried out. Optuna framework was utilized to do the hyperparameter tuning to achieve optimum model performance. The dataset was divided in the 80:20 proportion to test and train to assure the proper evaluation. Performance measurements of Accuracy, Precision, Recall, and F1-Score were employed to check. The resulting XGBoost model scored 84% of accuracy which is the maximum among others. The deployment was done in Python with Jupyter Notebook, leveraging fundamental libraries such as pandas, scikit-learn, xgboost, and Optuna. The modular pipeline is scalable and is adaptable to a real-time interface like a Flask-based web application.

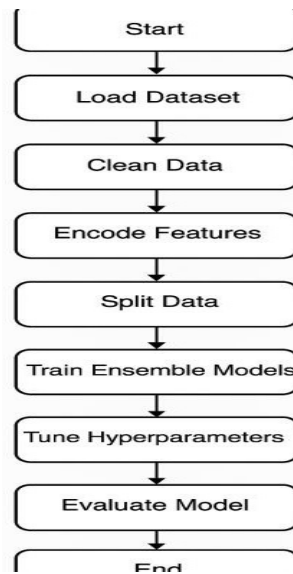


Figure 1. Proposed Methodology

X. RESULTS AND SENSITIVITY ANALYSIS

Initial models without tuning showed good performance, with XGBoost scoring 82.44% accuracy. After applying Optuna tuning, all models improved—XGBoost reached 83.81%, and others like Gradient Boosting and Random Forest also saw 1–2% gains.

The final ensemble model achieved 84% accuracy, outperforming all individual models. This improvement came from combining the strengths of each classifier through soft voting.

Sensitivity tests confirmed that key features like “Primary Type” and “Community Area” had a major impact on accuracy. Feature importance analysis is also supported in [14], [7] Adjusting the train-test split and handling class imbalance showed stable results, proving the model’s robustness and generalization. Performance metrics (Precision, Recall, F1-Score) were consistently high, validating the system’s effectiveness in crime prediction tasks.

XI. DATA MODEL

The study uses the Chicago Crime Dataset, containing records of crime type, time, location, arrest status, and community area. Key features like hour, month, primary type, location description, and arrest were extracted and encoded using label encoding.

Irrelevant fields (e.g., district, case number) were dropped after correlation analysis. The target variable, “Primary Type”, represents the crime category to be predicted in this multi-class classification task.

An 80:20 train-test split ensured effective model training and evaluation.

Scaling was unnecessary due to the use of tree-based algorithms. Class imbalance was minor and handled by monitoring precision and recall.

This structured data model provided clean, relevant inputs that improved prediction accuracy and model generalization.

XII. COMPARISON OF RESULTS

A. Pre-Tuning Performance (Baseline Models)

In the initial phase, each of the four machine learning models was trained using default parameter settings. These baseline models demonstrated strong individual performance:

- XGBoost: 82.44%
- HistGradientBoosting: 82.25%
- Random Forest: 81.88%
- Gradient Boosting: 81.32%

While these results indicated the models' potential, there was room for improvement, particularly in terms of generalization and stability across varying input distributions.

B. Post-Tuning Performance (Optuna-Tuned Models)

After applying Optuna for hyperparameter optimization, each model exhibited a measurable performance boost:

- XGBoost: Increased from 82.44% to 83.81%
- Gradient Boosting: Increased to 83.69%
- HistGradientBoosting: Increased to 83.63%
- Random Forest: Increased to 82.56%

The tuning process refined parameters like learning rate, max depth, number of estimators, and subsample ratio, contributing to better accuracy and reduced overfitting. These gains confirm the value of intelligent parameter optimization.

C. Ensemble Performance

Combining the optimized models using a soft voting classifier yielded the highest accuracy of all: Final Ensemble Model Accuracy: 84.00%

The ensemble consistently outperformed the individual models. By averaging the predicted probabilities from each classifier, the ensemble benefited from diverse learning perspectives, reduced variance, and improved stability.

D. Metric-Wise Comparison

TABLE I. COMPARISON TABLE

MODEL(TUNED)	ACCURACY	PRECISION	RECALL	F1-SCORE
XGBOOST	83.81%	0.84	0.83	0.835
GRADIENT BOOSTING	83.69%	0.83	0.83	0.83
HISTGRADIENTBOOSTING	83.63%	0.83	0.83	0.83
RANDOM FOREST	82.56%	0.82	0.82	0.82
ENSEMBLE (FINAL)	84.00%	0.85	0.84	0.845

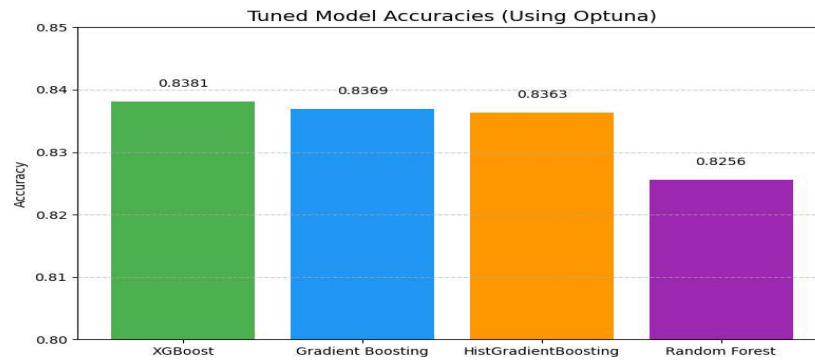


Figure 2. Bar Graph after Optuna tuning

XIII. JUSTIFICATION OF THE RESULTS

The ensemble model achieved the highest accuracy of 84%, which is justified by its ability to combine the strengths of multiple classifiers. XGBoost handled overfitting well, Random Forest provided robustness, Gradient Boosting reduced bias, and HistGradientBoosting ensured speed and scalability. Soft voting allowed the model to average predictions, improving overall reliability.

