Vantage: Journal of Thematic Analysis
(Vantage J. Them. Anal.) ISSN(E): 2582-7391 Vol. 7(1), April, 2026, 01-13

Original Research Article

Enhanced Approach for Crime Predictive Analysis Using NCRB Dataset

Sanjana Dwivedi,
Department of Computer Science, Shaheed Rajguru College of Applied Sciences for Women, University of Delhi, India.

Corresponding Author: sanjanadwivedi574@gmail.com

Received: 13/03/2026 | Accepted: 27/04/2026 | Published: 05/05/2026

ABSTRACT

The dynamic nature of socio-economic and spatial factors that affect crime patterns makes crime prediction and resolution complex. Even though the National Crime Records Bureau (NCRB) offers a vast amount of crime data in India, it is not utilized in predictive modelling, but in descriptive analysis. Integrated frameworks that use NCRB data to make long-term predictions and spatial analysis are lacking. This study solves this issue by concentrating on predictive modelling with NCRB data.

Keywords: Crime Prediction, Ensemble Learning, NCRB Dataset, Machine Learning, Time-Series Forecasting, Hotspot Analysis

1. INTRODUCTION

The prediction and resolution of crime is complicated. It is influenced by various factors including income, community composition, and location. Patterns in crime are geographical and not permanent. This renders the crime analysis important to the law enforcement agencies, the policymakers, and the planners in the field of public safety. These variations are important in a varied nation such as India. This demonstrates that we require predictive systems which can operate at both regional and national levels.

The National Crime Records Bureau (NCRB) of the Ministry of Home Affairs (National Crime Records Bureau, 2021) releases an annual research report named Crime in India. Such reports are very helpful as they give statistics of the criminal activities within states and union territories. The data in NCRB is one of the most valuable information resources to study the historical crime trends. But it is primarily used to do descriptive reporting and examine what happened in the past. It cannot be effectively used in crime prevention and long-term planning because it does not predict the future.

The latest developments on machine learning are promising. They may assist in the development of models of crime that comprehend patterns of time and space. International studies indicate that machine learning has the ability to forecast the trends of crime and establish risky regions. Nevertheless, there is a lack of studies that are specific to India. Specifically, methods of ensemble learning, long-term forecasting and structures that integrate state and national analysis are underdeveloped.

Several of the current crime prediction research is short-term predictions or specific models. They usually lose spatial associations.

Crime in a given area may spill over to other areas. In addition, the absence of interactive visualization tools increases the difficulty of allowing non-technical stakeholders, such as law enforcement and policymakers to retrieve prediction outcomes. This study fills these gaps by proposing a tailor-made crime forecasting model with the help of NCRB data. The framework employs ensemble learning to increase the predictive accuracy. It also uses feature engineering of higher order to obtain time trends, and spatial connections. We make projections over a span of years in order to forecast future crime trends. Statistical hot spot is also used in identifying high-risk areas. A dashboard that is interactive displays past and future data in an understandable format.

This research paper demonstrates how group learning aids in prediction of crimes on large-scale. It offers a specific approach based on data of addressing crime in India prior to its occurrence.

2. RELATED WORK

A review of existing literature indicates that traditional machine learning models such as Decision Trees, Support Vector Machines, and Random Forest have been widely used for crime prediction (Hossain et al., 2020; Abubakera et al., 2025). Recent studies highlight the effectiveness of ensemble and deep learning techniques in improving prediction accuracy, particularly in complex and imbalanced datasets (Mandalapu et al., 2023; Safat et al., 2021; Khan et al., 2022). However, limited research has focused on using NCRB datasets for long-term forecasting and spatial analysis, indicating a gap in the Indian context. Several additional studies and datasets (Yang, n.d.; Nicho et al., 2025; Kabiraj, 2023; National Crime Records Bureau, 2005) further support the development of crime prediction frameworks.

2.1 Characteristics of Traditional Machine Learning Approaches

Early studies that were done on the prediction of crime, mostly relied on conventional machine learning algorithms such as Decision Trees, Naive Bayes, K-Nearest Neighbours, Support Vector Machines, and Logistic Regression (Zhang et al., 2020).These approaches were based on the classification and regression tasks on historical crime records and basic spatial or temporal features. Several primary studies used these methods to identify crime patterns and hotspots with reasonable accuracy and low computational complexity (Hajela et al., 2020; Hossain et al., 2020; Khan et al., 2022; Abubakera et al., 2025).

While such models are easy to interpret and implement, most of the studies found limitations in representing long-term temporal dependences and complex spatial interrelations. Additionally, there have been limitations in the traditional methods of scaling out and generalizing them by building them based on short term prediction horizons and city level data sets.

