A Time-Aware Ensemble Learning Framework for Crime Rate Prediction and Risk Classification

1st Harinandan Singh
School of Engineering and Information Technology
Sanskriti University, Chhata
Mathura, Uttar Pradesh, India
harry.truba@gmail.com

2nd Satya Prakash Yadav
School of Engineering and Information Technology
Sanskriti University, Chhata
Mathura, Uttar Pradesh, India
spyadav2290@gmail.com

3rd Pankaj Kumar
Department of Technical Education
GPC, Patiala
Punjab, India
singla.pankaj4@gmail.com

4th Ramanjeet Kaur
Department of Applied Sciences, School of Engineering & Technology
CGC University
Mohali-140307, Punjab, India
rpanjeta@gmail.com

5th Danish Meiraj
School of Computer Science and Engineering
Lovely Professional University
Jalandhar, Punjab, India
er.danishmeiraj@gmail.com

***Abstract*—This study proposes to build a machine learning model to predict the crime rates of the Indian states and union territories of the Indian Penal Code (IPC) from the National Crime Records Bureau (NCRB) data from 2016-2023. The panel data of crime rates were constructed with different crime rate indicators for the states and years and the problem in the data (skewness and non-stationarity) are resolved by log transformation and train-test split by time. This is tested and compared with models like Linear Regression (LR), Support Vector Regression (SVR), Random Forest (RF) and Gradient Boosting (GB) and used classification models to classify the risk level. The result demonstrates various levels of accuracy and stability of the models. Random Forest is the best performing model (R^2 = 0.947) which has good balance between non-linearity and overfitting in the crime data. Our experiment shows the ensemble models are better; the top two models are Gradient Boosting (R^2 = 0.984, RMSE = 30.89, MAE = 18.27) and then Random Forest (R^2 = 0.947) and SVR (R^2 = 0.932), and linear regression is not good due to the linearity assumption. Logistic regression is the best model for the classification task with good accuracy (88.6%) and precision and recall for the high-risk group. The best two features are the murder rate and violent crime rate (feature importance and SHAP analysis). Finally, the model is able to predict and can be used for evidence-based crime prevention.**

Keywords— Crime Prediction, IPC Crime Rate, Ensemble Learning, Time-Aware Modelling, Model Interpretability, NCRB

I. INTRODUCTION

Crime forecasting plays an important role in evidence-based governance, especially in decentralized systems like India, where public safety is mainly handled at state level. Traditional statistical models often struggle with nonlinear relationships and also structural differences in data, whereas machine learning methods are more flexible and can handle complex social data in a better way. Previous research also supports data driven crime prediction. Bogomolov et al. [1] showed that contextual information can improve prediction accuracy. Kounadi et al . [2] found that machine learning performs better in spatial forecasting. Comparative studies by Tollenaar and van der Heijden [3] and Safat et al. [4] highlighted that ensemble and boosting models are more effective. Berk et al. [5] talked about importance of statistical rigor in high stakes applications while Mohler et al. [6] pointed out that temporal structure is also very important in crime modelling. However most of the existing studies are mainly focuses city level or hotspot based forecasting . State level longitudinal modeling in India is still limited . This kind of modeling is important for policy making and also for better resource allocation .

This study try to develop a machine learning framework to predict IPC crime rates across Indian states and union territories for the period 2016–2023 using official NCRB data. A unified state year dataset is constructed using multiple crime indicators. Since the data showed right skewness and some fluctuations during 2020–2021 IPC Crime Rate is log transformed before applying regression. Also , a time aware train test split (2016–2021 training and 2022–2023 testing) is used to maintain chronological order. The main objectives of this study are (i) to predict IPC crime rates using regression models (ii) to classify states into Low Medium and High-risk categories using tertile-based segmentation and (iii) to identify important predictors using interpretability techniques. The framework includes linear regression and support vector regression as baseline models along with hyperparameter tuned Random Forest and Gradient Boosting models. This study contributes in following ways (1) constructing a multi-year state level dataset using NCRB statistics (2) implementing time aware modelling with temporal cross validation (3) comparing regression and classification models and (4) using impurity-based permutation and SHAP based interpretability methods to improve policy relevance.