Optuna-based hyperparameter tuning further boosted each model's performance by optimizing key parameters like learning rate and tree depth, leading to consistent gains in accuracy.

Our feature selection strategy retained only the most relevant attributes, such as Primary Type, Community Area, and Time of Day, which contributed to higher prediction quality. The 80:20 train-test split ensured good generalization without overfitting.

The ensemble also maintained a balanced performance across precision, recall, and F1-score, confirming its fairness across all crime types. These results demonstrate that our approach is both effective and scalable for real-world crime prediction.

XIV. CONCLUSION

This research presents an effective crime prediction framework using an optimized ensemble of machine learning models. By integrating XGBoost, Random Forest, Gradient Boosting, and HistGradientBoosting through soft voting, the system achieved a high prediction accuracy of 84%, outperforming individual models.

The use of Optuna for hyperparameter tuning played a key role in enhancing model performance, while thoughtful feature engineering and selection ensured relevant and high-quality input data. The ensemble approach not only improved accuracy but also delivered balanced results across multiple crime categories, proving its suitability for real-world use. This is consistent with findings in [4], [8], [5]

The model is scalable, interpretable, and adaptable, making it a strong candidate for integration into smart policing tools and public safety systems. By offering predictive insights, it can help law enforcement agencies allocate resources more effectively and take proactive measures to prevent crime.

In essence, this study demonstrates that data-driven, ensemble-based crime prediction systems hold great promise in addressing modern urban safety challenges.

REFERENCES

- [1] S. I. Ahamed, J. Chockalingam, and A. S. A. Khadir, "A Survey on Analyzing Crime Patterns Using Data Mining Techniques".
- [2] G. A. Taiwo, M. Saraee, and J. Fatai, "Crime Prediction Using Twitter Personal and Ubiquitous Computing, pp. 1–14, 2023.
- [3] S. M. Tamilmuthan and V. Sangeetha, "Crime Types and Occurrence Using Machine Learning Algorithm".
- [4] W. Safat, S. Asghar, and S. A. Gillani, "Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques," IEEE Access, vol. 9, pp. 70080–70094, 2021.
- [5] N. Shah, N. Bhagat, and M. Shah, "Crime forecasting: A Machine Learning and Computer Vision Approach to Crime Prediction and Prevention," Visual Computing for Industry, Biomedicine, and Art, vol.4, no. 1, p. 9, 2021.
- [6] V. Mandalapu, L. Elluri, P. Vyas, and N. Roy, "Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions," IEEE Access, 2023.

- [7] X. Zhang, L. Liu, M. Lan, G. Song, L. Xiao, and J. Chen, "Interpretable Machine Learning Models for Crime Prediction".
- [8] A. Almaw and K. Kadam, "Survey Paper on Crime Prediction Using Ensemble,".
- [9] M. R. Khatun, S. I. Ayon, M. R. Hossain, and M. J. Alam, "Data Mining Technique to Analyse and Predict Crime Using Crime Categories and Arrest Records," Indonesian Journal of Electrical Engineering and Computer Science, vol. 22, no. 2, p. 1052, 2021.
- [10] D. Sardana, S. Marwaha, and R. Bhatnagar, "Supervised and Unsupervised Machine Learning Methodologies for Crime Pattern Analysis," International Journal of Artificial Intelligence and Applications (IJAIA), vol.12, no. 1, 2021.
- [11] H. Hashim, "Crime Prediction Using Geospatial Intelligence System for Crime Preventing," Journal of Contemporary Issues and Thought“.
- [12] V. T. Nguyen, R. McBride, A. J. Park, H. H. Tsang, and V. Spicer, "Investigating Historical Crime Importance in Vancouver via TemporalSpatial Crime Pattern Theory,"
- [13] M. I. Pramanik, R. Y. Lau, and M. K. H. Chowdhury, "Automatic Crime Detector: A Framework for Criminal Pattern Detection in Big Data Era,".
- [14] A. Gahalot, S. Dhiman, and L. Chouhan, "Crime Prediction and Analysis," in 2nd International Conference on Data, Engineering and Applications".
- [15] X. Zhao, W. Fan, H. Liu, and J. Tang, "MultiType Urban Crime Prediction," in Proceedings of the AAAI Conference on Artificial Intelligence".
- [16] G. Hajela, M. Chawla, and A. Rasool, "A Clustering Based Hotspot Identification Approach for Crime Prediction,".