International Journal of Engineering Research and Advanced Technology (IJERAT)

DOI: 10.31695/IJERAT.2025.1.1

E-ISSN : 2454-6135

Volume. 11, No.1
January – 2025

Machine Learning-Driven Crime Risk Modelling and Prediction

Mwita Isaac Makene[1], and Stanley Leonard Tito[1]

Computer Science and Engineering Department

Mbeya University of Science and Technology, Mbeya

Tanzania

ABSTRACT

Effective law enforcement and public safety strategies are essential for accurate crime analysis, and forecasting. This situation is a global challenge, especially in the region where diverse socioeconomic factors can influence complex crime patterns. This study presents a novel machine learning-based approach to address this critical issue. By leveraging publicly available historical crime data and relevant socio-demographic variables, we develop a predictive model to identify crimes in the area with high rates. The methodology involves data preprocessing, feature selection, model training, and validation using advanced machine learning techniques. Our proposed method attains a prediction accuracy of 0.52 over other competitive methods. These empirical results demonstrate the efficacy of the approach in forecasting crime hotspots, providing actionable insights for law enforcement agencies and policymakers. This research contributes to enhancing proactive measures for crime prevention and resource allocation in regions with diverse social economic factors.

Keywords: Crime Risk Modelling, Predictive Model, Random Forest, Machine Learning, XGBoost.

1. INTRODUCTION

After the Second World War, the world experienced a sequence of crime incidences that were deeply related to the nation's socioeconomic factors[1]. The regions characterized by diverse socioeconomic factors such as poverty, child sexual exploitation, and unemployment can influence complex crime patterns through criminal activities. Factors such as poverty, unemployment, and stark social inequalities are primary contributors. Additionally, recent years have seen a concerning rise in drug-related offenses and cybercrimes [2], illustrating the evolving nature of criminal behaviour in diverse regions. Different regions across the world have different global crime landscapes. For example, according to the Global Initiative's 2023 report, Tanzania ranks 42nd globally for overall criminality among 193 countries, placing it 11th in Africa and 5th in the East African region [3-5]. Likewise in Canada, the crime rate has increased by 2% for three consecutive years up to 2023 with about 21,417 reported child pornography incidents as shown in Figure 1. [2]. Limited studies have reported crime mitigation strategies for global crime incidents. This paper aims to present a mitigation strategy that employs machine learning algorithms to help address the crime rate incidences.

Understanding the global crime landscape of the Nation is pivotal for developing effective strategies that will promote sustainable development and enhance citizen well-being [6-8].

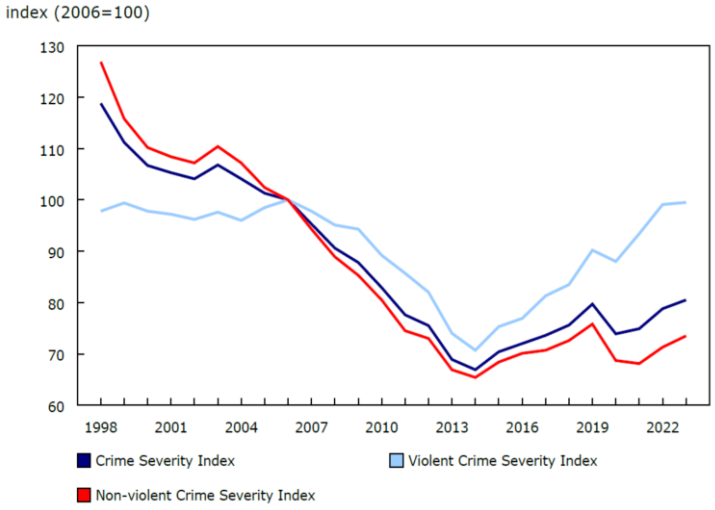

Figure 1: Police-reported Crime Severity Indexes, 1998 to 2023 Source Police Report in Canada,[2].

This study proposes an effective crime predictive model that analyzes historical crime data alongside socio-economic and geographical factors by utilizing machine learning algorithms. This holistic approach aims to generate predictive insights into areas prone to high criminal occasions [9-11]. By providing law enforcement agencies with actionable intelligence, the model seeks to facilitate proactive interventions and strategic resource allocation to improve public safety [12].

