https://dx.doi.org/10.4314/dujopas.v10i4a.34
ISSN (Print): 2476-8316
ISSN (Online): 2635-3490
Dutse Journal of Pure and Applied Sciences (DUJOPAS), *Vol. 10 No. 4a December 2024*

A Hyper-parameter Tuned Random Forest Algorithm-Based on Artificial Bee Colony for Improving Accuracy, Precision and Interpretability of Crime Prediction

Hauwa Abubakar*, Prof. Souley Boukari,
Prof . Abdulsalam Ya'u Gital, Dr. Fatima Umar Zambuk

Department of Computer Science ,
Faculty of Science,
Abubakar Tafawa Balewa University Bauchi,
Bauchi State
Nigeria.

E-mail: bbkrhauwa@gmail.com

Abstract
Crime prediction plays a crucial role in enhancing public safety and optimizing resource allocation for law enforcement. Traditional methods often fall short in addressing the complex and dynamic nature of crime data, relying on oversimplified assumptions and limited datasets that reduce accuracy and effectiveness. Advanced machine learning techniques, particularly a hyper-parameter tuned Random Forest model optimized using Artificial Bee Colony (ABC) algorithms, present a promising solution. This study proposes an enhanced crime prediction methodology that incorporates ABC-based hyper-parameter tuning and Recursive Feature Elimination with Cross-Validation (RFECV) to improve accuracy, interpretability, and robustness. The model leverages ensemble techniques to integrate diverse features from historical crime data, capturing intricate crime patterns more effectively. Performance evaluations will compare the proposed approach with existing models using metrics such as Predictive Accuracy Index (PAI), Predictive Efficiency Index (PEI), Recapture Rate Index (RRI), and SHapley Additive exPlanations (SHAP) values. By prioritizing accuracy, transparency, and stakeholder engagement, this research aims to develop reliable, interpretable, and data-driven crime prediction models, fostering informed decision-making and proactive crime prevention.The work emphasizes improving crime prediction models through advanced machine learning techniques, including enhanced model development, integration of diverse data sources, focus on interpretability, continuous optimization, and stakeholder engagement. These recommendations aim to create robust, interpretable, and data-driven models that support law enforcement decision-making while addressing biases and existing limitations.

Keywords: Random Forest, Artificial Bee Colony,Interpretability, Crime Prediction.

INTRODUCTION

Crime prediction leverages criminology, data science, machine learning, and sociology to forecast criminal activities using historical data. While advancements in big data and AI have enhanced predictive accuracy, challenges persist with dynamic crime data, including issues of over-fitting, under-fitting, and the lack of transparency in "black-box" models like deep learning (Du & Ding, 2023; Adeyemi *et al*., 2021; Zhang *et al.*, 2023).

Traditional models often face challenges in achieving reliable results due to the dynamic nature of crime data, which can lead to issues such as over-fitting or under-fitting (Alsayadi

Author for Correspondence

et al., 2022). Crime prediction methods include traditional approaches and modern machine learning techniques. Traditional methods, such as Geographic Information Systems (GIS), crime density mapping, regression analysis, and spatial clustering algorithms, are useful for identifying crime hot spots and analyzing spatial patterns (Kedia, 2016). However, they are limited by oversimplified assumptions, neglect of non-linearity, exclusion of diverse data sources, and a lack of predictive power (Wang *e*t *al*., 2023). Modern machine learning techniques, like Decision Trees (DTs) and Random Forests (RFs), address these limitations by handling complex relationships, enhancing accuracy, and reducing overfitting (Khan *et a*l., 2022; Oh *et a*l., 2022). Integrating machine learning and ensemble methods offers improved accuracy, interpretability, and predictive capabilities, supporting proactive crime prevention efforts. (Yao *et al*., 2020) .

This work aims to improve crime prediction by developing an enhanced Random Forest (RF) model. The key objectives are to Optimize Accuracy and Precision by the Use of Artificial Bee Colony (ABC) algorithm for hyper-parameter tuning, and Enhance Interpretability by applying Recursive Feature Elimination with Cross-Validation (RFECV) to improve model transparency and feature selection. These advancements target a more accurate, precise, and interpretable crime prediction model. The Taxonomy of Crime Prediction approaches can be categorized based on data sources, prediction techniques, types of crimes, and temporal/spatial dimensions. The taxonomy of the crime prediction as obtained from literature is illustrated in Figure 1.