2.2 Ensemble Learning Based Crime Prediction

In order to improve the performance of the prediction accuracy, several studies have been conducted to explore ensemble learning techniques, including Random Forest, AdaBoost, Gradient Boosting, and XGBoost (He & Zheng, 2021). These methods merge several weak learners to include a combination where the learning efficiency is much better and accurate. Ensemble-based models invariably outperformed single classifiers, especially while dealing with noisy or imbalanced datasets that contained DCR (Mandalapu et al., 2023; Safat et al., 2021; Khan et al., 2022).

Stacking based ensemble frameworks were also proposed in which one combines output of the different base models using a meta-learner. Such approaching results showed better accuracy of regression and smaller variance.

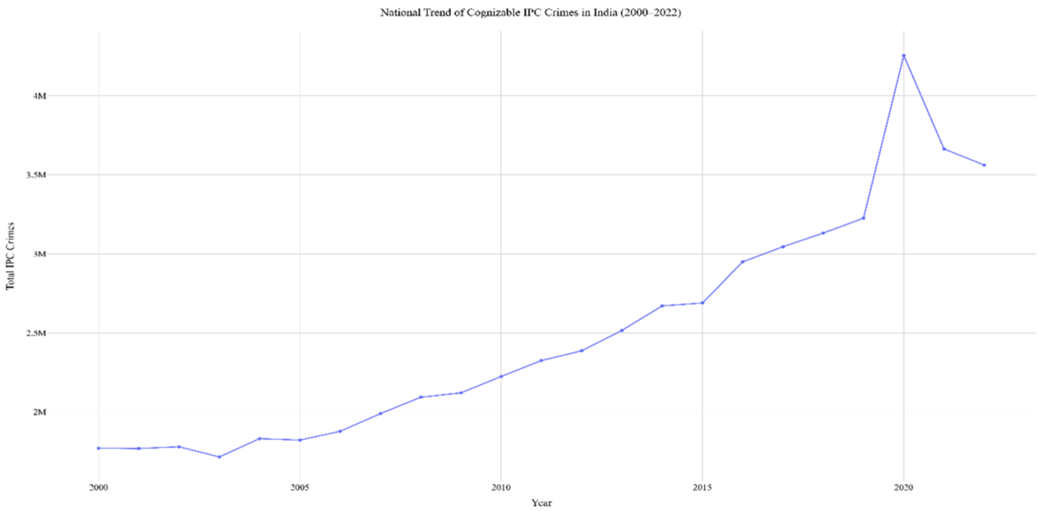

Figure 1: Country-wide trend in cognizable IPC crimes in India during the period 2000-2022 which demonstrates that the crime rate grew rather steadily over the years, reached its peak around the year 2020, and slightly dropped afterwards.

However, a few studies which made ensemble learning available with Indian crime data sets only turned out few where long-term forecasting NCRB data has been explored only little.

2.3 Deep Learning and Spatio Temporal Model

The recent research is using techniques in deep learning to model complex spatio temporal patterns of crimes. The recurrent neural networks, LSTMs and GRLs were popular to capture temporal dependence in the crime data (Mandalapu et al., 2023; Utsha et al., 2024; Shan et al., 2025). Convolution Neural Networks (CNNs) to accomplish spatial identification of hot spots and grids of crime (Jing et al., 2024; Chen & Guestrin, 2016).

More advanced studies involved Graph Neural Networks (GNNs), attention mechanisms, and transformer-based architectures in order to model the spatial adjacency and long range dependencies (Wang et al., 2025; Yang, n.d.). Although these models had achieved high predictive accuracy, they often necessitated large amount of the fine-grained data sets, huge computational power, thus restraining their applicability for aggregated and data poor environments like Indian NCRB type of data sets.

2.4 Indian Crime Prediction and NCRB Based Studies

A relatively small subset of the reviewed literature is specific to the prediction of crime in the Indian context. Studies using NCRB data, conducted mainly descriptive analysis, short-term forecasting, or classification type of prediction using conventional machine learning techniques. Very few studies tried ensemble learning or more than a year of forecasting at the national scale.

Furthermore, spatial correlation between the Indian states and union territories are most of the times neglected, although such correlations are important for understanding the propagation of crime. The lack of consolidated and machine-readable version of NCRB datasets further adds more work, or disconnect, to preprocessing and hamper reproducibility. It is evident through such limitations that a comprehensive and India-specific predictive framework needs to be developed.

2.5 Secondary Literature (Studies and Review Papers)

Several secondary and review-based studies are valuable in terms of crime prediction methodologies, data sets and trends from research. Systematic reviews and surveys provided a systemic analysis of the various machine learning and deep learning techniques, which revealed the key challenges, including the lack of data balance, the lack of interpretability, and the lack of compliance with ethical standards (Mandalapu et al., 2023; Kounadi et al., 2020; Jenga et al., 2023). These studies discuss the increasing importance of ensemble learning, explainable AI and spatio-temporal modelling in crime analytics.