2. METHOD

2.1 Model design

Exploration of machine learning algorithms in this study led the supervised learning as a compelling approach for crime analysis and prediction[13]. This choice finds its foundation in the very nature of the task of analyzing labelled historical data for crime risk prediction. The approach unravels patterns and relationships within the provided labelled data by examining the relationships among the data to allow the model to learn and link input features, such as location, temporal factors, and demographic characteristics with their corresponding crime outcomes. The architecture of the method is shown in (Figure 2).

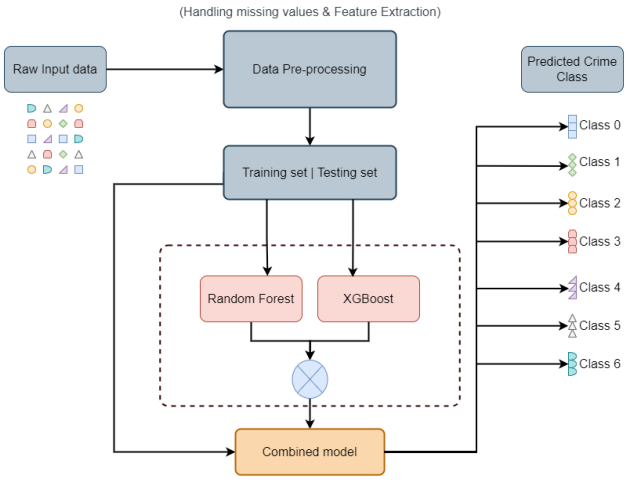

Figure 2: Architecture of the Learning Method

2.2 Feature engineering

The data pre-processing stage included a crucial step to address missing values. Missing values, often arising from sensor malfunctions or incomplete records, can significantly impact subsequent model training and data analysis. We

DOI : 10.31695/IJERAT.2025.1.1

utilized the median strategy approach available in the scikit-learn library to efficiently maintain data integrity and facilitate further analysis.

A thorough feature engineering process was conducted to improve the performance of our model by selecting and transforming relevant features to feature importance as shown in the (Figure 3).

We also worked to create new derived features based on domain knowledge and data exploration. For example, we calculated the rate of change for certain sensor readings over time, which helped to capture temporal dynamics that were otherwise not evident in the raw data. We also combined multiple features through arithmetic operations or categorical transformations to uncover higher-order relationships and patterns.

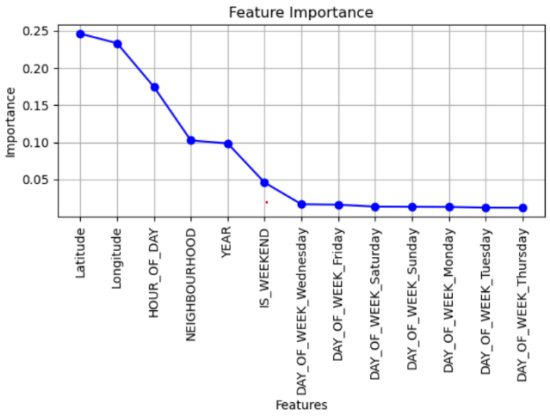

Figure 3: Feature importance attributes used in model training

The important values were derived from the model's ability to predict high-crime-risk areas accurately. For example, the latitude feature emerged as the most significant, with an importance value of approximately 0.25 which indicates that the geographical latitude of an area strongly influences the likelihood of it being a high crime risk zone. Similarly, the Day of the Week signifies that, the importance values are relatively low, hovering around 0.01 which is of less significant compared to other features.

2.3 Model Training

Random Forest (Figure 3) and XGBoost (Figure 4) machine learning algorithms were examined in constructing a crime risk analysis and predictive machine learning model. Initially, we trained the single Random Forest-based model illustrated in equation (1) followed by the XGBoost based model illustrated in equation (2). The results were recorded and presented in (Table 3)

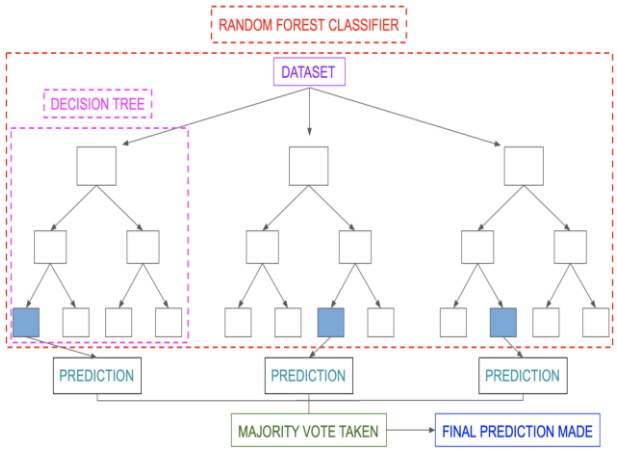

Figure 4: Structure of Random Forest classifier

The random forest model is presented in equation (1) as;