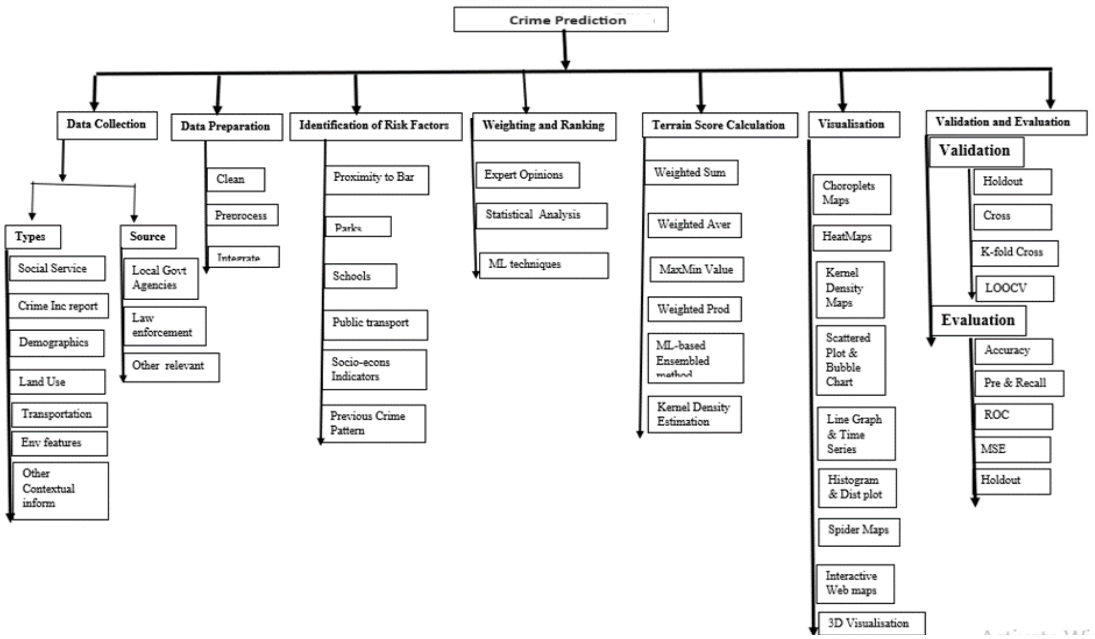

Figure 1. Taxonomy of crime prediction

Machine learning significantly enhances crime prediction by employing advanced algorithms to analyze data and provide predictive insights. Key applications like Predictive Modeling its on the view that Machine learning models use historical data and influential factors (e.g., time, day, environment) to forecast future crime occurrences with improved accuracy (Wheeler & Steenbeek, 2021). and key applications of Machine Learning models is good Feature selection, Identifying the most relevant features improves model performance, reduces overfitting, and optimizes computation time (Ahamad *et al*., 2022). These approaches help create more effective and efficient crime prediction systems.

METHODOLOGY

Material and methods

Limitations identified from the existing models includes Overfitting, Many models, like decision trees and support vector machines, are prone to overfitting without proper tuning *(Liao et al., 2022).*

Poor Interpretability: The opaque nature of ensemble methods, such as Random Forests, hinders understanding of predictions (Pfob *et a*l., 2022). Insufficient Optimization: Lack of effective hyper-parameter tuning affects model reliability and efficiency (Passos & Mishra, 2022).

Hyper-parameter Tuning is performed using Artificial Bee Colony optimization, followed by model evaluation with cross-validation. Hyper-parameter tuning is the process of optimizing the hyper-parameters of a machine learning model to improve its performance. (Bacanin *et al.,* 2023). Some common methods for hyper-parameter tuning includes Grid Search which evaluates all combinations of defined hyper-parameters but is computationally intensive. (Sumathi 2020). Random Search samples a subset of hyper-parameters, offering efficiency with good results at lower computational costs.(Ahamad *et al ., 2023).* Bayesian Optimization models the objective function probabilistic to evaluate promising hyper-parameters, suitable for high-dimensional spaces. (Liao *et al* 2024). Gradient-based Optimization uses gradients of validation errors for differentiable hyper-parameters, enabling efficient tuning. Evolutionary Algorithms apply biological evolution principles like mutation and selection for optimization.(Omotenhinwa & Oyewola 2023). Early Stopping halts training when validation performance plateaus, preventing overfitting. (Awad & Frathat , 2023).

Artificial Bee Colony (ABC) Optimization, developed by Karaboğa in 2005, is a nature-inspired algorithm based on honey bee foraging behavior. It involves three types of bees to balance exploration and exploitation effectively. ABC is simple, flexible, and requires fewer control parameters, making it a robust method for solving optimization problems. (Kayat *et al.,* 2022).