II. RELATED WORK

Machine learning in crime analytics has changed a lot over time, earlier it was mostly simple data mining but now it has moved towards more advanced prediction frameworks. Early studies showed that nonlinear models actually work better than traditional statistical ones. Saltos and Cocea [7] demonstrated that decision tree techniques can be more effective than linear regression in predicting the frequency of crime, whereas Sathyadevan et al.[8] demonstrated that both supervised and unsupervised learning can be useful in crime classification and pattern discovery .

Subsequent studies began to pay more attention to the comparison of models and ensemble methods. Safat et al. [4] have compared different models including logistic regression, SVM, random forest, boosting and deep learning and found

that the ensemble ones are more likely to yield better results. Kim et al. [9] and Kumar et al. [10] also report similar types of results suggesting that they had better classification and hotspot detection accuracy when they used supervised learning methods. Most of these works are however limited to hot spot and city level data and will need more short-term prediction and city data thus not general.

One of the directions in this direction is also spatio-temporal modelling. Mohler et al. [6] proposed self-exciting point process models to model temporal clustering, and demonstrated that there is an impact of the past on the near future probability of crime. Micro location forecasting models built by Jayawardana and Pathmaperuma [14] are based on the Random Forest algorithm with some engineered time characteristics, yet they are typically applied in small urban areas, rather than in the large scale. The modelling of the state level, using various populations, is not done in an appropriate way.

Ensemble learning has been a sort of a main focus of modeling complex social data. Random Forest [11], Gradient Boosting [12] and XGBoost [13] are able to capture nonlinear relations and feature variations. They can also deal with multicollinearity and outliers.These models are just a combination of a large number of weak learners to produce a more robust prediction system and thus they are effective in socio economic data where trends are neither very simple nor very clear.

Berk et al. [5] pointed out need of transparency and proper statistical rigor in such predictive models. Model agnostic methods like SHAP help to break predictions into feature level contributions, so it becomes easier to understand and also adds some accountability. Recent research is trying to combine explainability with prediction models so that results are not only accurate but also understandable, which is what is actually needed.

Even though there is already some good amount of research, still some gaps are there. Most of the work is focused on local urban prediction and not much on state level modeling. Multi-year datasets across large admin regions are not explored much, and interpretability is sometimes ignored when compared to accuracy , also very limited studies use time aware validation methods. This study tries to address these gaps by creating a state-wise dataset for India from 2016–2023 and applying time-aware ensemble models. Random Forest and Gradient Boosting are used with proper tuning, along with some baseline models. Interpretability is also included using feature importance and SHAP. Overall , the work focuses on state-level modelling and aims to give clearer insights with the predictions . It improves crime analytics, though there is still some scope for improvement.

III. Methodology

A. Data Sources

A state-year panel dataset for the period 2016–2023 was prepared using official crime statistics. The data was mainly taken from the National Crime Records Bureau (NCRB) Crime in India reports (2016–2023) [15]. We included different variables such as IPC Crime Rate, total IPC crimes, murder incidents and their rate, violent crimes and rate, culpable homicide, death by negligence, and population (in lakhs). The dataset was organized year-wise for each state.

B. Dataset Construction and Alignment

State names were standardized before merging the datasets to ensure consistency. Some administrative changes were also considered such as the 2019 reorganization of Jammu & Kashmir and the 2020 merger of Dadra & Nagar Haveli with Daman & Diu. The final dataset consists of observations and variables that are relevant. The panel is arranged in a chronological manner within each state (State × Year) , which helps in applying time-aware modeling. The

C. Structural Variability

Some noticeable discontinuities were observed during 2020–2021. This was probably due to pandemic-related disruptions and changes in reporting patterns , which affected the data consistency to some extent. These variations were kept to preserve the real behaviour of the dataset. Ensemble models were chosen for this study. They are generally more robust and can handle nonlinear variability and structural breaks better compared to simpler models.