DOI : 10.31695/IJERAT.2025.1.1

$$y_{R,i} = \frac{1}{N_T}\sum_{i=1}^{N_T} f_i^R(X_i) \quad \ldots\ldots(1)$$

Where;

N_T = Number of trees in the Random Forest model

$f_j^R(xi)$ = Prediction of the ith tree

Each tree $f_j^R(xi)$ is trained on a bootstrapped sample of the data and splits nodes based on a random subset of features, which ensures diverse trees to reduce the overall variance in predictions.

XGBoost model was trained based on gradient boosting principle, which is an ensemble strategy for constructing a strong predictive model by sequentially building several weak learners, often decision trees. XGBoost distinguishes itself with numerous major features that improve performance and scalability. These include improved regularization algorithms to prevent overfitting, parallelized tree for quicker computation, and the ability to handle missing values internally, which makes it very useful for large and complicated datasets.

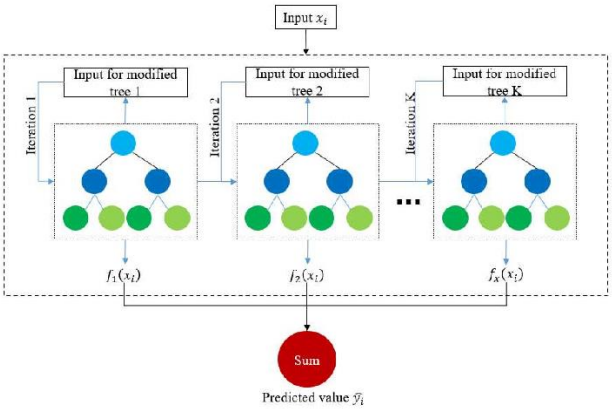

Figure 5: Structure of XGBoost classifier

We trained each decision trees under XGBoost sequentially, to capitalize the errors made by the previous tree. This iterative process minimizes the loss function, which measures the difference between the actual and predicted values, thus improving the model's accuracy over time. The XGBoost model is expressed in equation (2) as;

$$y_{G,i} = \sum_{k=1}^{N_T} \mu^k f_k^G(x_i) \quad \ldots\ldots(2)$$

Where;

N_T = Number of trees in the XGBoost model

μ^k Learning rate for the k^{th} tree

$f_k^G(xi)$ = Prediction of the k^{th} tree

XGBoost model was optimized using a combination of gradient descent. This minimized the loss function and enhanced regularization to prevent overfitting. The parameter μ^k adjusts the contribution of each tree, balancing the model's learning pace and robustness. To leverage the strengths of both the Random Forest and XGBoost models[14], a weighted average method was employed to combine their prediction capabilities as shown in equation (3). This approach allowed to mitigate the faults of individual models and improve the overall prediction accuracy.

$y_{c,i}=\alpha.\ y_{R,i}+(1-\alpha).$

$y_{G,i}$ ……(3)

This led to an expanded hybrid model expressed in equation (4) as;

DOI : 10.31695/IJERAT.2025.1.1

$$y_{C,i} = \alpha \cdot \frac{1}{N_T} \sum_{j=1}^{N_T} f_j^R(x_i) + (1 - \alpha) \cdot \sum_{k=1}^{N_T} \mu^k f_k^G(x_i)$$

……..……..(4)

Where;

α = Variable which controls the Random Forest and XGBoost predictions
N_R = Number of trees in the Random Forest
N_B = Number of trees in the XGBoost
$f_j^R(xi)$ = Prediction of the j^{th} tree from Random Forest
$f_k^G(xi)$ = Prediction of the kth tree from XGBoost.

The variable α ranges between 0 and 1, and it determines the emphasis on the Random Forest predictions versus the XGBoost predictions. By tuning α, we find an optimal balance that yields the best predictive performance at an accuracy of 0.52 as shown in (Table 3). This approach resulted in a robust and accurate prediction demonstrates improved performance compared to individual model performance.