The existing study demonstrates that Random Forests (RF), along with Decision Tree (DT) and K-nearest Neighbors (KNN) algorithms, can provide accurate long-term crime predictions by analyzing past crime data. DT is chosen for its simplicity and interpretability, helping to visualize key features and relationships. RF, an ensemble of decision trees, handles complex data relationships, improves accuracy, and reduces overfitting. KNN is used for classifying multiple crime types. The study emphasizes the importance of understanding non-linear factors in crime prediction and using machine learning models to enhance accuracy, but it does not address the interpretability challenges of these advanced models (Wubineh, 2024). The existing crime prediction system utilizes historical crime data, including features like crime type, location, time, demographics, and weather conditions. The data undergoes preprocessing, including handling missing values, encoding categorical variables, and normalizing numerical features. It is then split into training (80%) and testing (20%) sets. The system employs Random Forest (RF), which uses bootstrap sampling to create multiple subsets of the training data. Each subset trains a different decision tree, and a random subset of features is selected at each split to ensure trees are uncorrelated, reducing overfitting. The trees are fully grown without pruning to capture complex patterns. For classification, each tree votes on a class, and the most frequent class is selected as the final prediction. For regression, the mean of all tree predictions is used. The performance criteria include accuracy, precision, recall, F1 score, ROC-AUC, etc. The architecture is shown in Figure 2 and the framework is presented in Figure 3.

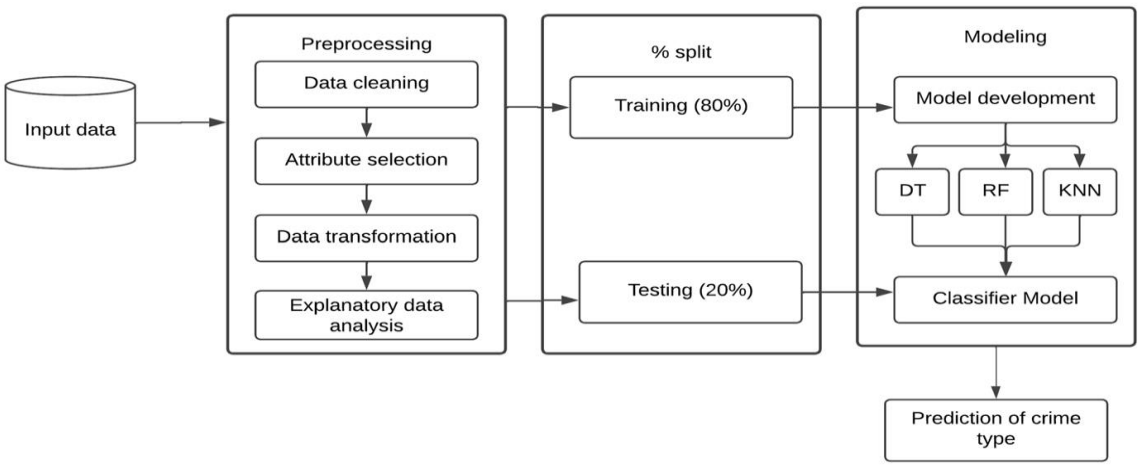

Figure 2. Architecture of the existing system (**Sourc**e:Wubineh, 2024).

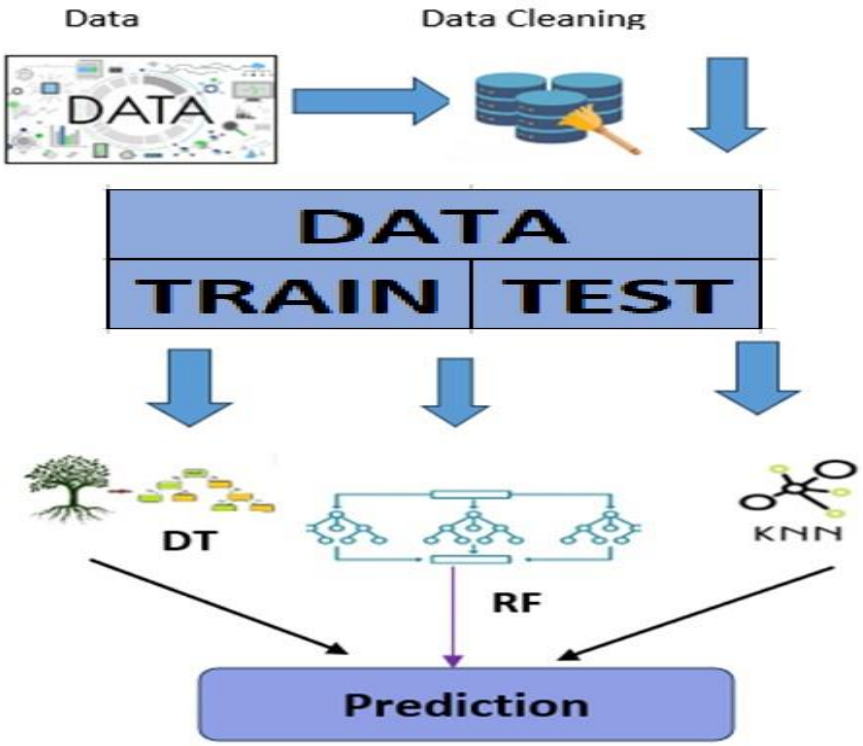

Figure 3. The framework of the existing system (**Sourc**e:Wubineh, 2024).

