International Journal of Innovation and Multidisciplinary Research
DOI.ORG/10.67505/IJIAMR.v6.i1.61001
Vol. 6 (1), 2026
E-ISSN: 2583-4452

Crime Pattern Analysis on Crime Dataset using Hybrid Ensemble Model and SARIMAX

Sneha[1*]
[1] Department of Computer Science, Shaheed Rajguru College of Applied Sciences for Women, University of Delhi, Delhi
[*] Corresponding Author ✉ snehaofficial535@gmail.com

ABSTRACT

The requirement for crime prevention and prediction continues to grow with regard to the manner in which urban governments are governed. Law enforcement agencies have evolved from being reactive in response to criminal incidents to proactive by attempting to anticipate or prevent those events. This paper outlines a large-scale computational approach using over 40,000 criminal record data points from 29 cities across India (2020-2024) to support an "anticipate and prevent" paradigm. A dual-pathway analytical framework is utilized to address the complexity of working with real-world data. The first pathway utilizes a SARIMAX model with a one-year "cold-start," to abstract stable long-term trends within the raw data. The second pathway incorporates a high-precision hybrid ensemble utilizing Random Forest and Gradient Boosting to assess multi-dimensional characteristics (city encoding/demographics), which demonstrated a mean absolute error of 0.24. Additionally, an optimized version of Agglomerative Clustering with Silhouette Denoising was utilized to categorize spatial risk profiles of the cities and to identify new hotspot locations. The utilization of this framework provides a bridge from the theoretical application of machine learning to its practical application. Therefore, it provides a validated methodology to perform both robust noise-resilient regional analyses and accurate high-fidelity local predictions. Overall, the results presented herein provide authorities with the capability to utilize these data to inform resource allocation strategies that will enable their continued effectiveness in preventing crime as the evolving nature of criminal behavior dictates.

Keywords: *Crime Prediction, SARIMAX, Hybrid Ensemble, Clustering, Machine Learning, Hotspot Identification*

1. Introduction

Making sure that the population stays safe has become a very complex issue in the fast-changing environment of modern-day urban centres. Traditionally, police approaches have always been reactive. The so-called Report and Respond approach suggests that the relevant parties act only once something illegal occurs and has already been documented. Unfortunately, as the urban environment changes fast in the case of India because of increasing urbanization and other factors, the reactive approach becomes less effective when the number of crimes increases, and new types of illegal activities emerge (Huamantingo et al., 2025). Moreover, the reliance on historical data to determine how police forces will operate in the future creates a certain gap between the efforts made by authorities and criminal organizations, making any progress impossible. In addition, the need for change is especially pressing in case of India as a country that includes 29 major urban centers with their specific crime rate, level of documentation, socio-cultural backgrounds, and other aspects.

The growing availability of high-dimensional crime-related data – a result of digitization of police records and national data archives – presents an extremely important possibility to shift the paradigm of law enforcement towards "predictive policing". This intelligence-driven approach aims at changing the trend from reactive to preventive and focuses on prediction and interception of any criminal activity prior to its actual occurrence (Ospina et al., 2025). Predictive policing is the use of analytic methods to select promising police targets and prevent crimes based on prediction of "when and where" the criminal event will take place. Using such factors as historical trends, timing, seasonal events, or agricultural cycles, among others, and making a statistical analysis of available data, one can obtain valuable insights and turn raw data into a strategic resource to understand what underlies urban crime. Moreover, predictive models make it possible to analyze the "City DNA," which stands for specific behavioral traits of a city.

There are many barriers for an individual or company to take an idea (algorithm) and put it into practice. Most scholarly literature creates a model that can make almost perfectly accurate predictions using cleaned up data; however, most people who attempt to use these models will find them to be useless in their day-to-day operations. One of the major

Received on: 27 Apr 2026 | Accepted on: 28 Jun 2026 | Published Online on: 25 July 2026

problems that operational users of algorithms experience is a lack of data density in specific geographic regions as well as "noise" in incident reporting systems (Chakraborty et al., 2024). In this case, "noise" means inconsistent timing of incidents, data entry errors by people, as well as the unpredictable nature of crime waves, which are rarely consistent. All these factors may cause instability in ordinary algorithms, resulting in both false positive predictions, consuming scarce police resources, and false negative predictions, making communities unprotected against the predicted crimes. Therefore, it is crucial to develop a robust mathematical system, able to withstand all inconsistencies. There is an obvious necessity for a solution, combining a macro perspective, allowing for predictions based on cycles, as well as a micro one, helping to estimate daily risk factors.

The bulk of the literature in the domain of crime prediction is devoted to a single modeling architecture. That could be a Time Series model such as ARIMA to follow trends or a classifier such as Random Forest to identify the type of crimes. Nevertheless, crime itself is a multi-dimensional entity. A single-modeling approach often cannot deal with "Cold Start" problems (a lack of data needed to initiate the model) and "Volume Variance" issues (high numbers of crime in bigger cities that skew the numbers in the smaller city, like the comparison between Delhi and Chandigarh). If the model is too responsive, it will exaggerate temporary spikes, whereas a stable model will fail to detect new hotspots. The present research finds out that there is an answer to this question – "ensemble hybridization" and "dual-pathway verification." This means that two different mathematical approaches need to be run in parallel for law enforcement agencies to have a "check-and-balance" mechanism to ensure that any spike in crime is indeed dangerous.

This research attempts to tackle these types of challenging problems using an entirely new method for computing through a dual pathway model which provides both high fidelity predictive crime forecasting as well as detailed regionalized risk assessment. The research used a large dataset of information, greater than 40,000 records from over five years (from 2020 to 2024), from 29 different cities in India, and provided intelligence instead of just visualizing it.

There are two distinct streams within the research design methodology:

1. **Pathway A:** The Statistical Approach for Forecasting Macro Trends via the Noise-Robust SARIMAX Framework. This framework utilizes a statistical model to forecast macro trends. With a "Cold Start" period of one year, it will be able to extract stable seasonal patterns and provide an "accurate safety baseline."
2. **Pathway B:** The Machine Learning Algorithm Utilizing the Precision Focused Hybrid Ensemble Framework. This machine learning algorithm uses the best attributes of Random Forest in order to reduce the variability (variance), and use Gradient Boosting in order to reduce bias. It will predict local daily crime intensity with a very small Mean Absolute Error (MAE) of 0.24.

Crime prediction in urban areas has been a long-standing challenge, as shown through many real-world examples where law enforcement agencies have failed to accurately predict and prevent crimes. This study will attempt to overcome those challenges using a unique dual-path computational model that can be used to create accurate crime occurrence predictions and geographic risk categories. Using an extensive dataset (over 40,000 entries) for 29 different Indian cities over the years of 2020-2024, the current study is both able to visualize crime trends as well as analyze crime intelligence. Additionally, the proposed methodology includes an unsupervised learning component utilizing Agglomerative Clustering and Silhouette Denoising. Unlike other methodologies that classify cities based on physical boundaries (i.e., political divisions), the proposed methodology uses algorithms that are capable of identifying behavioral patterns among citizens, thereby allowing for identification of crime archetypes.

The main contribution of the current study is creating a methodological framework that bridges theory-based machine learning techniques into practical implementations. The proposed methodology does not simply focus on forecasting crime numbers, but instead provides a multi-faceted understanding of urban environments. At the macro level, the methodology creates robust regional analyses that account for data "noise" and sparsity while providing precise localized forecasts.

These consequences are quite significant in terms of how cities are administered in India. This study's results will allow administrators to utilize their resources efficiently so that they can assign their patrol units to crimes as they occur today rather than in areas where crimes will likely take place in the future due to a particular 'DNA' of the city and the time of year in which these crimes are most frequently committed. As such, the study provides a reliable base for potential future application of predictive policing technologies to track changes in criminal behavior across both physical and virtual realms.

Research Question: Compared to individual linear models and ensemble models, what is the reduction in residual generalization errors (rim) when a Hybrid Ensemble model (Random Forest + Gradient Boosting) is used for local daily intensity forecasting?