2.6 Summary and Identified Research Gap

Existing studies on crime prediction are largely based on short-term forecasting, single-model approaches, or non-Indian datasets (Mandalapu et al., 2023; Safat et al., 2021; Jenga et al., 2023). Very few studies utilize NCRB datasets for long-term forecasting or incorporate spatial dependencies among states.

In addition, most NCRB-based studies focus on descriptive analysis rather than predictive modelling. The lack of structured and machine-readable NCRB datasets further complicates preprocessing and limits reproducibility.

To address these limitations, this research paper makes the following principal contributions:

- Development of an ensemble-based crime prediction framework using NCRB datasets.
- Integration of XGBoost and CatBoost with Ridge Regression stacking (Dash et al., 2025).
- Incorporation of temporal and spatial features.
- Long-term forecasting up to 2030.
- Identification of crime hotspots using z-score analysis.
- Development of an interactive visualization dashboard.

3. DESCRIPTION AND COLLECTION OF DATA

3.1 Dataset Description

The statistics of crimes utilized in this research are the ones provided by the National Crime Records Bureau (NCRB) on the annual basis. The NCRB publishes a report named as Crime in India which provides detailed crime statistics of all the Indian states and union territories.

The data used were designed in two sets related to this study:

State-wise IPC Crime Dataset (2010-2022): This data contains **annual number of IPC offenses reported** by the Indian Penal Code (IPC) (National Crime Records Bureau, 2022) against each state and union territory. It allows us to see the way offense is transformed throughout a region.

National IPC Sub-Category Dataset (2000-2022): This data set contains specific crime figures in lots of IPC sub-categories all over the entire country. The adjustable longer period (2000-2022) allows us to investigate the trends in the long-term and to make predictions.

Combined, these datasets can allow us to model the differences between crime across states and its time dependence. The model in our framework follows an understandable workflow and transforms raw crime information into credible predictions.

Besides, spatial charts were created based on GeoJSON boundary maps of the Indian states that were also used to indicate hotspots of crime.

3.2 Data Collection Process

The downloadable file available on NCRB is not a single file with all the years of data. Therefore, it was needed to select the required numbers of crime per year out of the annual report, crime in India. This was done by taking table after table in each year, which was copied and inserted in clear and well-organized tables. In so doing it was

Table 1: Model Performance Comparison

Model	RMSE	MAE	R^2
XGBoost	11683.15	6268.41	0.984
CatBoost	16887.54	7882.48	0.967
Ensemble (XGB + CatBoost)	11311.96	5865.43	0.985

rechecked that the state names and classifications of crimes were typed in the same manner each year. Any discrepancies in the manner the data was coded were adjusted in such a manner that the data remained similar over time. Such a structured set created one dataset that could be studied by machine-learners and used to predict in the long run.

4. METHODOLOGY

4.1 Overview of the Framework

The suggested framework will turn the structured crime statistics into predictive data through a multi-stage modelling pipeline. It can be broadly described as data preprocessing, feature engineering, models development, ensemble integration, forecasting, hotspots identification and visualization. The data were split according to a time split method, meaning that the data to be used in the training as well as testing condition is ordered with time and does not become polluted by the information about the further occurrences. Instead of using one algorithm, the research takes a multi model approach to increase its robustness and predictive stability.

4.2 Data Preprocessing

Data was standardized and cleaned up before composing the model in order to ensure some sort of uniformity across the years and states. The preprocessing was performed in the following steps:

- Unification of column names and elimination of unnecessary white space.
- Year by year harmonisation of state and union-territory names.
- Disaggregation of data in a broad format into a long format through melting and reshaping process.
- Handling of missing values using forward-fill strategies when it is appropriate.
- Logarithmic transformation (log1p) would be used to correct instability in variance.

All these preprocessing will ensure the dataset is presented in a format that can be utilized in regression-based machine-learning models.

4.3 Feature Engineering

To extract temporal variations and spatial relationships of the crime information, feature engineering was done. The following features were being constructed:

4.3.1 Lag Features

Lag_1, Lag_2, and Lag_3 variables had been simulated to include the past years values of crimes as input predictors and thus allow the model to be learned to learn the continuity of time (Hossain et al., 2020).

$$\text{Lag}_k(t) = \text{Crime}(t-k) \tag{1}$$

where k is 1,2,3.

The lag function as shown in Equation (1) is used to capture temporal dependency, assuming that past values influence future trends in time-series data.

For example, if crime in 2020 is 1000 and in 2019 is 900, then Lag_1 = 900. This helps the model learn temporal continuity.

Table 2: Prediction Model Performance Evaluation of Crime Prediction Model.