Although Random Forest (RF) outperforms Decision Trees (DT) and K-nearest Neighbors (KNN), there is still significant potential to improve its accuracy. RF is prone to overfitting, especially if its hyper-parameters are not properly optimized, and when the data contains noise or irrelevant features (Rodrigues *et al*., 2023). It can also struggle with poor generalization to unseen data. Additionally, RF can be memory-intensive due to storing multiple copies of the training data and numerous trees . The algorithm may face challenges with imbalanced datasets, leading to biased predictions towards the majority class, requiring techniques like resampling or weighting. While RF provides feature importance, it may be biased towards features with more levels or higher variability, potentially leading to misleading interpretations if not carefully analyzed.

Overview of the Proposed Model

The proposed model enhances crime prediction by integrating Recursive Feature Elimination with Cross-Validation (RFECV) and Artificial Bee Colony (ABC) optimization to fine-tune the hyper-parameters of a Random Forest model. This hybrid approach addresses the limitations of traditional crime mapping and existing Random Forest models by incorporating advanced optimization, diverse data sources, and a focus on fairness and stakeholder engagement. Key steps include data collection, feature engineering, model training, evaluation, and comparative analysis, with ABC optimization providing a significant advantage in improving model performance and accuracy. Figure 4 shows the overview of the process in the proposed model.

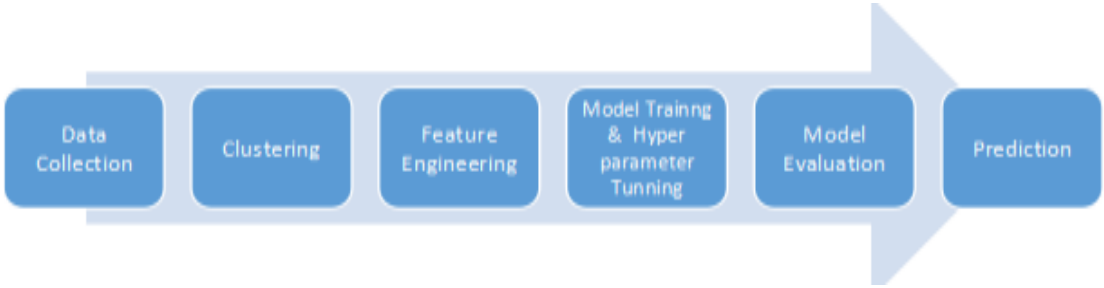

Figure 4: Overview of Proposed Model

Architecture of the proposed model

The proposed model integrates Artificial Bee Colony (ABC) optimization for tuning the hyper-parameters of the Random Forest model. Random Forest, an ensemble learning method, constructs multiple decision trees during training and outputs the mode of the classes (for classification) or the mean prediction (for regression) from the individual trees, improving accuracy and preventing overfitting. The ABC algorithm involves a population of artificial bees, employed bees, onlookers, and scouts, who iteratively search for the optimal set of hyper-parameters, enhancing model performance. Additionally, clustering techniques are used to group similar data points, helping identify and remove outliers, improve data distribution, and create homogeneous subsets of data, which can further enhance the accuracy of the Random Forest model. Figure 5 shows the architecture of the proposed system.