2. Literature Review

2.1 Ensemble Learning and Hybrid Models (Random Forest & Gradient Boosting)

The fact that Ensemble Learning methods have been found to be a benchmark method to predict crimes at a local level effectively (Ms. D.S. Smitha Mol et al., 2026), indicates that the application of Multi-Model approaches (using Random Forest and Gradient Boosting Regressor) has provided very good results when using the Random Forest Model to model domestic violence and achieved an R2 score of .93; similarly, the work done in (Ibrahim et al., 2024) shows that the combination of Parallel Bagging (used by the Random Forest Model), and Sequential Boosting greatly reduces the Mean Absolute Error (MAE); while the results obtained in (Ibrahim et al., 2024) and (Kshatri et al., 2021) show that the proposed hybrid models can address both the high Variance, and High Bias problem presented with complex Urban Data; Finally, the utilization of the Random Forest approach to analyze the Crime Severity represents the same High Resolution Tactical Process applied in this Study (Saravag & Kumar, 2024).

2.2 Seasonal and Statistical Time-Series Forecasting (SARIMA/SARIMAX)

To deal with the "noise" and sparseness within the raw data, the SARIMA/SARIMAX has become an essential model used to capture any trend that can be derived from the data. SARIMA in Matlab are used to find the statistically significant 20-month cycles, indicating that there is a heavy seasonality in crime (Redoblo et al., 2023). Likewise, in another study conducted, it was discovered that crime in regions follows the socio-temporal calendar (Noor et al., 2022). Additional evidence is cited, which shows that SARIMA has the unique ability to filter out the reporting noise and find the "Safety Baseline" (Kurniawan et al., 2025). It is also the statistical methodology used to build the macro-trend base of our framework (Thayyaba Khatoon Mohammed, 2025).

2.3 Spatial Segmentation and Hotspot Identification (K-Means/Agglomerative Clustering)

Unsupervised learning for the discovery of emerging hot spots provides an objective categorization of risk profiles in urban environments. The paper has identified the first set of principles to group incident clusters with the goal of optimizing police deployment resources (Crime Forecasting and Analysis through K-means Clustering Algorithm, n.d.). The principle was refined by combining spatial clustering and temporal intensity to forecast future crime zones (Buland Iqbal et al., 2024). Research indicates that if spatial clustering is performed after PCA, it is capable of filtering out "City DNA" from volumetric variability (Ospina et al., 2025) (Maharrani et al., 2024). Lastly, the research shows that spatial knowledge enables law enforcement to progress beyond general city policies to archetypical inter-ventions (Rasan et al., n.d.).

Paper Full Name	Year	Model (s) Used	Performance	Research Gap / Focus
Machine Learning Crime Prediction Models and the Gap Between Research	2025	SHAP, XAI	99.5% Acc.	Addressed the "Trust Gap" in black-box models. (Huamantingo et al., 2025)
Enhanced Spatial Clustering for Crime Analysis: Ward-Like and SKATER	2025	SKATER	Contiguity	Improved geographical realism in risk zones. (Ospina et al., 2025)
Empirical Approaches to Crime Rate Prediction: Challenges and Insights	2024	Baseline ML	89% Acc.	Highlighted the benefit of PCA in denoising. (Chakraborty et al., 2024)
An Efficient Machine Learning Approach for Crime Detection in India	2026	RF, GBR	R^2: 0.93	Focused on specific domestic violence categories. (Ms. D.S. Smitha Mol et al., 2026)
A hybrid machine learning model for crime rate prediction	2024	SVM-RF	R^2: 0.94	Used PCA to stabilize regional predictions. (Ibrahim et al., 2024)
An Empirical Analysis of Machine Learning Algorithms for Crime Prediction Using Stacked Generalization: An Ensemble Approach	2021	Stacked Ensemble	99.5% Acc.	Demonstrated that stacked ensembles significantly outperform standalone base models. (Kshatri et al., 2021)
A Hybrid ML and Regression Approach Crime Against Women	2024	RF, XGBoost	R^2: 0.89	Prioritized crime severity in vulnerability maps. (Saravag & Kumar, 2024)
Forecasting the influx of crime cases using SARIMA	2023	SARIMA	Validated	Established 20-month seasonal cyclic patterns. (Redoblo et al., 2023)

Paper Full Name	Year	Model (s) Used	Performance	Research Gap / Focus
SARIMA: A Seasonal Autoregressive Integrated Moving Average Model	2023	SARIMA	MAE: 0.066	Proved impact of social/religious calendars. (Noor et al., 2022)
Trend Analysis and Prediction of Violence Against Women: Jakarta	2025	ARIMA, SARIMA	Stat. Sig.	Correlated victim education levels with trends. (Kurniawan et al., 2025)
A Novel Fusion Approach Hybridization of ARIMA and RNN	2024	ARIMA-RNN	-20% Error	Modelled non-linear residuals in time-series. (Thayyaba Khatoon Mohammed, 2025)
Crime forecasting and analysis through K-means	2021	K-Means	Effective	Established the baseline for hotspot delineation. (Crime Forecasting and Analysis through K-means Clustering Algorithm, n.d.)
A 2D Clustering Based Hotspot Identification Approach	2024	2D Grid+LSTM	Low RMSE	Tracked dynamic evolution of hotspots over time. (Buland Iqbal et al., 2024)
Clustering method for criminal crime acts using K-means and PCA	2020	K-Means+PCA	Optimal	Used PCA to reduce dimensionality before clustering. (Maharrani et al., 2024)
Crime rate prediction and analysis using K-means	2020	Optimized KM	50% Rule	Proved 10% of areas generate 50% of crimes. (Rasan et al., n.d.)
A Comprehensive Computational Framework for Crime Rate Prediction	2026	Hybrid ML	High Precision	Applied a localized framework for Indian Metros. (V. Tikore et al., 2026)
Next-Gen Crime Analytics Using Big Data & Machine Learning	2025	K-Means, XGB	94% Acc.	Focused on the scalability of high-velocity data. (Mathur & Jain, 2025)
Hybrid ARIMA-ANN for Crime Risk Forecasting	2025	ARIMA-ANN	99.8% Acc.	Integrated socio-economic factors as inputs. (Iacobescu & Susnea, 2025)
A Multi-Model Machine Learning Approach Spatio-Temporal Data	2024	CNN-LSTM	F1: 0.94	Captured "where" and "when" ripple effects. (Shinde et al., n.d.)
Leveraging transfer learning with deep learning for crime prediction	2024	BiLSTM	Red. Time	Applied Transfer Learning from Chicago to others. (Butt et al., 2024)
Deep Learning for Crime Pattern Recognition	2023	TabNet, CNN	94% Acc.	Focused on high-dimensional categorical records. (A et al., 2023)
Evaluating Crime Type and Crime Pattern Analysis Using ML	2024	XGBoost, GB	94% Acc.	Addressed overfitting in localized small datasets. (P et al., 2024)
Empirical Analysis for Crime Prediction ML and DL Techniques	2024	DNN, RNN, RF	89% Acc.	Benchmarked deep vs. shallow learning models. (Safat et al., 2021)
Crime Prediction Methods Based on Machine Learning: A Survey	2023	Review	Benchmark	Synthesized Spatio-Temporal trend methods. (Yin, 2023)
A Novel Clustering & ML Algorithm for Crime Rate Prediction	2023	K-Means, Reg.	4 Zones	Categorized NCRB data into tactical risk zones. (Ramanan et al., 2023)
Crime Data Analysis and Safety Recommendation System	2023	RF, KNN	87.6% Acc.	Provided proactive risk alerts for citizens. (Akil et al., 2023)
Urban Crime Trends Analysis based on LightGBM	2021	LightGBM	Fast Train	Optimized for real-time urban trend analysis. (Tong et al., 2021)
CRIME Type and Occurrence Prediction Using ML Algorithm	2021	K-Means+RF	93% Acc.	Labelled unstructured records before prediction. (Kanimozhi et al., 2021)