Model	Technique Used	Evaluation Metric	Result
Trend-based Model	Linear Regression	RMSE	Moderate accuracy
Time-Series Model	ARIMA (SARIMAX)	RMSE	Improved accuracy
Machine-Learning Model	XGBoost(Lag-based)	RMSE	Best predictive performance

Table 3: States with Significant Z-Score Deviations

State	Year	Z-Score	Interpretation
A & N Islands	2020	-2.64	Significant decline
Assam	2020	-2.84	Significant decline
Bihar	2020	2.64	High-growth hotspot
Chandigarh	2020	-2.58	Significant decline
D & N Haveli Daman & Diu	2020	2.78	High-growth hotspot
Delhi	2020	3.01	High-growth hotspot
Gujarat	2020	2.83	High-growth hotspot
Haryana	2020	2.78	High-growth hotspot
Karnataka	2020	-2.88	Significant decline
Kerala	2020	2.72	High-growth hotspot
Lakshadweep	2020	2.63	High-growth hotspot
Maharashtra	2020	2.99	High-growth hotspot
Sikkim	2020	-2.91	Significant decline
Tamil Nadu	2020	2.94	High-growth hotspot
Tripura	2020	-2.85	Significant decline
Uttar Pradesh	2020	2.77	High-growth hotspot
Uttarakhand	2020	-2.57	Significant decline

4.3.2 Trend-Based Features

Normalised year index and the year-on-year percentage change were added to get long-term trends.

4.3.3 Differential Features

To model the patterns of growth or decline, the difference between crime numbers in the two years was computed.

4.3.4 Spatial Correlation Features

To calculate the relationships between states, pivot correlation-based k-nearest-neighbor (k - NN) features, individual states were not considered but instead, the relationship between the state and its region was measured.

These constructed variables make the learning algorithms to be able to capture the spatial and the temporal patterns.

4.4 Model Development

Another type of base learners was chosen to be two gradient-boosting:

- **XGBoost Regressor**: This is selected due to its capacity to form nonlinear associations and to utilize the structured tabular data effectively.
- **CatBoost Regressor**: Chosen due to its inductiveness qualities and stability when regressing.

The two models were trained on a time-based split in order to be chronologically intact.

4.5 Ensemble Learning Strategy

A stacking-based ensemble method was adopted in order to improve on predictive performance.

Ridge Regression is used as the meta-learner in the stacking ensemble due to its ability to handle multicollinearity and prevent overfitting through L2 regularization. It combines the predictions of base models by assigning optimal weights while penalizing large coefficients. This results in improved generalization and more stable predictions, making it suitable for crime forecasting tasks.

The Ridge Regression was used to combine the predictions made by XGBoost and CatBoost as given in Equation (2). This methodology lessens prejudice in models and variations though a combination of the capabilities of more than one learner. The individual models were used to compare the ensemble model with their own metrics of RMSE, MAE, and R^2 metrics.

$$\widehat{y}_{ensemble} = w_1\widehat{y}_{XGBoost} + w_2\widehat{y}_{CatBoost} + b \quad (2)$$

4.6 Forecasting Procedure

The rolling forward forecast strategy was used to come up with multi-year forecasting. In this approach:

- With prediction of year t, an input (lag feature) is taken as an input of year t.
- This is a cyclic process that will go on till the year 2030.
- This plan enables long-term prediction, as well as retaining time-related relations.

4.7 Hotspot Identification

Z -score normalisation was used formula is given as Equation (3) to identify the statistically important crime growth between states, in terms of the year-to-year percentage change in the crime rates. The z-score was calculated as:

$$Z = \frac{X - \mu}{\sigma} \quad (3)$$

where:

X= crime value of a state

μ= mean crime value

σ= standard deviation

States where Z > 1.5 were categorized to have high-growth crime hotspots, and this denotes considerably large increases as compared to variation within the nation. On the other hand, those states with a Z < -1.5 were passed as heavily detracting areas. This statistical methodology allows the objective identification of hotspots and not by acting the absolute crime figures (Hajela et al., 2020).

4.8 Visualization Framework

Streamlit was used to create an interactive dashboard to visualise:

- Historical trends
- Forecasted crime levels
- General Spatial distribution Choropleth maps
- Identified hotspots

The presentation of results in an understandable format facilitates decision-making and boosts the interpretability of the results, which is done through the visualisation layer.

The implementation was carried out using Python libraries including Pandas and NumPy for data processing, Scikit-learn for modelling, XGBoost (Chen & Guestrin, 2016), CatBoost (Prokhorenkova et al., 2018), and Streamlit for visualization.

5. EXPERIMENTAL WORK

5.1 Experimental Setup

To maintain the chronological integrity, a split on the basis of time was used. Training was

Table 4: State-wise Distribution of IPC Crimes (latest available year, 2022)

Rank	State	Total IPC Crimes	Observation
1	Uttar Pradesh	High	Major crime hotspot
2	Maharashtra	High	Consistently high crime
3	Madhya Pradesh	High	Rising trend
4	Rajasthan	Moderate	Stable crime rate
5	Tamil Nadu	Moderate	Urban influence

done in previous years and testing was done in successive years so that information leakage would not occur. The experimental setup was implemented using Python libraries such as Scikit-learn, XGBoost, CatBoost, Pandas, and NumPy.