3. EXPERIMENTS

3. 1 Experiment Setup

Developing a machine learning model for crime prediction required meticulous attention to computational resources, encompassing hardware, software, and programming languages. We employ Python 3.12.1 libraries (Scikitlearn, Pandas, NumPy and Matplotlib) in Jupyter Notebook 7.0.8 environment for interactive development and experimentation.

3.2 Experimental Process

Raw Input Data: Data from diverse areas were used. The dataset spans a period of 14 years, from 2003 to 2017, and comprises 435,422 rows. This raw input data forms the foundation upon which all subsequent steps were built, requiring thorough understanding, and pre-processing to ensure its suitability for analysis and modeling.

Data Understanding: The first phase in the data understanding process involved integrating the raw data to create a cohesive dataset. This integration was followed by a detailed feature description phase, where each feature within the dataset was meticulously described and understood. Critical columns such as 'MONTH', 'DAY', 'HOUR', and 'MINUTE' were examined to gain insights into their characteristics and distributions, setting the stage for effective pre-processing and feature engineering.

Data pre-processing: In the data pre-processing phase, feature selection was a key activity. This step focused on handling missing values, a common challenge to large datasets. The SimpleImputer class from scikit-learn was employed, using the 'median' strategy to replace missing entries in relevant columns. This approach ensured that the dataset was complete and ready for the subsequent stages of analysis. Following feature selection, feature creation was undertaken, involving the development of new features and the transformation of existing ones to better represent the underlying relationships within the data.

Modelling: The modeling phase involved training multiple models to predict outcomes using the pre-processed and engineered features. Two prominent models, Random Forest, and XGBoost, were utilized in this step. These models were trained on the dataset, leveraging their unique strengths to capture different aspects of the data. Following the individual model training, a combined model was created to integrate the predictions of both Random Forest and XGBoost, aiming to enhance overall performance and accuracy.

Evaluation: The evaluation phase was critical in assessing the performance of the trained models. Various evaluation metrics were employed, including accuracy, F1 score, precision, recall, and ROC-AUC scores. These metrics provided

DOI : 10.31695/IJERAT.2025.1.1

a comprehensive view of how well the models distinguished between different crime types, offering insights into their effectiveness and reliability. The evaluation results highlighted the strengths and weaknesses of each model, guiding further refinements and improvements.

This section provides all the necessary procedures to conduct experiments. We present clear conditions of the experiments and simulation environments that may assist other researchers in replicating the experimental results. We presented a diagram that shows the flow of experiments, arrangements, and settings in Figure 6.

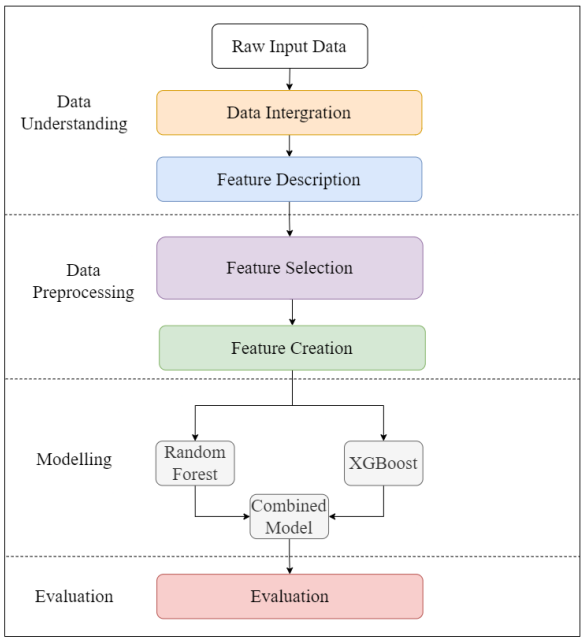

Figure 6: Experiment Process

4. RESULTS AND DISCUSSION

The dataset used was a comprehensive collection of crime-related information spanning several years, sourced from various region. Key statistics of the dataset were as follows:

Total number of data points: 435,422

Types of crimes: Theft from vehicle, mischief, break and enter residential/other, offence against a person and other types of theft.

Geographic distribution: Covers multiple regions.

Temporal distribution: Data spans multiple years, capturing seasonal and temporal trends (2003-2017).