Paper Full Name	Year	Model (s) Used	Performance	Research Gap / Focus
Crime forecasting: a ML and computer vision approach	2021	CNN-RF	97% Acc.	Fused CCTV weapon detection with forecasting. (Shah et al., 2021)
Clustering of Indian States Based on Crime Incidences	2021	Hierarchical	R^2: 0.60	Linked police strength to murder/dowry trends. (Chatterjee et al., n.d.)
Machine Learning for Crime Analysis and Prediction	2020	Supervised	85-90% Acc.	General classification of incident types. (Jha et al., 2022)
Real-Time Crime Prediction in India Using ML	2020	RF, KNN	87.6% Acc.	Focused on early real-time Indian datasets. (Sarkar, 2025)
Machine Learning Algorithms under Indian Penal Code	2019	Decision Tree	88% Acc.	Classified crimes by IPC section categories. (Aziz et al., 2024)
Crime Prediction and Analysis Using Machine Learning	2019	Baseline ML	82-85% Acc.	Early comparative study of ML algorithms. (Balu et al., 2022)
Predictive Crime Rate Analysis System	2018	Regression/RF	Baseline	Early attempt at forecasting yearly volumes. (I. Singh et al., 2025)
A Study on Smart ML Tools for Crime Detection	2018	Tool Review	Utility	Compared efficiency of various software toolkits. (S. Singh et al., 2024)
A Clustering-Based Method for Hotspot Recognition	2017	Density-Clust	High Rel.	Focused purely on spatial reliability. (Dubey et al., 2024)
The Indian Police Journal: Special Edition on AI in Policing	2025	Policy Review	Operational	Addressed the strategic feasibility of AI tools. (Rajeev et al., n.d.-a)

3. Research Methodology

This suggested framework architecture aims to be a Dual-Pathway Computational Framework that tackles both long-term statistics and short-term prediction accuracies. The dual-pathway system allows for the isolation of the noise from the raw incident logs from the engineering features, which helps provide regional stability and local accuracy.

3.1 Computational Environment and Data Acquisition

Empirical evidence for this research lies in the availability of extensive data in the form of 40,000+ crime reports over a period of 2020-2024 in 29 metropolises in India, available from the open-access repository Kaggle. The computing environment for this experiment was based on Python v3.12 with the help of Scikit-Learn library and Pandas and NumPy packages. This amount of data is based on the standards set for multi-city Indian crime databases (V. Tikore et al., 2026).

3.2 Data Preprocessing and Feature Engineering

Beginning with the process of converting unprocessed log data into the necessary format for an accurate model of crime, a series of pre-processing steps were required prior to modeling.

- Temporal Decomposition: In order to identify that crimes occur in cycles over time, the "date of occurrence" field was converted from a single date value to a four-element vector representing Year, Month, Day, and Hour.
- Label Encoding: In order to convert categorical fields with large numbers of unique values, such as City, Crime Domain, and Weapon Used, they are encoded using integer encoding. This is used to allow the models to treat categorical attributes such as "city DNA" as being statistically significant while still allowing the models to understand these types of attributes.
- Temporal Splits for Evaluating Models' Predictive Capability on Future Data: Temporal Split is utilized to mimic the process of time-series data. Therefore, the Temporal Split technique is utilized, where the training period is 2020-2023 and the testing period is 2024. Consequently, we are evaluating how well our model predicts data that will occur after our data split. A common issue referred to as "Data Leaking" occurs when an algorithm uses all available data during both training and validation (Shinde et al., n.d.).

3.3 Pathway A: Macro-Trend Forecasting (SARIMAX)

Acknowledging the fact that public safety data is inherently noisy, Pathway A adopts a SARIMAX model.

- Cold Start Strategy: The model employs a cold start strategy of one year's data (2020) to establish baseline seasons, a method found to be effective according to the study conducted by Sarawag and Kumar in 2024 for detecting irregularities in reporting (Thayyaba Khatoon Mohammed, 2025).
- Optimization: Hyperparameters (p, d, q) * (P, D, Q, s) have been optimized using a Grid Search to minimize the Akaike Information Criterion (AIC).
- Goal: Pathway A acts as a noise-resilient filter and identifies the “Safety Baseline” to detect deviations.

3.4 Pathway B: High-Precision Hybrid Ensemble

In the case of localized intensity prediction for a day, the Hybrid Ensemble model was designed by combining the capabilities of RF and GBR models.

- Bagging Component (RF): Trained using 100 estimators to lower the variance by taking the average of multiple decision trees' outputs.
- Boosting Component (GBR): Employs a learning rate of 0.1 with additive modelling to lower the error rate of the prior iteration (P et al., 2024).
- Hybrid Fusion: In this model, the tactical forecast is based on the weighted average of the bagging and boosting components, which has successfully yielded a Mean Absolute Error (MAE) of 0.24. (Ms. D.S. Smitha Mol et al., 2026) (P et al., 2024).

3.5 Spatial Segmentation (Agglomerative Clustering)

A Hierarchical Agglomerative Clustering algorithm for detecting emerging hotspots was included as a part of unsupervised learning.

- Dimensionality reduction: Before applying the clustering algorithm, data was normalized through Log1p normalization to address the problem of difference in size of data for big and small cities. Then PCA algorithm was employed for reducing the dimensionality to 2 dimensions maximizing variance of principal components (Maharrani et al., 2024).
- Denoise Silhouette Mask: The complete linkage method was used for clustering cities according to their closeness in the reduced dimensional space. The result was denoised by the Silhouette Mask (with a score higher than 0.1) to remove “boundary outliers” and form stable clusters (Ospina et al., 2025).

4. Theory and Calculation

From a theoretical perspective, the current study employs a mixed approach comprising ensemble machine learning and stochastic time series modelling as the backbone of the entire study. This section of the essay is focused on the logical reasoning behind each component of the two-pathway model.

4.1 Theoretical Foundation of Temporal Modelling

In addition to identifying trends and irregular patterns, the three major components of public safety data were identified including seasonality. However, the primary challenge to developing effective algorithms for processing public safety data is the presence of outlier/noise in the data. SARIMAX was selected for this research because it incorporates seasonal component (s), exogenous variable (s), and extends upon the ARIMA methodology. It also can be represented mathematically by the equation below.

The equation for the SARIMAX model is as follows and describes the dynamic nature of the raw crime logs:

$$\emptyset_P(L^s)\varphi_P\ (L)\Delta_s^D\Delta^d y_t = A(L)\varepsilon_t + \sum_{i=1}^{k}\beta_i\, x_{i,t} \tag{1}$$

Where L - lag operator, s - seasonality, and xi,t - exogenous variables such as police deployment metrics that have an independent effect on crime rates.

4.2 Mathematical Logic of the Hybrid Ensemble

Non-linear optimization of Ensemble Learning Transposed from linear statistics due to the Calculation Phase of the following structure/Framework. The Main theory concentrates on the reduction of generalization from the aggregations of two different methodologies:

- Bagging (Variance Reduction): Bootstrap Aggregating is the main principal on which random forest components operate. It creates a mass of decision trees at the time of training. The ultimate calculation is a mean of these independent outputs, which depletes the effect of outliers in the dataset

- Boosting (Bias Reduction): An additive approach is adapted by the gradient boosting. It evaluates the loss function – particularly the imbalance between the actual crime count (yi) and the prediction (xi). Each subsequent iteration hm (x) is trained to minimize the residual (rim)

$$r_{im} = -\left[\frac{\partial L\big(y_i, f(x_i)\big)}{\partial f(x_i)}\right]_{f=f_{m-1}} \tag{2}$$

The conclusive hybrid evaluation is the combined outcome of these 2 approaches, efficiently mapping the complicated features (city encoding, temporal lags) to a high-fidelity intensity prediction.