5.2 Evaluation Metrics

These metrics were used to gauge model performance:

- **Root Mean Square Error (RMSE)**
- **Mean Absolute Error (MAE)**
- **Coefficient of Determination (R^2 Score).**

5.3 Model Performance Comparison

The ensemble model had the lowest RMSE (11311.96) and MAE (5865.43) and the highest R^2 score (0.985), which indicates that the ensemble model provides better predictive stability and performance compared to individual models.

5.4 National-Level Sub-Crime Forecasting

Particularly, the forecast analysis performed on IPC has been done on both the state and IPC sub-category level throughout the nation using both ARIMA/SARIMAX models and lag-based XGBoost features. The framework upholds the short term and multi-year forecasting horizon, both at the state and nation levels.

5.5 Observations

- The improved performance of the ensemble model is attributed to Ridge Regression, which reduces overfitting by applying regularization and effectively combines predictions from XGBoost and CatBoost, resulting in improved generalization and lower prediction error (Dash et al., 2025).
- Time can be divided into lag and trend variables that helped in learning.
- Hotspots identification was effectively achieved with the help of spatial correlation characteristics.
- The model was found to be stable compared to dissimilar temporal segments.

The findings presented in **Table 1** prove that machine-learning model is better than the conventional model. **Table 2 presents a** comparative analysis of various predictive models. The XGBoost-based model is more efficient because it can assess non-linear nature of a crime. In Figure 1, long-term national trend of cognizable IPC crimes is presented with characteristics of steady increase throughout the 20-year period with its peak at approximately 2020. Figure 1: Country-wide trend in cognizable IPC crimes in India during the period 2000-2022 which demonstrates that the crime rate grew rather steadily over the years, reached its peak around the year 2020, and slightly dropped afterwards.

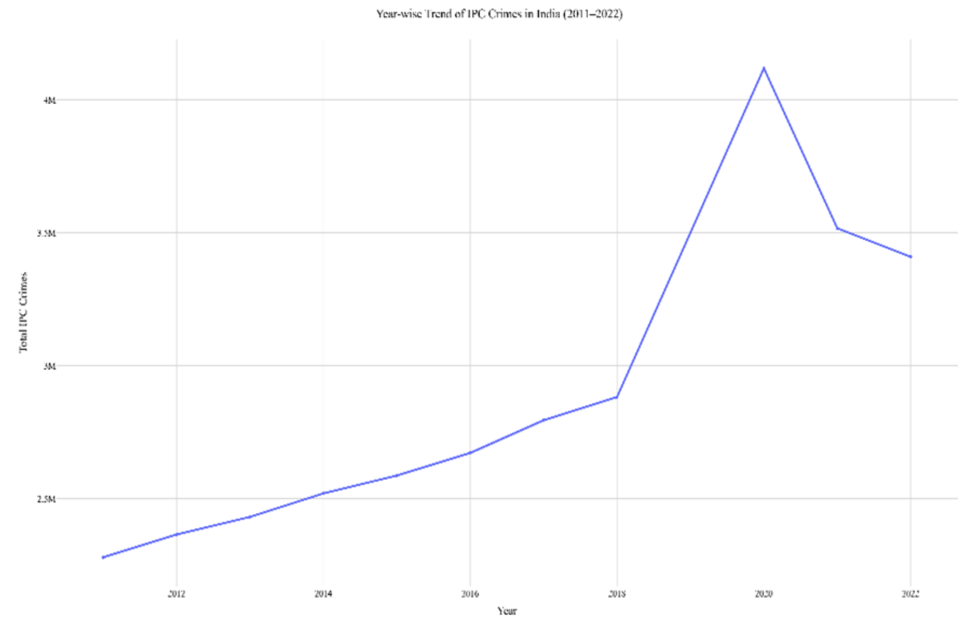

Figure 2: Year-wise trend of IPC crimes across Indian states (2010–2022), illustrating temporal variations and inter-year changes.