Table 1: Dataset splitting

Dataset	Dataset (%)	Dataset in figures
Training Set	80%	348,338
Testing Set	20%	87,084
Total	100%	435,422

4.1 Model Performance

The combined model was trained on the training dataset (x_train, y_train) and then used to predict on both the training and test datasets (x_test, y_test). The performance metrics were calculated for both datasets.

The accuracy of the model on the training dataset was found to be 0.79, indicating that the model has learned the patterns in the training data effectively. The detailed classification report showed high precision, recall, and F1-scores across all classes, further confirming the model's strong performance on the training data as shown in (Table 2).

DOI : 10.31695/IJERAT.2025.1.1

Table 2: Classification report on training set

Class	Combined Model Training Classification Report			
	Precision	Recall	F1 Score	Support
0	0.89	0.56	0.69	26,890
1	0.87	0.74	0.80	48,804
2	0.93	0.62	0.74	56,306
3	0.91	0.92	0.92	41,658
4	0.70	0.98	0.81	136,734
5	0.97	0.39	0.56	20,464
6	0.98	0.53	0.69	17,481
Accuracy			**0.79**	**348,337**
Macro Avg	**0.89**	**0.68**	**0.74**	**348,337**
Weighted Avg	**0.83**	**0.79**	**0.78**	**348,337**

The accuracy on the test dataset was slightly lower than on the training dataset as shown in (Table 3), suggesting some generalization error, which is expected. The classification report for the test dataset also showed good precision, recall, and F1-scores, though slightly lower than the training dataset, indicating that the model performs well on unseen data but with some room for improvement.

Table 3: Classification report on testing set

Class	Combined Model Training Classification Report			
	Precision	Recall	F1 Score	Support
0	0.49	0.22	0.30	6952
1	0.44	0.35	0.39	12055
2	0.38	0.14	0.20	13854
3	0.80	0.76	0.78	10505
4	0.50	0.85	0.63	34157
5	0.50	0.06	0.11	5156
6	0.67	0.11	0.19	4406
Accuracy			**0.52**	**87085**
Macro Avg	**0.54**	**0.36**	**0.37**	**87085**
Weighted Avg	**0.52**	**0.32**	**0.47**	**87085**

4.2 Evaluation metrics

To evaluate and compare the performance of the machine learning models for the predictive task, we considered two classifiers Random Forest and XGBoost. Each model was assessed using multiple evaluation metrics, including accuracy, precision, F1-score, and recall as shown in (Table 4).

DOI : 10.31695/IJERAT.2025.1.1

Table 4: Accuracy, Precision, F1 score, and Recall for Random Forest classifier and XGBoost Classifiers

Evaluation Set	Techniques		
	Random Forest Classifier	XGBoost Classifier	Combined Model
Accuracy	0.48	0.53	0.52
Precision	0.45	0.52	0.52
F1 score	0.48	0.53	0.52
Recall	0.45	0.48	0.47

XGBoost outperforms individual classifiers with high accuracy and F1 score, indicating a balance between precision and recall. The Combined Model offers balanced performance, although not as good, highlighting the potential benefits of ensemble methods.

4.3 Multiclass Crime Prediction

The ROC (Receiver Operating Characteristic) curves was used to evaluate the performance of multiclass crime prediction model. Using the One-vs-Rest strategy, where a separate ROC curve was computed for each class.