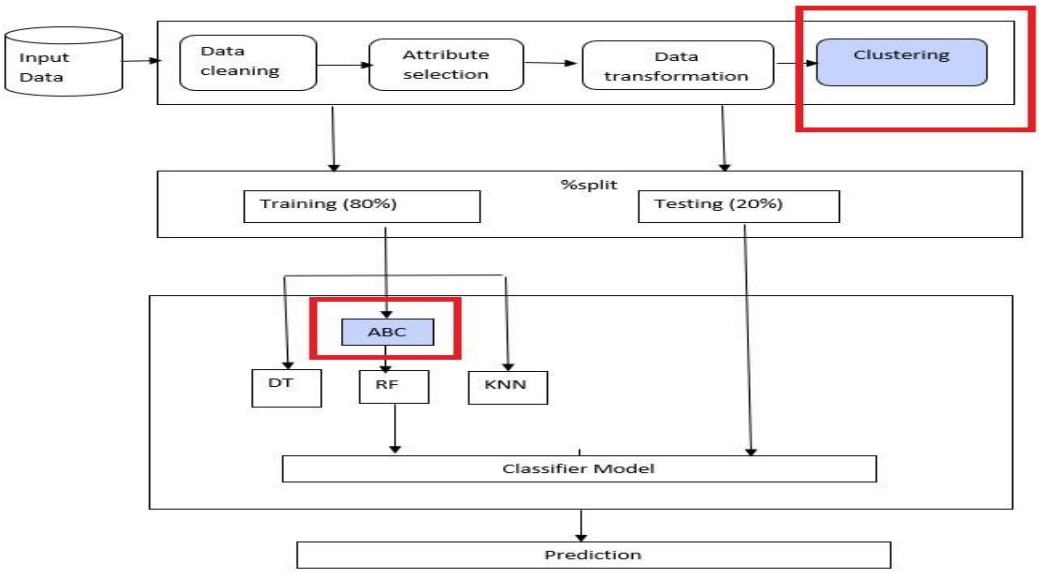

Figure 5: Architecture of the proposed RF- ABC

Working principle of proposed framework

The Ensemble Random Forest framework for crime prediction consists of several key components that work together to produce accurate predictions. It begins with Data Preprocessing, where the data will be cleaned and transformed. Feature Engineering follows, where relevant features will be selected to enhance model performance. The Ensemble Construction step creates multiple base models using techniques like bootstrap sampling and random feature selection. Each base model is a Random Forest, with predictions combined through methods like voting or averaging. The Ensemble Combination component aggregates predictions from all base models, improving accuracy. Finally, the Prediction component will use the trained ensemble model to make predictions on new data, resulting in a robust and accurate crime prediction model.

Figure 6 shows the framework of the proposed model. The flow chart is presented in Figure 7.

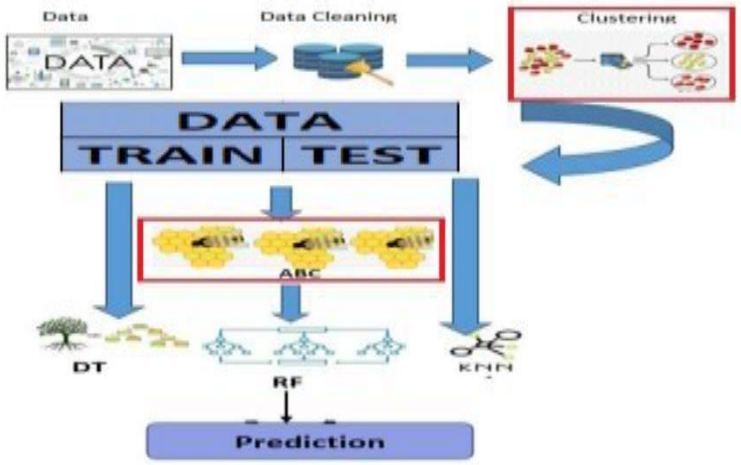

Fig. 7: Framework of the proposed RF – ABC Model

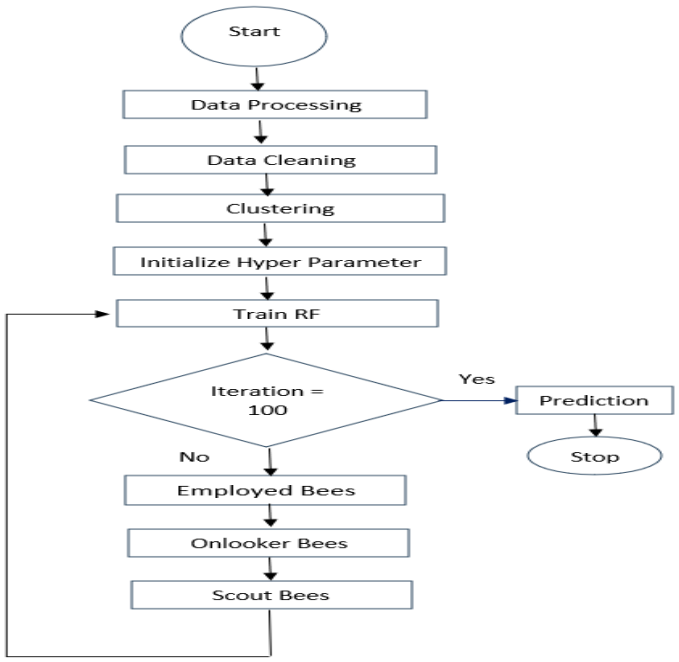

Figure7: Flow Chart of the Proposed Model

The performance of the Random Forest model depends on tuning several hyper-parameters such as the number of trees, the maximum depth of trees, and the number of features for splits. The Artificial Bee Colony (ABC) algorithm will be used to optimize these hyper-parameters. The process involves initialization of a bee colony (representing possible solutions), exploration by employed bees, selective exploitation by onlooker bees, and random searches by scout bees to avoid local optima. This iterative process continues until the best set of hyper-parameters is found. Two models will be developed and tested:

- **Model-1:** ABC tuning the hyper-parameters of a homogeneous Random Forest model.
- **Model-2:** ABC tuning hyper-parameters of a heterogeneous model consisting of Random Forest, KNN, and Decision Trees.

These models will be evaluated to determine the impact of hyper-parameter optimization on crime prediction accuracy.

System Requirement

The proposed ensemble clustering method requires specific hardware and software resources to ensure efficient processing and accurate crime predictions:

Hardware requirement

A multi-core processor (at least 4-6 cores) is recommended to handle the computational demands of the ensemble clustering and optimization steps. A minimum of 8-16 GB of RAM is necessary, with 16 GB or more for larger datasets to ensure smooth processing. Adequate storage of at least 500 GB for larger datasets, with 1-2 GB required for smaller to medium-sized datasets.

Software Requirements:

Python: Primarily used for developing Random Forest models with libraries such as Scikit-learn, which provides an efficient implementation for model training and testing. **MATLAB**: An alternative option for training and testing Random Forest models, using the Statistics and Machine Learning Toolbox.

Performance Evaluation Metrics:

Accuracy, Precision, Recall, F1-score, AUC, and MAP will be used to evaluate the performance of the developed ensemble model, assessing its ability to predict crime rates and types. These metrics will be compared to baseline and existing methods.

Predictive Accuracy Index (PAI):

PAI will be used to evaluate the accuracy of crime hot spot predictions. It measures how well the model predicts spatial crime patterns by comparing the variance of predicted responses at review and development. PAI provides a comprehensive metric for evaluating crime prediction models, accounting for both correctly and incorrectly identified crime hot spots.

The PAI is calculated as follows:

$$PAI = \left(\frac{TP+TN}{TP+TN+FP+FN}\right) \quad (1)$$

Where:

- TP: True Positives (correctly predicted crime hot spots)
- TN: True Negatives (correctly predicted non-crime hot spots)
- FP: False Positives (incorrectly predicted crime hot spots)
- FN: False Negatives (incorrectly predicted non-crime hot spots)

Interpretation of PAI:

- PAI > 0.5: Indicates good prediction accuracy, suggesting the model effectively identified both actual crime hot spots and non-hot-spots.
- PAI = 0.5: Represents chance performance, suggesting the model does not significantly improve upon random guessing.
- PAI < 0.5: Indicates poor prediction accuracy, suggesting the model frequently misidentified crime hot spots.

Predictive efficiency index (PEI)

The **Predictive Efficiency Index (PEI)** evaluates the performance of crime forecasting models by considering both the **accuracy** of crime predictions and the **resources** required to generate them. It is calculated as the ratio of the number of crimes in forecast hot spots to the number of crimes in actual hot spots. This metric helps assess how well a forecasting model performs relative to its resource usage, offering a comprehensive view of its efficiency in predicting crime.

The PEI is calculated as follows:

$$PEI = \left(\frac{TP+TN}{TP+TN+FP+FN}\right) * \left(\frac{T}{C}\right) \quad (2)$$

Where:

- TP is the number of true positives (correctly predicted crimes)
- TN is the number of true negatives (correctly predicted non-crimes)
- FP is the number of false positives (incorrectly predicted crimes)
- FN is the number of false negatives (incorrectly predicted non-crimes)
- T is the total number of predicted crimes
- C is the total number of actual crimes

The **Predictive Efficiency Index (PEI)** ranges from 0 to 1, with higher scores indicating better model performance. It is a useful metric because it evaluates both **accuracy** and **resource efficiency**, making it easy to calculate and compare different models. However, PEI has limitations: it is sensitive to the distribution of crimes and may not be suitable for all crime forecasting models. Additionally, it does not account for the **severity** of crimes.

Recapture rate index (RRI)

The Recapture Rate Index (RRI) is a metric used to assess the precision of crime forecasting models, focusing specifically on their ability to recapture crime hot spots in a future period. RRI is a measure of the percentage of actual crimes that are correctly predicted. The RRI is calculated as follows:

HRR: Hot-spots Recaptured Ratio = Number of predicted crime hot spots that were actual hot spots in the future period / Total number of predicted crime hot spots.