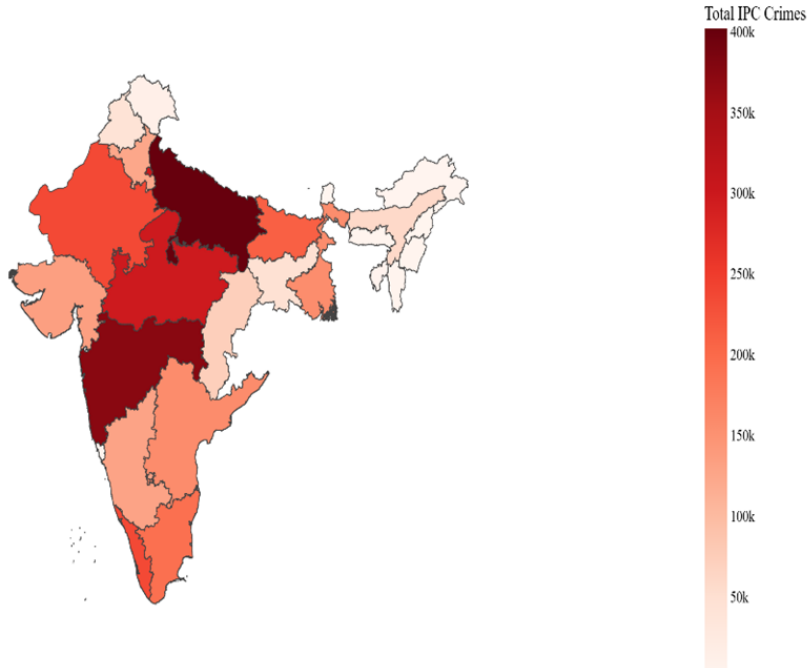

Figure 3: Spatial distribution of IPC crime across Indian states based on the latest available data. The visualization highlights relative variations in crime intensity, where states with higher values are represented with stronger colour intensity.

6. RESULTS AND DISCUSSION

In this section, the key findings of performing machine-learning and ensemble-based crime forecasting models are outlined. These results also prove the effectiveness of the suggested model to forecast crime tendencies and define hotspots that emerge in the Indian states.

6.1 State-wise Crime Forecasting Results

The trained models were then used to generate crime forecasts to all States and Union Territories

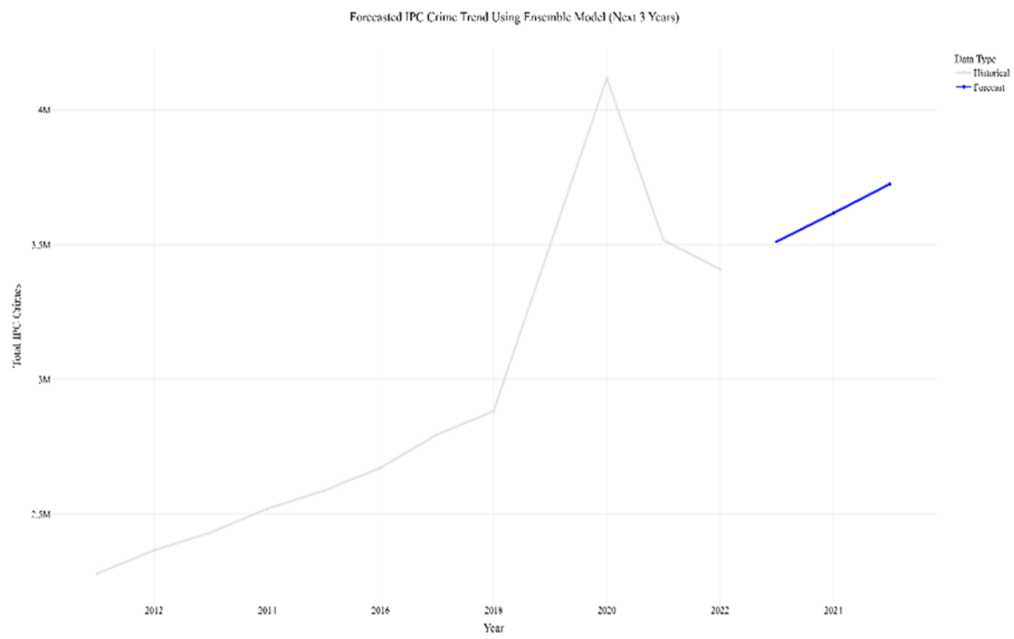

Figure 4: Forecasted trend of IPC crimes for the next three years (2023–2025), based on historical data, indicating a gradual increasing pattern.

of India that covers the year 2023 to 2030. The Ensemble Model was successful in giving smooth and continuous trend lines and this confirms its forecasting ability over the long-term. Those state authorities that had had high levels of crime in the past still obtained high levels of prediction, which validated the ability of the model to learn the temporal patterns.

The smooth prediction trends indicate that the ensemble model effectively captures long-term temporal dependencies in the data.

6.2 Comparative Performance Analysis

The findings show that the ensemble model was statistically superior to the individual models on every measurement metric. The decrease in the values of the RMSE and MAE error indices proves the improvement of predictive accuracy, whereas the high score of the R^2 supports interpretative accuracy.