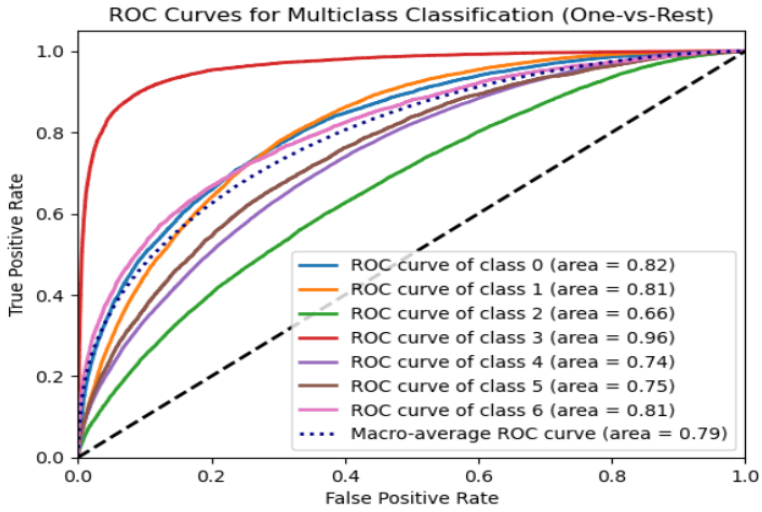

Figure 7: Multiclass ROC Curve

The ROC curves in (Figure 7) illustrate the performance of a multiclass classification model across seven distinct classes. Each curve represents the model's ability to discriminate a specific class from the others. The area under each curve (AUC) quantifies this discrimination capability, with higher AUC values indicating better performance. Notably, class 3 exhibits exceptional performance with an AUC of 0.96, while class 2 shows the lowest performance with an AUC of 0.66.

To assess the overall model performance, a macro-average ROC curve is computed by averaging the true positive rates across all classes for each unique false positive rate. This aggregated curve yields a macro-average AUC of 0.79, summarizing the model's average discriminatory power. This comprehensive analysis allows for identifying the model's strengths and weaknesses across different classes, guiding targeted improvements for enhanced multiclass classification accuracy.

4.4 Model 1 Comparison with the state-of-the-art models

To further evaluate the effectiveness of various machine learning techniques in the context of crime prediction, we conducted a comparative analysis of multiple models. These models include various machine-learning techniques. In (Table 5) below, Model-1 represented Random Forest (RF), Model-2 was XGBoost (XGB), Model-3 was Logistic Regression (LR), Model-4 was LightGBM, and Model-5 was Hybrid technique.

DOI : 10.31695/IJERAT.2025.1.1

The green ticks, symbolize that the particular model used that respective technique while the red cross symbolizes that the respective technique was not used in that specific model. This comparison provides insights from three different models, Model 1 being our own and Model 2 and 3 obtained from [15-16] respectively

Table 5: Comparison with the state-of-the-art models

Model Name	Technique Used					Accuracy	F1 Score	Precision	Recall
	1	2	3	4	5				
Model-1	✓	✓	✗	✗	✓	0.52	0.52	0.52	0.47
Model-2	✓	✗	✓	✓	✗	0.42	0.29	0.31	0.43
Model-3	✗	✓	✗	✗	✗	0.50	0.50	0.48	0.54

The (Figure 7) below shows that our combined model attains a competitive accuracy and maintains strong precision and recall values compared to the models by [15] and [16].

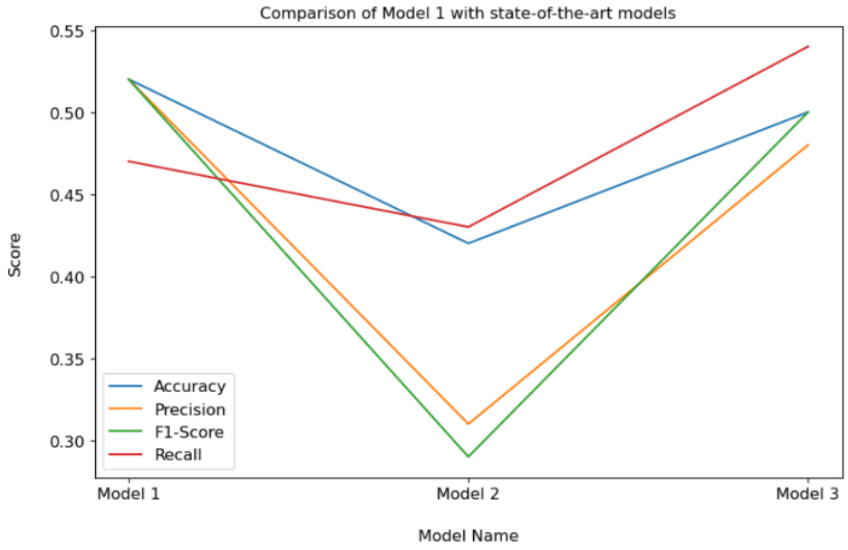

Figure 1.7: Comparisons with the state-of-art models

This analysis underscores the advantage of using ensemble methods to improve model performance and achieve a competitive crime prediction.

5. CONCLUSION

This study demonstrates that ensemble methods, are highly effective in predicting high crime risk areas. The superior performance of the combined model suggests that integrating multiple classifiers can enhance predictive accuracy.