D. Feature Engineering

The following features were constructed:

1) Crime Growth Rate: It is defined in (1). Let C_t denote the IPC crime rate in year t.

$$\text{Crime Growth Rate}_t = \frac{C_t - C_{t-1}}{C_{t-1}} \times 100 \quad (1)$$

2) Violent Crime Share: It is defined in (2).

$$\text{Violent crime share} = \frac{\text{Violent crimes}}{\text{Total IPC crimes}} \quad (2)$$

3) Murder Share: It is defined in (3).

$$\text{Murder share} = \frac{\text{Murder incidents}}{\text{Total IPC crimes}} \quad (3)$$

Raw counts were used only for constructing ratios. They were not included directly in the regression inputs, mainly to avoid redundancy and also issues like multicollinearity with rate-based predictors.

E. Outlier Treatment and Target Transformation

Based on the exploratory analysis, it was observed that the data were skewed to the right and had extreme values in IPC Crime Rate, particularly in 2020-21.This could affect the model performance. To handle this, a logarithmic transformation was applied so that variance becomes more stable and the impact of extreme spikes is reduced, as shown in (4).

$$IPC\ Log = \log(IPC\_Crime\ Rate) \quad (4)$$

Regression models were trained on the transformed target. Predicted log-values were exponentiated to evaluate performance on the original IPC Crime Rate scale.

F. Final Model Inputs

Regression and classification models used the following set of predictors:

- Population (Lakhs)
- Murder Rate
- Violent Crime Rate
- Crime Growth Rate
- Violent Crime Share
- Murder Share

For regression, the target variable was the log-transformed IPC Crime Rate. In classification, states were divided into risk

categories (Low, Medium, High) using tertile-based segmentation with quantiles (pd.qcut, q = 3). To avoid target leakages, only those predictors were used which do not directly reconstruct IPC Crime Rate. This is important because IPC Crime Rate itself is calculated separately by NCRB as total IPC crimes per lakh population, so it should not be indirectly derived from input variables. The block diagram of proposed state-wise crime prediction framework is shown in Fig. 1.

G. Model Development

1)) Regression Models: IPC Crime Rate was predicted using:

- Linear Regression
- Support Vector Regression (SVR)
- Random Forest Regressor
- Gradient Boosting Regressor

Hyperparameter tuning for Random Forest and Gradient Boosting was done using GridSearchCV with Timeseries Split cross-validation (3 splits) on the training data to keep the time order of the data intact. Tree-based ensemble methods [11], [12] were chosen for this study , as they can handle nonlinear relationships quite well.

2) Classification Models: States were categorized into high, medium, and low risk groups using tertile-based quantile segmentation on IPC Crime Rate to keep the class distribution balanced. Classification models included:

- Logistic Regression
- Random Forest Classifier
- Support Vector Machine (SVM)

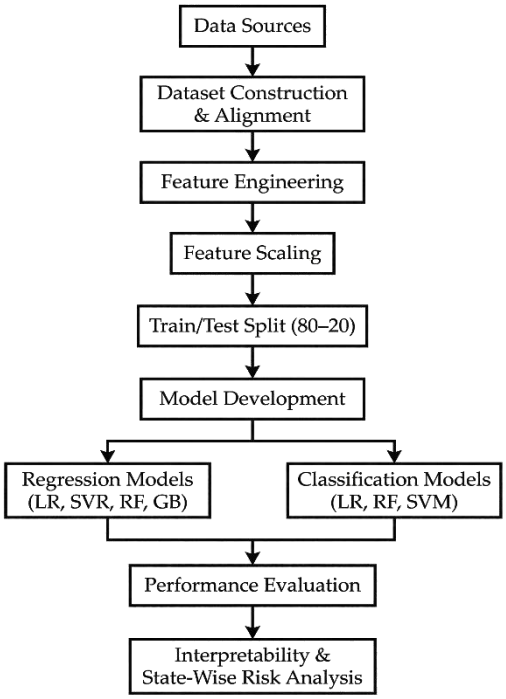

Fig. 1 Block Diagram of Proposed State-Wise Crime Prediction Framework

H. Model Training and Evaluation

Continuous predictors were standardized using Standard-Scaler for Linear Regression, SVR, Logistic Regression and SVM. Tree-based ensemble models (Random Forest and Gradient Boosting) were trained on unscaled features , since they are not affected much by feature scaling. A time-aware train-test split was used in this study. Data from 2016–2021 was taken for training . 2022–2023 was used for out-of-sample testing. This helps in maintaining the correct time order and also avoids leakage of future information. For hyperparameter tuning, timeseries split cross-validation was applied within the training set.