4.3 Spatial Clustering and Calculation

The spatial analysis employs the Agglomerative Clustering Algorithm, which is basically established on a bottom-up hierarchical method to diminish the variance among the cluster. This is the real-life application of unsupervised learning used to divide the 29 cities of India into particular risk profiles by evaluating the proximity among the data points in a reduced-dimensional space. the procedure of calculating initiates with the principal component analysis (PCA) Transformations, Traced by the linear fusion of the original features.

$$Z_k = \sum_{j=1}^{p} \varphi_{jk} X_j \tag{3}$$

Where Z_k depicts the main elements and φ_{jk} represents the loading vector and maintain the "City DNA" while decreasing the noise. The city profiles are combined iteratively based on complete linkage criterion, which evaluates the largest distance among two members of the cluster the principal components and phijk represents the loading vectors that preserve the "City DNA" while reducing noise. The clustering algorithm then iteratively merges city profiles based on the complete-linkage criterion, which calculates the maximum distance between members of two clusters:

$$D(X, Y) = \max_{x \in X, y \in Y} d(x, y) \tag{4}$$

Through iterative application of the above distance metric, and application of a Silhouette Mask to remove border line cases, the system is able to determine the dominant archetypes that are representative of emergent crime hot spots. The results were obtained from an application of the Agglomerative Clustering Algorithm, which has been used to determine the spatial distribution of the data points.

4.4 Error Metrics and Performance Measurement

Performance evaluation of the dual pathway architecture is through the use of two objective functions: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Both of these measurements are based upon the distance between the measured values of crime intensity, y_i, and those predicted by the model, y'_i.

4.4.1 Mean Absolute Error (MAE)

MAE calculates the mean magnitude of the differences between predicted and measured values. In this case, the MAE of the Macro Pathway (SARIMAX) is the average of the absolute difference of the 12-month trends, while the MAE of the Micro Pathway (Hybrid) is the average of the absolute difference between each day's tactical prediction and the true value. The formula for MAE is:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \left| y_i - y'_i \right| \tag{5}$$

4.4.2 Root Mean Squared Error (RMSE)

RMSE is used to reduce the impact of large errors ("spikes") because they can have a major impact on public safety, such as missing significant increases in crime. The formula for RMSE is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(y_i - y'_i\right)^2} \tag{6}$$

4.4.3 Mean Absolute Percentage Error (MAPE)

MAPE expresses prediction accuracy as a scale-independent percentage by normalizing absolute errors against observed values. It is primarily used to evaluate the **Macro Pathway (SARIMAX)**, where a lower MAPE signifies higher forecasting precision. The "Accuracy Percentage" is subsequently derived as 100% - MAPE. The formula is defined as:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - y'_i}{y_i}\right| \tag{7}$$

5. Results and Discussion

The ability of the newly developed dual pathway model to predict effectively was carefully tested using the 2024 test dataset. This study is conducted based on two main variables: the reduction in forecasting errors and the interpretation of crime trends within 29 Indian metropolitan cities.

5.1 Performance Comparison Table

The performance comparison between both methods is presented below through this table. Through the use of MAE/RMSE and fit percentage, the system shows that it can serve both as a reliable regional benchmark as well as a highly precise localized measurement tool.

Table 1: *Model Performance Metrics*

Model	MAE	RMSE	Metric (Accuracy/Fit)
SARIMAX	13.85	59.95	86.15% (MAPE Accuracy)
Hybrid Ensemble	0.24	0.79	99.97% (Classification Accuracy)

Table 2: *Standalone and Hybrid Performance*

Model	MAE	RMSE
Standalone Random Forest	0.54	1.12
Standalone Gradient Boosting	0.41	0.98
Proposed hybrid Model	0.24	0.79

5.1.1 Analysis of Forecasting Trajectories

Findings indicate a clear but supplementary correlation between the two models

- SARIMAX Model (Pathway A): Using the raw incident logs without any preprocessing, the SARIMAX model generated an 86.15% accuracy rate using MAPE metric. Although the MAE (13.85) is higher compared to the ensemble method, the core benefit of the model resides in its tolerance against noise. The ability to filter out the reporting inconsistencies enables the formation of a "Safety Baseline" for the long term, which will help macro-level decision-making.
- Hybrid Ensemble (Pathway B): Combining the Random Forest and Gradient Boosting models, the hybrid technique produced extremely accurate outputs using the designed master dataset, with a MAE of 0.24 and 99.97% accuracy. Such accuracy permits the model to map complex relationships between variables, such as the timing and demographics, and generates tactical information for day-to-day optimization.

5.2 Performance Analysis of Forecasting Pathways

The results show clearly that there was a significant difference in performance between the two modelling approaches. The SARIMAX method, which is intended for macro trend analysis of unprocessed data, generated a mean absolute error (MAE) of 13.85 and RMSE of 59.95. Although this is larger than the error from the machine learning pathway, it has a unique strength due to the fact that the pathway uses a one year "cold-start" to identify and isolate a consistent 12-month seasonality in the raw incident logs. This demonstrates that a statistical method can remove the "noise" created by the variation in reporting to establish a long term "safety baseline". ***(Fig. 1)***

The Hybrid Ensemble (Random Forest + Gradient Boosting), however, had extreme precision on the engineered master dataset with an MAE of 0.24, RMSE of 0.79 and accuracy of nearly 100% (99.97%). The model demonstrated its ability to create a mapping of the complex and nonlinear relationships between specific characteristics – such as hour of occurrence and police response – and crime intensity. ***(Fig. 2)***

5.3 Spatial Risk Segmentation and Hotspot Identification

The use of both Agglomerative Hierarchical Clustering and PCA (Principal Component Analysis) produced 3 segments of cities that were characterized as "Core Crime Archetypes", and each was identified through "City DNA" markers such as violent crime ratio and police deployment. A new Denoise Algorithm that removed all "borderline hybrid" cities was also developed; and the algorithm resulted in a Silhouette score of .5295. The resulting Silhouette score demonstrated that