DOI : 10.31695/IJERAT.2025.1.1

These findings highlight the potential of advanced machine learning techniques to improve public safety and resource allocation.

Despite the challenges and complexities in developing effective crime risk predictive models, this research underscores the transformative potential of machine learning in crime prevention. The insights gained offer a data-driven approach to enhance public safety and inform future research in crime prediction.

Future research should focus on enhancing dataset, model accuracy and exploring additional variables that may influence crime risk. Continued efforts in this area can further improve predictive models and contribute to safer communities. Speculatively, integrating real-time data and enhancing model interpretability could be valuable steps forward.

REFERENCES

[1] Burton and Andrew, "Jamii ya wahalifu. The growth of crime in a colonial African urban centre: Dar es Salaam, Tanganyika, 1919-1961," *http://journals.openedition.org/chs*, vol. 8, no. Vol. 8, n°2, pp. 85–115, Nov. 2004, doi: 10.4000/CHS.465.

[2] S. Conroy, "Recent trends in police-reported clearance status of sexual assault and other violent crime in Canada 2017 to 2022", *Juristat: Canadian Centre for Justice Statistics*, 1-44.

[3] S. Caneppele and A. da Silva Cybercrime. In Research handbook of comparative criminal of justice (pp. 243-260). Edward Elgar Publishing. DOI: https://doi.org/10.4337/9781839106385.00024

[4] R. Nord, Y. Sobolev, D. G. Dunn, A. Hajdenberg, and N. Hobdari, "EliScholar-A Digital Platform for EliScholar-A Digital Platform for Tanzania: The Story of an African Transition Tanzania: The Story of an African Transition," 2009. [Online]. Available: https://elischolar.library.yale.edu/ypfs-documents/10223

[5] "Crime worldwide - Statistics & Facts | Statista." Accessed: Jan. 31, 2024. [Online]. Available: https://www.statista.com/topics/780/crime/#topic Overview

[7] "The Global Organized Crime Index 2023 | Global Initiative." Accessed: Jul. 29, 2024. [Online]. Available: https://globalinitiative.net/analysis/ocindex-2023/

[8] Global Initiative, "Global Organized Crime Index 2021," *Global Organized Crime: A 21st Century Approach: Second Edition*, 2021.

[9] Y. Lu, "Crime Prediction Utilizing ARIMA Model," *BCP Business & Management*, vol. 38, 2023, doi: 10.54691/bcpbm.v38i.3721.

[10] S. Sharma, B. K. Rai, G. Kumar, A. Prajapati, and V. Kumar, "Crime Visualization and Forecasting Using Machine Learning," in *Lecture Notes in Electrical Engineering*, Springer Science and Business Media Deutschland GmbH, 2023, pp. 307–320. doi: 10.1007/978-981-99-2058-7_28.

[11] S. Wu, C. Wang, H. Cao, and X. Jia, "Crime prediction using data mining and machine learning," *Advances in Intelligent Systems and Computing*, vol. 905, pp. 360–375, 2020, doi: 10.1007/978-3-030-14680-1_40.

[12] X. Zhang, L. Liu, L. Xiao, and J. Ji, "Comparison of machine learning algorithms for predicting crime hotspots," *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.3028420.

[13] M. Khan, A. Ali, and Y. Alharbi, "Predicting and Preventing Crime: A Crime Prediction Model Using San Francisco Crime Data by Classification Techniques," *Complexity*, vol. 2022, 2022, doi: 10.1155/2022/4830411.

DOI : 10.31695/IJERAT.2025.1.1

[14] Y. Lamari, B. Freskura, A. Abdessamad, S. Eichberg, and S. de Bonviller, "Predicting spatial crime occurrences through an efficient ensemble-learning model," *ISPRS Int J Geoinf*, vol. 9, no. 11, 2020, doi: 10.3390/ijgi9110645.

[15] A. Alsubayhin, M. S. Ramzan, and B. Alzahrani, "Crime Prediction Model using Three Classification Techniques: Random Forest, Logistic Regression, and LightGBM," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 1, 2024, doi: 10.14569/IJACSA.2024.0150123.

[16] J. Alghamdi and T. Al-Dala'in, "Towards spatio-temporal crime events prediction," *Multimed Tools Appl*, vol. 83, no. 7, 2024, doi: 10.1007/s11042-023-16188-x.

DOI : 10.31695/IJERAT.2025.1.1