All models were evaluated strictly on unseen test data. For regression, predictions generated on the log-transformed scale were exponentiated to compute performance metrics on the original IPC Crime Rate scale, ensuring interpretability in practical units. Model interpretability was performed using impurity-based feature importance (Random Forest and Gradient Boosting), permutation importance (SVR), and SHAP (Tree Explainer) for global and local explanations. Interpretability analysis followed final model selection to ensure consistency with optimized configurations. Additionally, model selection was based on comparative out-of-sample performance metrics to ensure robustness and prevent overfitting.

1) Regression Metrics: The various regression metrics are defined in (5), (6), (7).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (5)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (6)$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (7)$$

2) Classification Metrics: Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. The various classification metrics are defined in (8), (9), (10), (11).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

$$F1 = 2x\frac{Precision \cdot Recall}{Precision+Recall} \quad (11)$$

IV. Experimental Results

A. Regression Performance

Regression models were evaluated on the out-of-sample test period (2022–2023). Performance metrics were computed after exponentiating the log-transformed predictions to restore IPC Crime Rate to its original scale.

TABLE I REGRESSION PERFORMANCE ON TEST SET (2022–2023)

Model	RMSE	MAE	R^2
Linear Regression	27181.73	3298.16	-12439.05
SVR	63.66	30.63	0.932
Random Forest	56.19	31.18	0.947
Gradient Boosting	30.89	18.27	0.984

As shown in Table I, Gradient Boosting achieved the best performance (R^2 = 0.984), which clearly outperforming the other models. Random Forest also showed strong results (R^2 = 0.947) and SVR had moderate performance (R^2 = 0.932) .

Linear Regression performed very poorly (R^2 = -12439.05) which indicates serious model misspecification. This is mainly because it cannot capture nonlinear relationships in longitudinal crime data. The extremely negative R^2 value also shows that the model is highly sensitive to distribution changes and retransformation errors under log scaling. The better performance of Gradient Boosting can be explained by its sequential residual correction process. It is able to model nonlinear interactions between crime-related variables and demographic factors more effectively.

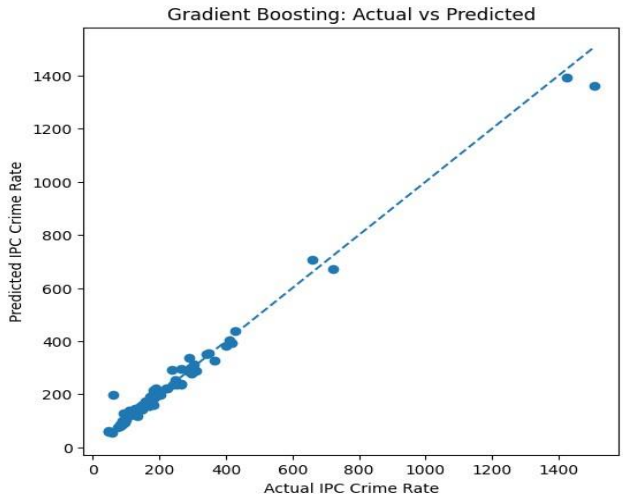

Fig. 2 Actual vs Predicted IPC Crime Rate (Gradient Boosting)

As seen in Fig. 2, the predicted values are closely aligned with the 45-degree reference line. This suggests good calibration with very little systematic bias. The spread of values is also limited even at higher crime rates which indicates strong out-of-sample performance. Overall, ensemble methods perform much better than linear and kernel-based baseline models for state-level crime rate prediction , showing that capturing nonlinear patterns is important for improving prediction accuracy in such datasets.

B. *Classification Performance*

States were categorized into Low, Medium, and High-risk groups using tertile-based quantile segmentation.

TABLE II CLASSIFICATION MODEL COMPARISON ON TEST SET

Model	**Accuracy**	**Weighted (F1)**
Logistic Regression	0.886	0.886
Random Forest	0.871	0.874
SVM	0.843	0.848

As shown in Table II , Logistic Regression achieved the highest overall accuracy (88.6 %) and weighted F1-score (0.886), performing better than Random Forest and SVM. Hence, Logistic Regression was selected as the final classification model.