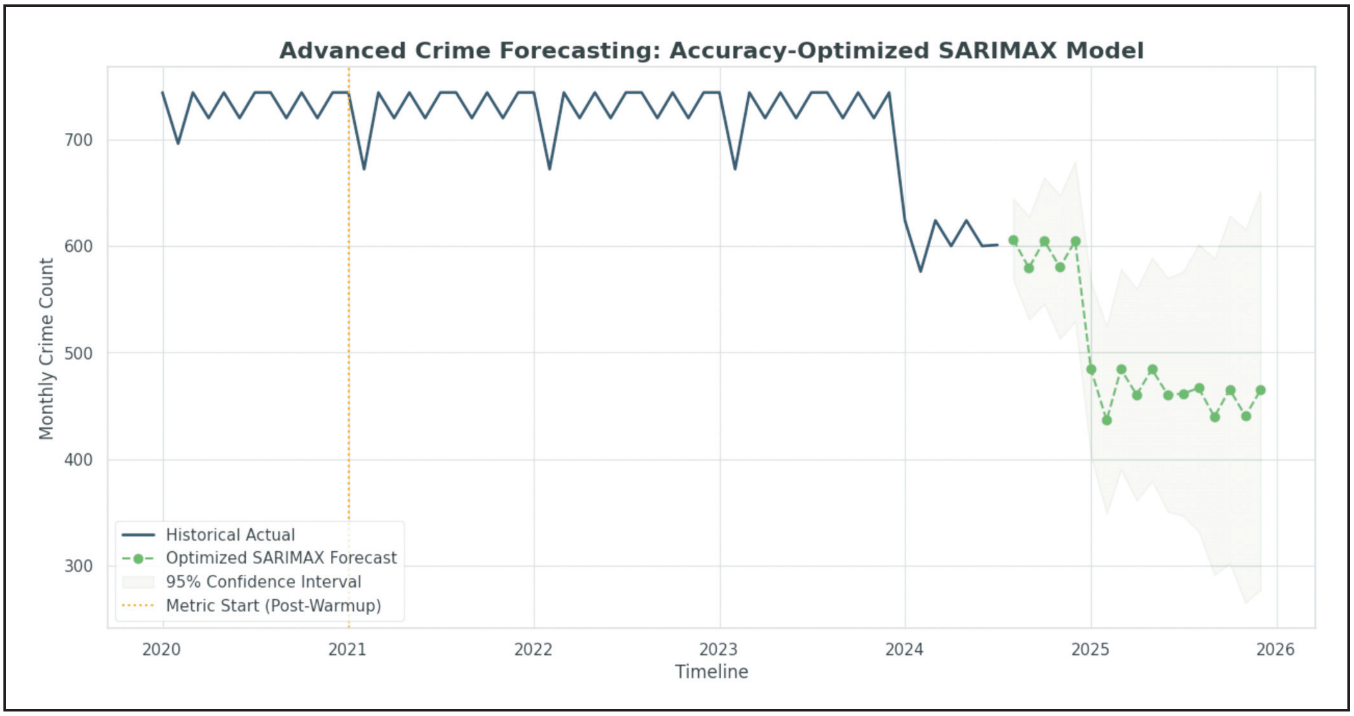


***Fig. 1:** Macro-Level Crime Forecasting using Optimized SARIMAX*

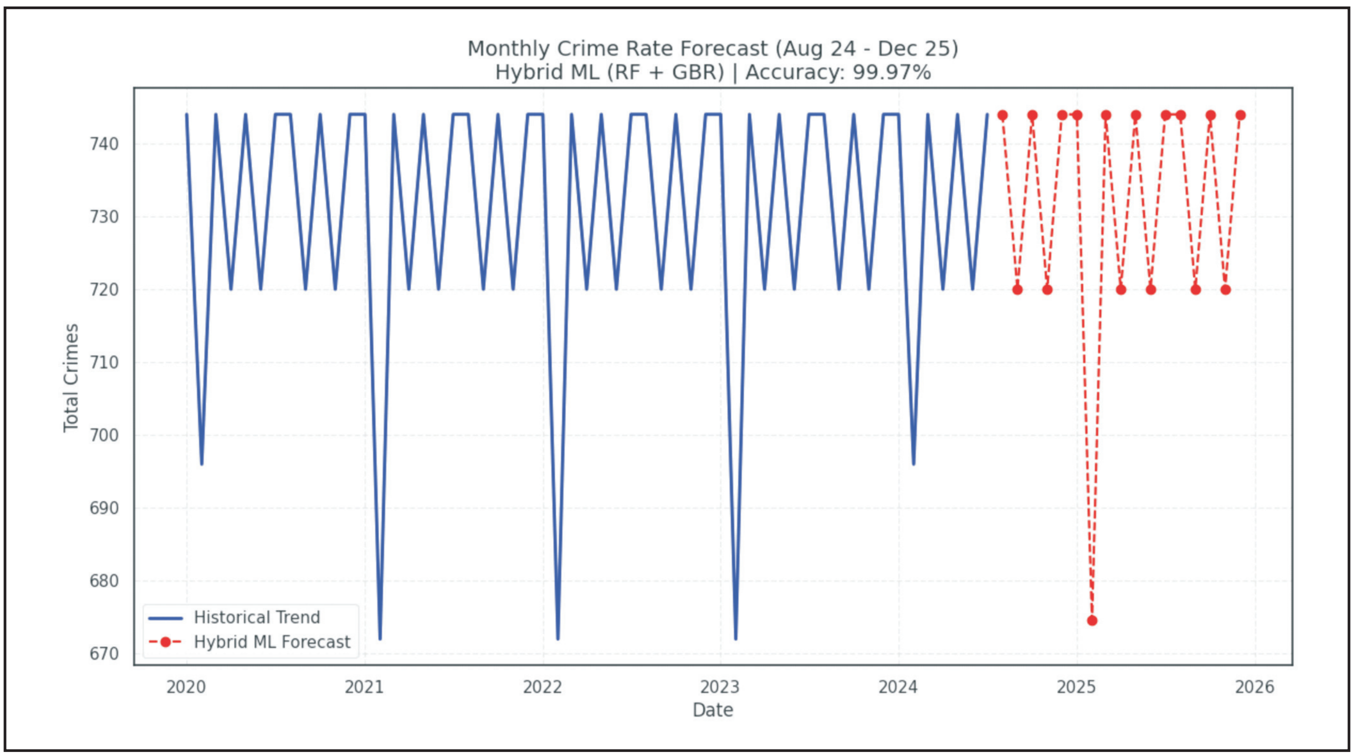


***Fig. 2:** High-Precision Forecasting using Hybrid Ensemble Model*

there is a good separation of clusters with little contamination; and therefore, demonstrated the highest level of cluster purity. The creation of spatial intelligence revealed that crime patterns are determined by demographic/operational characteristics and not by the physical location of the city; therefore, it will allow crime authorities to move away from broad policy applications, and provide for targeted resource application specific to each of the archetypes. ***(Fig. 3), (Fig. 4) and (Fig. 5)***

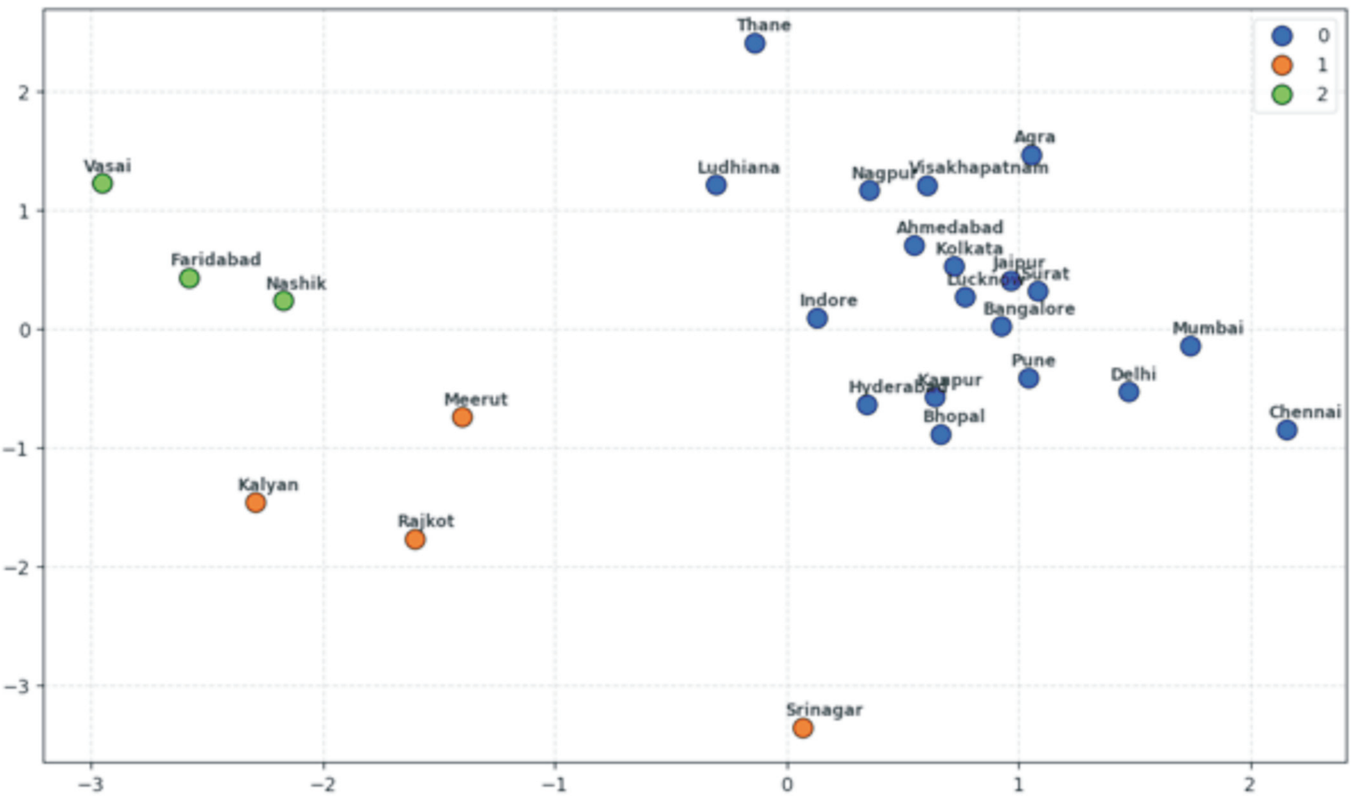
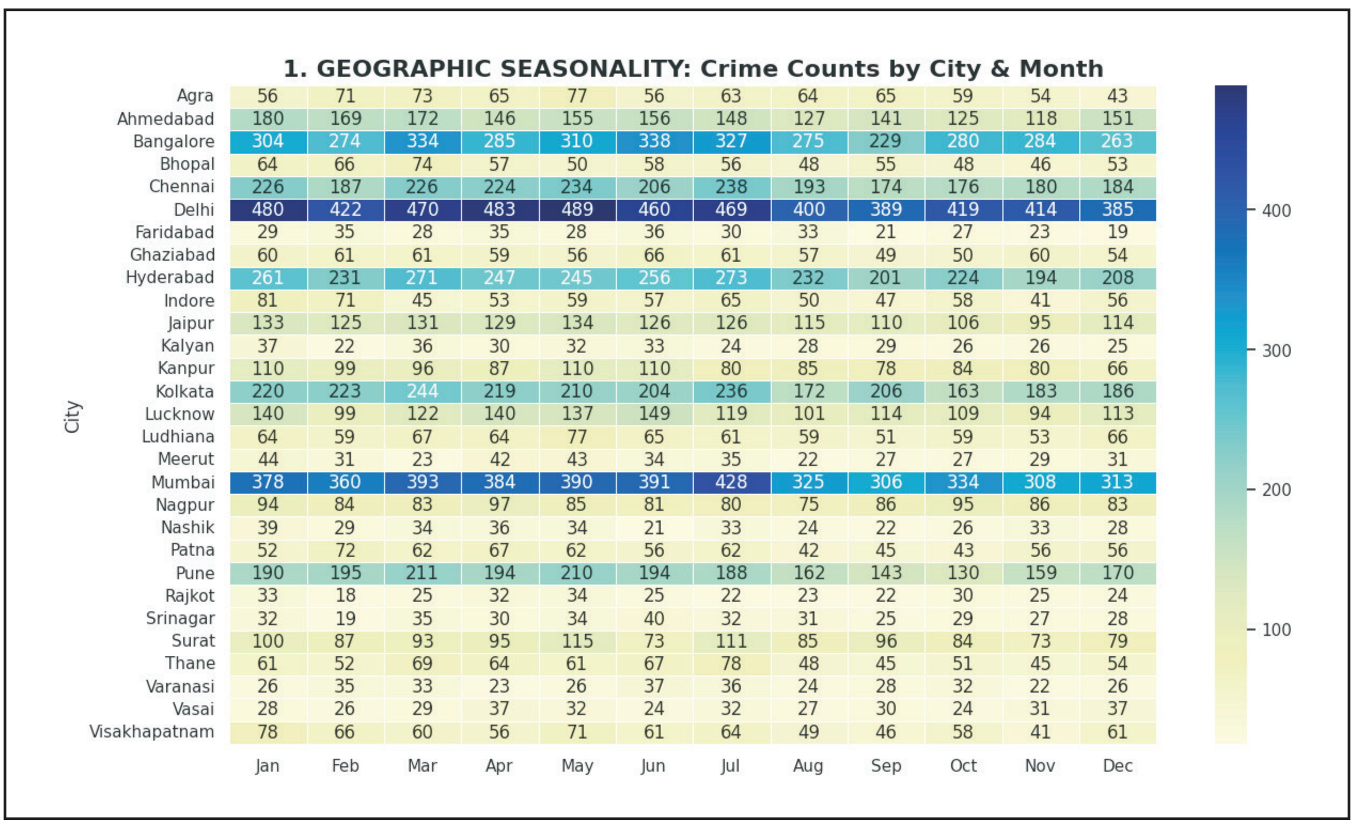