6.3 Hotspot Identification Results

The Z-score-based analysis was performed as shown in Table 3 to examine year-on-year variations in crime growth across Indian states and union territories. The Z-score standardizes the percentage change in crime rates, allowing comparison across regions irrespective of scale differences. From the analysis, several states exhibited statistically significant deviations from the national trend. States such as Delhi (Z = 3.01), Maharashtra (Z = 2.99), Tamil Nadu (Z = 2.94), Gujarat (Z = 2.83), Haryana (Z = 2.78), and Uttar Pradesh (Z = 2.77) were identified as high-growth hotspots, indicating a sharp increase in crime rates during the observed period. Additionally, Bihar, Kerala, Lakshadweep, and D&N Haveli Daman & Diu also demonstrated strong positive deviations, reinforcing the presence of emerging high-risk regions. Conversely, several states showed significant negative deviations. Sikkim (Z = −2.91), Karnataka (Z = −2.88), Tripura (Z = −2.85), Assam (Z = −2.84), A&N Islands (Z = −2.64), Chandigarh (Z = −2.58), and Uttarakhand (Z = −2.57) were classified as regions experiencing a notable decline in crime trends. The remaining states fall within the Z-score range of −1.5 to +1.5 and are therefore categorized under normal variation. These states do not exhibit statistically significant changes and indicate relatively stable crime patterns over time. Overall, the analysis highlights the presence of both rapidly increasing crime hotspots and regions with declining trends.

This spatial variation emphasizes the importance of region-specific policy interventions and supports the need for predictive modelling to assist in proactive crime management.

Note: The remaining states fall within normal variation ($-1.5 \leq Z \leq 1.5$).

6.4 National Sub-Crime Trend Analysis

At the national level forecasting of IPC subcategories, it is observed that some categories have a steady growth whereas others depict signs of stabilising or even depreciating.

The observed trends reflect non-linear patterns in different crime categories, which justifies the use of ensemble learning methods for improved prediction accuracy.

6.5 Visualization Outcomes

The dashboard built in Streamlit served as a good visualisation of the trends of the past as well as forecasting the future, choropleth map and hotspots. State-wise interactive comparison is more interpretable, and results will be accessible to non-technical stakeholders.

Certain states such as Jammu & Kashmir and Odisha are not prominently highlighted as hotspots because their crime values do not exceed the defined statistical threshold for hotspot classification. This indicates that while these states may have notable crime levels, they are comparatively lower than high-intensity regions such as Uttar Pradesh and Maharashtra.

The map represents relative differences rather than absolute categorization. Table 4 complements this figure by presenting the corresponding numerical values for the latest available year (2022).

Table 4 presents the state-wise distribution of IPC crimes for the latest available year (2022). The values correspond to the spatial patterns observed in Fig. 3, providing a quantitative representation of high and moderate crime regions.

In case one is looking at the past three years, Figure 4 shows the trend of the historical data of IPC crimes.

Figure 4: Forecasted trend of IPC crimes for the next three years (2023–2025), based on historical data, indicating a gradual increasing pattern.

The forecasts presented below are useful in giving an idea on the budget of the future crime and thus aid the policymakers in developing counter-measures. The graphical representation of the comparative values of the past and future provides an insight into the path the history will take to the forecasted values.

Table 5: Forecasted National IPC Crime Levels with Percentage Change

Year	Predicted Crime	% Change	Trend
2023	3,511,364	+3.00%	Increasing
2024	3,616,705	+3.00%	Increasing
2025	3,725,206	+3.00%	Increasing

Actual crime data for the forecasted years is not available at the time of study, as NCRB data is published with a delay. Therefore, only predicted values are presented in this table.

The forecasted results indicate a steady increasing trend in national IPC crime levels over the projected period. The predicted values show a consistent annual growth of approximately 3%, suggesting a gradual rise in crime under current patterns. This trend highlights the need for proactive policy measures and improved crime prevention strategies to mitigate future risks. As summarized in Table 5, future predictions of crime can be used to help in visualising the picture of Figure 4.

7. CONCLUSION

The new upgraded machine learning structure with an addition of an ensemble approach to the original one has been introduced in the analysis of crime prediction based on the data collected by the National Crime Records Bureau (NCRB). The present study detects and addresses some of the limitations present in previous studies of the crime prediction in India. The effectiveness of the stacking methodology is supported by the fact that the ensemble model has a better predictive ability as compared to other methods. The fact that the model can determine the crime levels to 2030, identify the hotspots and show the results through the dashboard highlights its applicability. All these findings support a scaled, data-driven design that can be used to make evidence-based decisions. The ensemble model achieved RMSE of 11311.96, MAE of 5865.43, and R^2 score of 0.985. This study demonstrates the effectiveness of ensemble learning in large-scale crime prediction and provides a scalable framework for data-driven crime prevention in India.

Acknowledgement

The author gratefully acknowledges the Department of Computer Science, University of Delhi, Delhi, India, for providing essential facilities and a supportive environment for this research.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of use of generative AI and AI-assisted technologies in the article

The author did not use AI and AI-assisted technologies in writing this article.

Conflict of Interest

The author has no conflict of interest to declare.