TABLE III LOGISTIC REGRESSION CLASSIFICATION PERFORMANCE

Class	**Precision**	**Recall**	**F1-score**	**Support**
High	0.93	0.96	0.94	26
Low	0.95	0.80	0.87	25
Medium	0.77	0.89	0.83	19
Macro Avg	0.88	0.89	0.88	70
Weighted Avg	0.89	0.89	0.89	70
Accuracy		0.89		

As shown in Table III, logistic regression performs well in identifying high-risk states (Precision = 0.93 , Recall = 0.96). Most of the misclassifications happen between Low and Medium categories, which suggests some overlap in risk levels rather than a clear model bias.

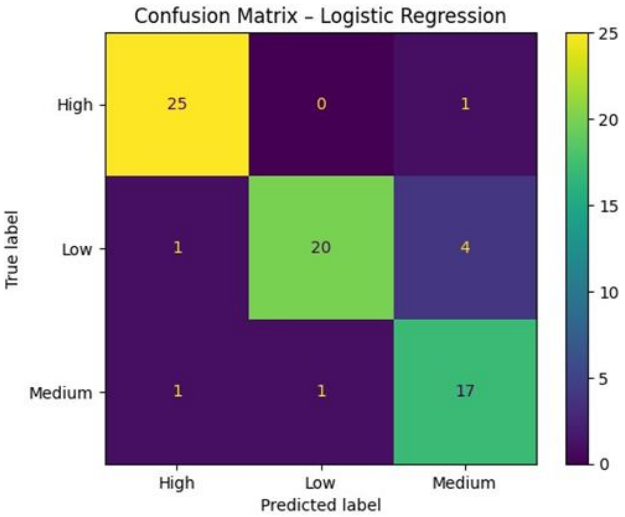

Fig. 3 Confusion Matrix – Logistic Regression

Fig. 3 shows limited cross-class leakage with especially clear separation for High-risk states.

V. MODEL INTERPRETABILITY AND STATE-WISE RISK ANALYSIS

A. *Feature Importance Analysis*

Interpretability was done using feature importance and SHAP which help in understanding how the model works and which features are important.

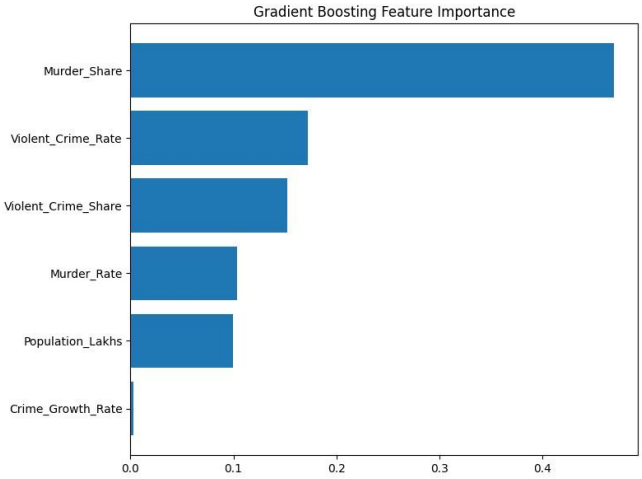

Fig. 4 Gradient Boosting Feature Importance

From Fig. 4, the main predictors are:

- Murder Share

- Violent Crime Rate
- Violent Crime Share
- Murder Rate
- Population Lakhs
- Crime Growth Rate

It can be seen that variables of crime composition would have a stronger role than just the demographic factors in determining the intensity of IPC crime.

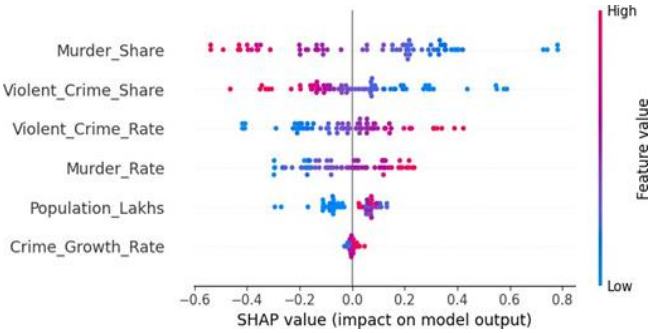

Fig. 5 SHAP Summary Plot with Global Feature Contributions

Fig. 5 demonstrates that SHAP analysis shows that increased value of Murder Share and Violent Crime Rate shifts the predictions upwards whereas Crime Growth Rate has a comparatively lower influence. These findings are in line with the feature importance results with similar trends across methods. The consistency of the contribution to features among observations is also an indication that the model is reliable in its behaviour. This also justifies the need to concentrate on variables that have high impact in order to plan and make better decisions on policy.