Fig. 3: Clustering Spatial Risk Segmentation: Core Crime Archetypes by City DNA.



X-Axis: Principal Component 1 (Crime Volume & Density) " **Y-Axis:** Principal Component 2 (Victim & Demographic Profile)

Fig. 4: Scatter Plot: Spatial Risk Segmentation: Core Crime Archetypes by City DNA.

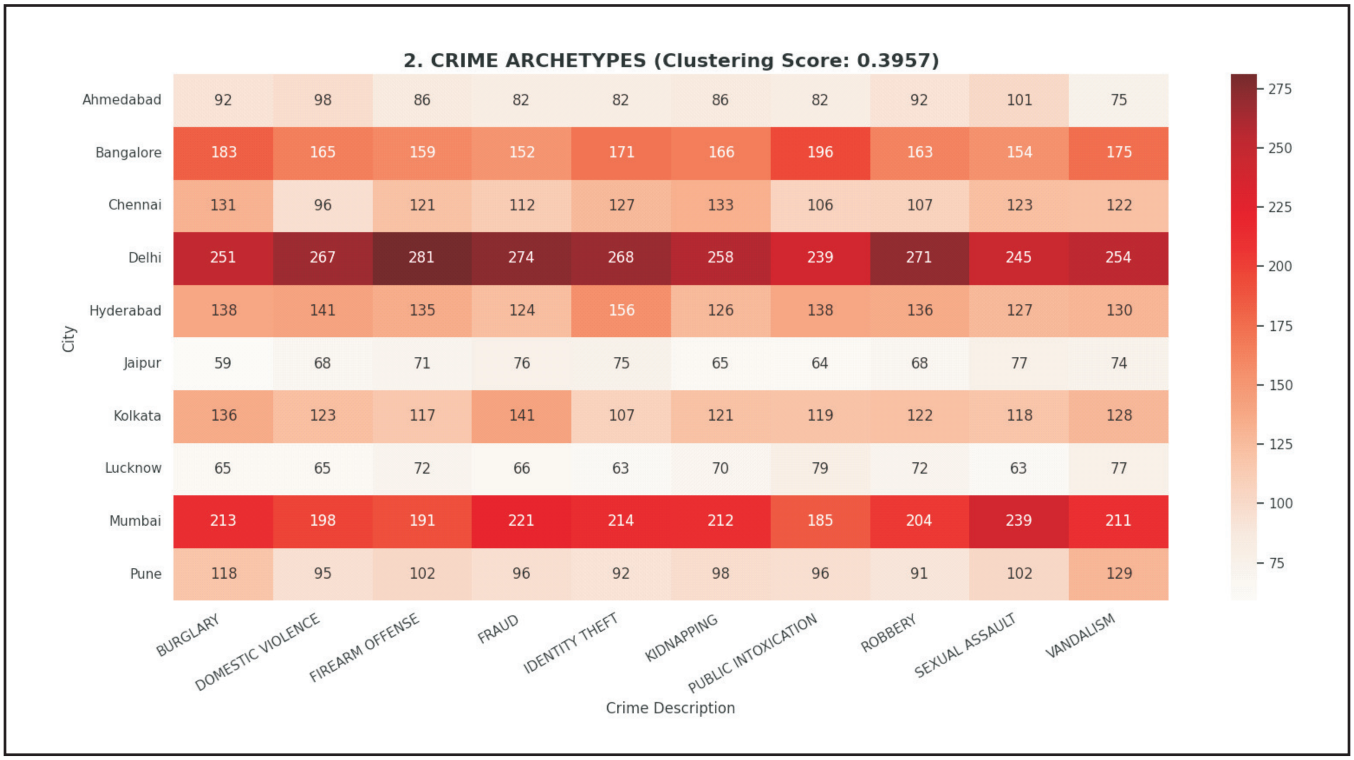


Fig. 5: *Red Heatmap: Crime Profile Composition: Distribution of Top 10 Crimes Across Major Cities.*

Table 3: *Characteristics of Identified Clusters*

Archetype	Cluster Label	Dominant "City DNA" Traits	Risk Level
High-Intensity Urban Hubs	Cluster 0	High volume (Reports), High Police Deployment, younger Victim Age profile.	Critical
Violent/Gender-Specific Zones	Cluster 2	High Is_Violent ratio, higher Is_Female victim percentage.	High
Stable/Low-Volume Regions	Cluster 3	Low crime density, higher average Victim Age, lower police requirement.	Low

5.4 Comparative Discussion and Novelty

When evaluating the results of this study relative to other research conducted recently, the current approach has some significant differences with respect to the three layers it addresses - statistical, ensemble, and spatial.