- HD: Historical Density = Average crime density across the entire study area in the historical period.
- HNR: Hot-spots Not Recaptured Ratio = Number of predicted crime hot spots that were not actual hot spots in the future period / Total number of predicted crime hot spots.

$$RRI = \left(\frac{HRR*HD}{HRR+HNR}\right) \quad (3)$$

Where:

Interpretation of RRI:

- RRI > 1: Indicates an increase in crime density in predicted hot spots, suggesting the model accurately identified potential hot spots.
- RRI =1: Indicates that no change in crime density in predicted hot spots, meaning the model did not predict future crime patterns effectively.
- RRI < 1: Indicates a decrease in crime density in predicted hot spots, suggesting the model overestimated future crime activity.

SHapley additive exPlanations (SHAP)

SHAP values are a method to interpret the output of machine learning models. They provide insight into how each feature in the datasets influences the model's predictions. SHAP values are based on the concept of Shapley values from cooperative game theory, ensuring a fair distribution of the total output among the features (Song *et al.,* 2023). They ensure that each feature's contribution is fairly assessed by considering all possible combinations of features. Positive SHAP values indicate that a feature increases the prediction, while negative values suggest a decrease. The magnitude of SHAP values reflects the strength of the feature's impact. SHAP values provide a powerful and flexible way to interpret machine learning models, ensuring trust in model predictions (Nohara *et al.,* 2022).

SHAP Value is illustrated as:

$$g(z) = \emptyset_0 + \sum_{j=1}^{p} \emptyset_0 z_j^{"} \quad (4)$$

Where g is the explanation method, p is the number of features, and $z' \in \{0,1\}$ is the coalition vector that indicates the on or off state of each feature.

RESULT AND DISCUSSION

At the end of this research it is expected that, the research proposes the development of an ensemble clustering crime prediction model utilizing the Random Forest algorithm to enhance prediction accuracy. Hyper-parameters of the model will be optimized using Artificial Bee Colony (ABC) techniques, while interpretability will be improved through the Recursive Feature Elimination with Cross-Validation (RFECV) method. The ensemble model will be integrated into a framework, with performance evaluation and bench-marked against existing models using metrics such as Predictive Accuracy Index (PAI), Predictive Efficiency Index (PEI), Recapture Rate Index (RRI), and SHapley Additive exPlanations (SHAP) values.

CONCLUSION

The work concludes by emphasizing the significance of developing an enhanced crime prediction model that will improve accuracy, precision, and interpretability through the use of a hyper-parameter tuned Random Forest algorithm optimized by the Artificial Bee Colony (ABC) technique. we outline the limitations of traditional crime prediction methods, which often oversimplify complex crime patterns and fail to capture non-linear relationships and diverse data sources. By advocating for the use of advanced machine learning models, the paper underscores the importance of addressing inherent biases and ensuring ethical use in predictive modeling. The conclusion stresses that improved interpretability and transparency in crime prediction models are essential for gaining public trust and ensuring that law enforcement agencies can effectively utilize the predictions for resource allocation and proactive crime prevention strategies . In future, we plan to implement the proposed RF-ABC model through advanced tuning using Artificial Bee Colony (ABC) optimization techniques. The focus will also include a detailed comparison of the developed models' performance against existing crime prediction approaches. The ultimate goal is to refine the ensemble crime prediction methodology and contribute to more effective, transparent, and actionable crime prevention strategies.

ACKNOWLEDGMENT

We will like to express our profound gratitude to Tertiary Education Trust Fund (TETFund). We appreciate the collective efforts and support received from our department of Computer Science, Faculty of Computing Abubakar Tafawa Balewa University, Bauchi . Bauchi State , Nigeria.

REFERENCES

Adeyemi, R. A., Mayaki, J., Zewotir, T. T., & Ramroop, S. (2021). Demography and Crime: A Spatial analysis of geographical patterns and risk factors of Crimes in Nigeria. *Spatial Statistics*, *41*, 100485.

Ahamad, G. N., Shafiullah, Fatima, H., Imdadullah, Zakariya, S. M., Abbas, M., ... & Usman, M. (2023). Influence of optimal hyperparameters on the performance of machine learning algorithms for predicting heart disease. *Processes*, *11*(3), 734.

Alsayadi, H. A., Khodadadi, N., & Kumar, S. (2022). Improving the regression of communities and crime using ensemble of machine learning models. *J. Artif. Intell. Metaheuristics,* 1(1), 27-34.