VI. Conclusion and Future Scope

This study examines the potential of a machine learning framework to predict the rates of IPC crimes in Indian states from time series data from the National Crime Records Bureau (NCRB) 2016-2023. The framework addresses the issues of skewness, variations and seasonality in data, using temporal modelling, feature extraction and log transformation. As the results show, ensemble methods, like Gradient Boosting, outperform traditional linear models and kernel-based models with high accuracy and generalisation ability in the out-of-sample prediction test. In addition, the classification model using Logistic Regression accurately predicts the risk levels for each state, thereby helping policymakers to identify the riskiest states. Explainability techniques, such as feature importance and SHAP, also explain the model to policymakers by pointing out important features (murder share and violent crime rate), thus improving its utility.

The above-mentioned works are crucial but there is still much more that can be done. More socio-economic, demographic and environmental characteristics such as unemployment rate, education level, urbanisation and migration patterns can enhance model prediction and interpretation. Including higher-frequency data can improve the policy-making process. Also, the use of deep learning, spatio-temporal and hybrid ensemble models could be explored for complex modeling. Further, a local or urban level analysis will provide more detail. Finally, explainable and interpretable AI techniques will make models more interpretable and trustworthy and allow use in high-stakes public safety.

References

[1] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi and A. Pentland, “Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data,” in Proc. 16th ACM Int. Conf. Multimodal Interaction (ICMI), Istanbul, Turkey, 2014, pp. 427–434.

[2] O. Kounadi, A. Ristea, A. Araujo Jr. and M. Leitner, “A systematic review on spatial crime forecasting,” Crime Science, vol. 9, Art. no. 7, 2020.

[3] N. Tollenaar and P. G. M. van der Heijden, “Which Method Predicts Recidivism Best A Comparison of Statistical, Machine Learning and Data Mining Predictive Models,” J. Royal Statistical Society: Series A, vol. 176, no. 2, pp. 565–584, 2013.

[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] R. Berk, L. Sherman, G. Barnes, E. Kurtz and L. Ahlman, “Forecasting Murder Within a Population of Probationers and Parolees,” J. Royal Statistical Society: Series A, vol. 172, no. 1, pp. 191–211, 2009.

[6] G. Mohler et al., “Self-Exciting Point Process Modeling of Crime,” J. American Statistical Association, vol. 106, no. 493, pp. 100–108, 2011.

[7] G. Saltos and M. Cocea, “An Exploration of Crime Prediction Using Data Mining on Open Data,” Int. J. Information Technology & Decision Making, vol. 16, no. 5, pp. 1155–1181, 2017.

[8] S. Sathyadevan, M. S. Devan and S. S. Gangadharan, “Crime analysis and prediction using data mining,” in Proc. Int. Conf. Networks & Soft Computing (ICNSC), Guntur, India, 2014, pp. 406–412.

[9] S. Kim et al., “Crime Analysis Through Machine Learning,” in Proc. IEEE IEMCON, Vancouver, Canada, 2018, pp. 415–420.

[10] R. S. Kumar et al., “Empirical Analysis on Crime Prediction using Machine Learning,” in Proc. ICCCI, Coimbatore, India, 2023, pp. 1–5.

[11] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[12] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[13] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proc. ACM SIGKDD, San Francisco, CA, USA, 2016, pp. 785–794.

[14] H. Jayawardana and M. H. Pathmaperuma, “Suraksha: Spatio-Temporal Crime Forecasting and Micro-Location Analysis,” Journal of Electrical Systems, vol. 20, no. 9s, pp. 1635–1641, 2024.

[15] National Crime Records Bureau, “Crime in India Reports 2016–2023,” Government of India. [Online]. Available: https://ncrb.gov.in