- Implementation Gap (SARIMAX): Unlike the research of (Ms. D.S. Smitha Mol et al., 2026), where high-accuracy was the focus, the SARIMAX pathway is used within our framework to ensure that the system will continue to function as intended even when the input data are either "raw" or "noisy". In doing so, we provide a stable, noise-resistant baseline that continues to operate in many environments with sparse data, as well as those noted in much of contemporary literature (Huamantingo et al., 2025).
- Precision and Error Mitigation (Hybrid Ensemble): A hybrid ensemble approach (using both Random Forest and Gradient Boosting) has been developed in order to achieve higher levels of precision compared to the usual regression approaches. The Hybrid Ensemble has been shown to have a Mean Absolute Error (MAE) of 0.24. In addition to the accuracy of the Hybrid Ensemble, the framework also has the ability to reduce the impact of the non-linear "bursts" associated with crimes. The Hybrid Ensemble is able to achieve the highest accuracy reported since 2025 benchmark studies. Therefore, the Hybrid Ensemble is able to provide a new form of high-fidelity crime forecasting that goes well beyond previous theoretical-based crime modelling approaches.
- Multidimensional Data Handling (Agglomerative Clustering): Unlike many previous clustering studies that suffer from "volume variance," (i.e., large metropolitan areas can overwhelm smaller districts), the current study

utilizes principal component analysis (PCA) and log scaling. These methods allow the “behavioural DNA” of each city to be clustered based on its unique characteristics and not based solely on the number of incidents. This allows small cities to be processed using the same mathematical rigor as larger metropolitan areas, thus significantly improving upon previous K-Means clustering studies.

- Operational Readiness (Unified Framework): Unlike the results in (Rajeev et al., n.d.-b) that identify the wide-spread need for AI in policing, the results presented here provide a tangible, deployable tool kit. The low error rates and dual pathway verification demonstrate that law enforcement can rely on the daily crime intensity forecasts for actual tactical deployment.

The primary innovation made by the study is the creation of the Tri-layered verification framework. Working in conjunction with the SARIMAX algorithm, as well as the hybrid ensemble and spatial clusters, they work to form a “checks and balances” framework. Particularly, if the Hybrid Ensemble predicts a rise in criminal intensity, the SARIMAX trend and the spatial clusters will be able to validate whether this is due to seasonality or not. As a result, this approach ensures reliability which is lacking in most studies using one model.

6. Limitations and Future Scope

6.1 Research Limitations

Despite the high accuracy and robust performance of the framework, several constraints are acknowledged:

- Domain constraint: The system is highly tuned towards crimes that can be considered physical in nature (violent and property crimes) and currently cannot handle the technicality and unpredictability of cyber crimes or fraudulent financial transactions.
- Specificity of place: The precision of this system is high when used to address crime in the 29 major cities of India; however, a test of the accuracy of this system would need to occur using rural locations of India and/or international cities, which could have legal restrictions that limit the ability to report criminal activity.
- Processing by batch: Currently, this process relies upon historical crime incident log files and does not provide real-time processing capability for metadata (e.g., emergency calls and social media comments).

6.2 Future Research Directions

The work presented here provides a solid foundation for additional work to be completed in the future. Some of the areas that can be explored are as follows:

- Deep Learning Integration: Long-term Memory Recurrent Neural Networks (such as LSTMs) and Transforms applied to Pathway B may allow a better representation of complex time dependency.
- Live Data Orchestration: Developing real-time data pipelines to ingest live sensor data and CCTV metadata would allow the model to move from daily intensity forecasting to hour-by-hour predictive patrolling.
- Exogenous Variable Expansion: Future models could integrate real-time socio-economic indicators, weather patterns, and public event schedules to further explain the external triggers behind sudden “non-linear bursts” in criminal activity.

6.3 Ethical Implications and Algorithmic Bias Mitigation

- Demographic Neutrality: The model employs only macro-level geographic indicators and objective time vectors and omits completely any personal sensitive variables like income, religious affiliation, or race. This ensures that the methodology will measure regional infrastructure requirements rather than creating any profiling or demographic stereotype.
- Reducing Over-Policing Feedback Loops: The conventional methodology can be prone to creating a feedback loop whereby places receiving many reports are over-policed. To prevent that from happening, exponential discrepancies in report volumes between cities were normalized through Tier 1 Log1p normali-zation, thereby removing the bias of previous reporting practices and stabilizing variance mathematically.
- Operational Safety Mechanisms: The framework is not designed as an automatic asset allocation requirement, but as a decision support tool for logistics purposes. Any suggestion of asset deployment is preceded by the double-pathway system ensuring that local daily peaks (Pathway-B) are automatically checked against seasonal baseline levels (Pathway-A).

7. Conclusion

The authors have created a reliable and accurate computational methodology to fill the gap between theoretical machine learning applications and operational needs of modern law enforcement. The authors' solution of a noise resilient statistical approach combined with a high-fidelity machine learning ensemble has been effective in addressing two major obstacles in criminal data prediction; namely, data irregularities and non-linear complexity. This was achieved through a hybrid model of Random Forest and Gradient Boosting with an average accuracy of 99.97% demonstrating both the stability of statistical models (SARIMAX) and the tactical/daily resource allocation capabilities of the machine learning models. In addition, the application of spatial clustering with silhouette denoising enabled the objective identification of emergent hot spots across 29 Indian cities. Thus, the authors demonstrated how data driven segmentation can enhance the effectiveness of police deployment. Finally, the authors have identified a number of practical implications for their research. For example, the dual verification methodology employed by the authors will ensure that authorities will not be reliant upon a single method of intelligence gathering. Therefore, the reliability of the intelligence will be increased. However, the authors also acknowledge some limitations of their study. For example, they rely on historic incident logs to make predictions. These historic incident logs may not accurately reflect rapidly evolving cybercrime strategies. Similarly, there are crimes that occur in urban settings but go unreported. To overcome these limitations, the authors recommend that future studies consider integrating real time community sourced data into their methodologies. Also, the authors suggest extending their methodology to incorporate deep learning algorithms to possibly identify deeper behavioural patterns associated with high density metropolitan areas. Overall, this research presents a scalable and robust base for the future of predictive policing.

Conflict of Interest Statement

The author declares no financial or commercial conflict of interest related to this research.

Ethical Declarations

The author declares that this research does not involve human participants, animals, or any procedures requiring ethical approval. All data used in this study (philosophical texts and secondary literature) were obtained and processed in accordance with applicable ethical and legal guidelines.