REFERENCES

Abubakera, H., Muchtar, F., Degan, K. S., Mohd Azmi, K. H., Abd Hamid, F. K., Binti Along, N. Z., & Bin Khairuddin, A. R. (2025). Crime Prediction based on Classification Approaches. *Procedia Computer Science*, *259*, 1407–1415. https://doi.org/10.1016/J.PROCS.2025.04.095

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785

National Crime Records Bureau. (2005). *Crime in India 2005*. Ministry of Home Affairs, Government of India. https://www.data.gov.in/catalog/crime-india-2005

National Crime Records Bureau. (2021). *Crime in India - 2021*. Open Government Data (OGD) Platform India. https://www.data.gov.in/catalog/crime-india-2021

National Crime Records Bureau. (2022). *Crime in India - 2022*. Open Government Data (OGD) Platform India. https://www.data.gov.in/catalog/crime-india-2022

Dash, T., Jain, T., Shukla, A., V., B. C., & Samieiyeganeh, M. (2025). *Crime Forecasting Using Ensemble Learning*. 1–6. https://doi.org/10.1109/IACIS65746.2025.11211145

Hajela, G., Chawla, M., & Rasool, A. (2020). A Clustering Based Hotspot Identification Approach For Crime Prediction. *Procedia Computer Science*, *167*, 1462–1470. https://doi.org/10.1016/J.PROCS.2020.03.357

He, J., & Zheng, H. (2021). Prediction of crime rate in urban neighborhoods based on machine learning. *Engineering Applications of Artificial Intelligence*, *106*, 104460. https://doi.org/10.1016/J.ENGAPPAI.2021.104460

Hossain, S., Abtahee, A., Kashem, I., Hoque, M. M., & Sarker, I. H. (2020). Crime prediction using spatio-temporal data. *Communications in Computer and Information Science*, *1235 CCIS*, 277–289. https://doi.org/10.1007/978-981-15-6648-6_22/TABLES/7

Jenga, K., Catal, C., & Kar, G. (2023). Machine learning in crime prediction. *Journal of Ambient Intelligence and Humanized Computing 2023 14:3*, *14*(3), 2887–2913. https://doi.org/10.1007/S12652-023-04530-Y

Jing, C., Lv, X., Wang, Y., Qin, M., Jin, S., Wu, S., & Xu, G. (2024). A deep multi-scale neural networks for crime hotspot mapping prediction. *Computers, Environment and Urban Systems*, *109*, 102089. https://doi.org/10.1016/J.COMPENVURBSYS.2024.102089

Kabiraj, P. (2023). Crime in India: a spatio-temporal analysis. *GeoJournal*, *88*(2), 1283–1304. https://doi.org/10.1007/S10708-022-10684-7

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*(1), 4830411. https://doi.org/10.1155/2022/4830411

Kounadi, O., Ristea, A., Araujo, A., & Leitner, M. (2020). A systematic review on spatial crime forecasting. *Crime Science*, *9*(1). https://doi.org/10.1186/S40163-020-00116-7

Mandalapu, V., Elluri, L., Vyas, P., & Roy, N. (2023). Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions. *IEEE Access*, *11*, 60153–60170. https://doi.org/10.1109/ACCESS.2023.3286344

Nicho, M., Hamed, A., Gaber, T., & Al Arimi, J. H. (2025). A-XGBoost: a resilient machine learning technique for predicting crimes against women across cultures on low cardinality crime data. *Cogent Social Sciences*, *11*(1). https://doi.org/10.1080/23311886.2025.2527392

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, *2018-December*, 6638–6648.

Safat, W., Asghar, S., & Gillani, S. A. (2021). Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques. *IEEE Access*, *9*, 70080–70094. https://doi.org/10.1109/ACCESS.2021.3078117

Shan, M., Ye, C., Chen, P., & Peng, S. (2025). Ada-GCNLSTM: An adaptive urban crime spatiotemporal prediction model. *Journal of Safety Science and Resilience*, *6*(2), 226–236. https://doi.org/10.1016/J.JNLSSR.2024.11.003

Utsha, R. B., Alif, M. N., Rayhan, Y., Hashem, T., & Ali, M. E. (2024). *Deep Learning Based Crime Prediction Models: Experiments and Analysis*. http://arxiv.org/abs/2407.19324

Wang, S., Zhang, Y., Piao, X., Lin, X., Hu, Y., & Yin, B. (2025). MRAGNN: Refining urban spatio-temporal prediction of crime occurrence with multi-type crime correlation learning. *Expert Systems with Applications*, *265*, 125940. https://doi.org/10.1016/J.ESWA.2024.125940

Zhang, X., Liu, L., Xiao, L., & Ji, J. (2020). Comparison of machine learning algorithms for predicting crime hotspots. *IEEE Access*, *8*, 181302–181310. https://doi.org/10.1109/ACCESS.2020.3028420

COPYRIGHT AND LICENSE

HOW TO CITE?

Sanjana Dwivedi. (2026). Enhanced Approach for Crime Predictive Analysis Using NCRB Dataset. *Vantage: Journal of Thematic Analysis*, 07(01), 01-13. https://vantagejournal.com/