Anyanwu, G. O., Nwakanma, C. I., Lee, J. M., & Kim, D. S. (2023). Novel hyper-tuned ensemble random forest algorithm for the detection of false basic safety messages in internet of vehicles. *ICT Express*, *9*(1), 122-129.

Awad, M., & Fraihat, S. (2023). Recursive feature elimination with cross-validation with decision tree: Feature selection method for machine learning-based intrusion detection systems. *Journal of Sensor and Actuator Networks*, *12*(5), 67.*l of Modern Science*, *8*(1), 1-19.

Bacanin, N., Stoean, C., Zivkovic, M., Rakic, M., Strulak-Wójcikiewicz, R., &Stoean, R.(2023). On the benefits of using metaheuristics in the hyperparameter tuning of deeplearning mAwad, M., &Fraihat, S. (2023). Recursive feature elimination with cross validation with decision tree: Feature selection method for machine learning-based intrusion detection systems. *Journal of Sensor and Actuator Networks*, *12*(5), 67.

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., ... & Lindauer, M. (2023).Hyper parameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *13*(2), e1484.

Kaya, E., Gorkemli, B., Akay, B., &Karaboga, D. (2022). A review on the studies employing artificial bee colony algorithm to solve combinatorial optimization problems. *Engineering Applications of Artificial Intelligence*, *115*, 105311.

Kedia, P. (2016). Crime mapping and analysis using GIS. *International Institute of Information Technology*, *1*(1), 1-15.

Khan, M., Ali, A., & Alharbi, Y. (2022). Predicting and preventing crime: A crime prediction model using San Francisco crime data by classification techniques. *Complexity*, *2022.*

Liao, M., Wen, H., Yang, L., Wang, G., Xiang, X., & Liang, X. (2024). Improving the model robustness of flood hazard mapping based on hyperparameter optimization of random forest. *Expert Systems with Applications*, *241*, 122682.

Nohara, Y., Matsumoto, K., Soejima, H., & Nakashima, N. (2022). Explanation of machine learning models using shapley additive explanation and application for real data in hospital. Computer *Methods and Programs in Biomedicine*, 214, 106584.

Oh, G., Song, J., Park, H., & Na, C. (2022). Evaluation of random forest in crime prediction: Comparing three-layered random forest and logistic regression. *Deviant Behavior*, *43*(9), 1036-1049.

Omotehinwa, T. O., & Oyewola, D. O. (2023). Hyper-parameter optimization of ensemble models for spam email detection. *Applied Sciences*, *13*(3), 1971..

Passos, D., & Mishra, P. (2022). A tutorial on automatic hyper-parameter tuning of deep spectral modelling for regression and classification tasks. *Chemometrics and Intelligent Laboratory Systems*, *223*, 104520.

Pfob, A., Lu, S. C., & Sidey-Gibbons, C. (2022). Machine learning in medicine: a practical introduction to techniques for data pre-processing, hyper-parameter tuning, and model comparison. *BMC medical research methodology*, *22*(1), 282.

Rodrigues, A., González, J. A., & Mateu, J. (2023). A conditional machine learning classification approach for spatio-temporal risk assessment of crime data. *Stochastic Environmental Research and Risk Assessment*, 1-14.

Song, Y., Wang, Q., Xi, Y., Ma, W., Zhang, X., Dong, L., & Wu, Y. (2023). Interpretability study on prediction models for alloy pitting based on ensemble learning. *Corrosion Science*, 111790.

Sumathi, B. (2020). Grid search tuning of hyper-parameters in random forest classifier for customer feedback sentiment prediction. *International Journal of Advanced Computer Science and Applications*, *11*(9).

Wang, D., Ding, W., Lo, H., Stepinski, T., Salazar, J., & Morabito, M. (2023). Crime hotspot mapping using the crime related factors a spatial data mining approach. *Applied intelligence*, *39*, 772-781.

Wheeler, A. P., &Steenbeek, W. (2021). Mapping the risk terrain for crime using machine learning. *Journal of Quantitative Criminology*, *37*, 445-480.

Wubineh, B. Z. (2024). Crime analysis and prediction using machine-learning approach in the case of Hossana Police Commission. *Security Journal*, 1-16.

Yao, S., Wei, M., Yan, L., Wang, C., Dong, X., Liu, F., & Xiong, Y. (2020, August). Prediction of crime hotspots based on spatial factors of random forest. In 2020 15th International *Conference on Computer Science & Education* (ICCSE) (pp. 811-815). IEEE.

Zhang, H., Gao, Y., Yao, D., & Zhang, J. (2023). Interaction of Crime Risk across Crime Types in Hotspot Areas. ISPRS *International Journal of Geo-Information*, 12(4), 176.