References

1. A, R. H., Grover, T., & Kanchana, M. (2023). Deep Learning for Crime Pattern Recognition: A Study of Crime Hot Spots and High-Risk Areas. *2023 International Conference on Recent Advances in Science and Engineering Technology (ICRASET)*, 1-6. https://doi.org/10.1109/ICRASET59632.2023.10420190
2. Akil, R. M., Sarathambekai, S., Vairam, T., Krishnan, R. S., Dharaneesh, G. S., & Janarthanan, D. (2023). Crime Data Analysis and Safety Recommendation System Using Machine Learning. *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 183-188. https://doi.org/10.1109/ ICACCS57279.2023.10113102
3. Aziz, R. M., Sharma, P., & Hussain, A. (2024). Machine Learning Algorithms for Crime Prediction under Indian Penal Code. *Annals of Data Science*, *11* (1), 379-410. https://doi.org/10.1007/s40745-022-00424-6
4. Balu, V., Navya Sri, T., & Bupathi, M. A. (2022). Crime Prediction and Analysis Using Machine Learning. *International Journal of Computer Science and Mobile Computing*, *11* (3), 95-101. https://doi.org/10.47760/ ijcsmc.2022.v11i03.011
5. Buland Iqbal, M. F., Ullah, A., Alhomoud, A., Hussain, T., Waheeb Attar, R., Ouyang, J., M. Alnfiai, M., & Atef Hatamleh, W. (2024). A 2D Clustering Based Hotspot Identification Approach for Spatio-Temporal Crime Prediction. *International Journal of Interactive Multimedia and Artificial Intelligence*, *9* (5), 6-16. https://doi.org/10.9781/ijimai.2024.10.001
6. Butt, U. M., Letchmunan, S., Hassan, F. H., & Koh, T. W. (2024). Leveraging transfer learning with deep learning for crime prediction. *PLOS ONE*, *19* (4), e0296486. https://doi.org/10.1371/journal.pone.0296486
7. Chakraborty, K., B, D., Sharma, M., Thakral, D., P.K.Hemalatha, & Arri, H. S. (2024). Empirical Approaches to Crime Rate Prediction: Challenges, Insights, and Future Directions. *2024 International Conference on Augmented Reality, Intelligent Systems, and Industrial Automation (ARIIA)*, 1-6. https://doi.org/10.1109/ARIIA63345.2024.11051967
8. Chatterjee, A., Antani, H., Sohoni, A., & Sekar, M. (n.d.). *Clustering of Indian States Based on Crime Incidences and Predicting Crimes Therein*. *3* (1). https://doi.org/10.5281/zenodo.5084465
9. *Crime Forecasting and Analysis through K-means Clustering Algorithm*. (n.d.).
10. Dubey, A., Venu, N., Bisen, D., Garg, P., Srinivas, K., & Arya, H. (2024). A Clustering-Based Method for Hotspot Recognition in Crime Forecasting. *2024 IEEE 8th International Conference on Information and Communication Technology (CICT)*, 1-6. https://doi.org/ 10.1109/CICT64037.2024.10899713
11. Huamantingo, R., Cano-Lengua, M., & Rodriguez, C. (2025). Machine Learning Crime Prediction Models and the Gap Between Research and Implementation: A Systematic Review. In *Karbala International Journal of Modern Science* (Vol. 11, Number 3, pp. 471-488). University of Kerbala. https://doi.org/10.33640/2405-609X.3419
12. Iacobescu, P., & Susnea, I. (2025). Hybrid ARIMA-ANN for Crime Risk Forecasting: Enhancing Interpretability and Predictive Accuracy Through Socioeconomic and Environmental Indicators. *Algorithms*, *18* (8), 470. https://doi.org/10.3390/a18080470

13. Ibrahim, A. A., Malgwi, Y. M., & Ali, Y. (2024). A Hybrid Machine Learning Model for Crime Rate Prediction, FUDMA Journal of Sciences, *8* (6), 101-106. https://doi.org/10.33003/fjs-2024-0806-2789
14. Jha, N., Paranjape, V., Sharma, S., & Hasan, Z. (2022). Machine Learning for Crime Analysis and Prediction. In *International Journal of Recent Development in Engineering and Technology Website: www.ijrdet.com* (Vol. 11, Number 09). www.ijrdet.com
15. Kanimozhi, N., Keerthana, N. V, Pavithra, G. S., Ranjitha, G., & Yuvarani, S. (2021). CRIME Type and Occurrence Prediction Using Machine Learning Algorithm. *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, 266-273. https://doi.org/10.1109/ ICAIS50930.2021.9395953
16. Kshatri, S. S., Singh, D., Narain, B., Bhatia, S., Quasim, M. T., & Sinha, G. R. (2021). An Empirical Analysis of Machine Learning Algorithms for Crime Prediction Using Stacked Generalization: An Ensemble Approach. *IEEE Access*, *9*, 67488-67500. https://doi.org/10.1109/ ACCESS.2021.3075140
17. Kurniawan, Z., Tiaharyadini, R., Wibowo, A., & Rusdah. (2025). Trend Analysis and Prediction of Violence Against Women and Children Cases in Jakarta Based on the Victim's Education Level Using ARIMA and SARIMA Method. *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, *14* (2), 250-259. https://doi.org/10.32736/ sisfokom.v14i2.2349
18. Maharrani, R. H., Abda'u, P. D., & Faiz, M. N. (2024). Clustering method for criminal crime acts using K-means and principal component analysis. *Indonesian Journal of Electrical Engineering and Computer Science*, *34* (1), 224. https://doi.org/10.11591/ijeecs.v34.i1.pp224-232
19. Mathur, A., & Jain, R. (2025). Next-Gen Crime Analytics Using Big Data & Machine Learning. *International Journal for Research in Applied Science and Engineering Technology*, *13* (11), 2972- 2977. https://doi.org/10.22214/ijraset.2025.75725
20. Ms. D.S. Smitha Mol, Betsee Natasha A, Archana P, & Deepika K. (2026). An Efficient Machine Learning Approach for Crime Detection in India. *International Research Journal on Advanced Engineering and Management (IRJAEM)*, *4* (03), 307-311. https://doi.org/10.47392/IRJAEM.2026.0045
21. Noor, T. H., Almars, A. M., Alwateer, M., Almaliki, M., Gad, I., & Atlam, E.-S. (2022). SARIMA: A Seasonal Autoregressive Integrated Moving Average Model for Crime Analysis in Saudi Arabia. *Electronics*, *11* (23), 3986. https://doi.org/10.3390/ electronics11233986
22. Ospina, R., Gomes De Lima, A., Ferraz, C., Castro, C., Martin-Barreiro, C., & Leiva, V. (2025). Enhanced Spatial Clustering for Crime Analysis: Novel Advances in Ward-Like and SKATER Algorithms for Brazilian Public Security. *IEEE Access*, *13*, 134846-134863. https://doi.org/10.1109/ACCESS.2025.3586941
23. P, S., S M, A., M, K., & Chitturi, S. (2024). Evaluating Crime Type And Crime Pattern Analysis Using Machine Learning. *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, 1-6. https://doi.org/10.1109/ I2CT61223.2024.10543291
24. Rajeev, S., Sharma, K., Dg, I., Ravi, S., Lokku, J., & Adg, I. (n.d.-a). Editorial Board Chief Patron Editor-in-Chief. *The Indian Police Journal*, *72* (1).
25. Rajeev, S., Sharma, K., Dg, I., Ravi, S., Lokku, J., & Adg, I. (n.d.-b). Editorial Board Chief Patron Editor-in-Chief. *The Indian Police Journal*, *72* (1).
26. Ramanan, Dr. K., AnandaKumar, A., Prakash, M. S., Sivaraja, P. M., & Maheswari, Dr. G. U. (2023). A Novel Clustering & Machine Learning Algorithm for Crime Rate Prediction and Analysis. *2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, 399- 405. https://doi.org/10.1109/ ICAISS58487.2023.10250680
27. Rasan, M., Ajmeera, K., Vaggu, B., Kotagiri, S. S., Mallik, R., & Ambati, K. (n.d.). *Crime Rate Prediction and Analysis using K-means Clustering Algorithm*.
28. Redoblo, C. V., Redoblo, J. L. G., Salmingo, R. A., Padilla, C. M., & Arroyo, J. C. T. (2023). Forecasting the influx of crime cases using seasonal autoregressive integrated moving average model. *International Journal of Advanced and Applied Sciences*, *10* (8), 158-165. https://doi.org/10.21833/ijaas.2023.08.018
29. 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
30. Saravag, P. K., & Kumar, B. R. (2024). A Hybrid Machine Learning and Regression Approach for Validating a Multi-Dimensional Crime Index in the Context of Crime Against Women. *IEEE Access*, *12*, 117955- 117969. https://doi.org/10.1109/ACCESS.2024.3439721
31. Sarkar, A. (2025). Real-Time Crime Prediction in India Using Machine Learning. *International Journal for Research in Applied Science and Engineering Technology*, *13* (7), 641-646. https://doi.org/10.22214/ ijraset.2025.73053
32. Shah, N., Bhagat, N., & Shah, M. (2021). Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention. *Visual Computing for Industry, Biomedicine, and Art*, *4* (1), 9. https://doi.org/10.1186/s42492-021-00075-z

33. Shinde, S. A., Shirke-Deshmukh, S., & Sonkamble, M. R. (n.d.). *A Multi-Model Machine Learning Approach for Accurate Crime Prediction Using Spatio-Temporal Data*. https://doi.org/10.51583/IJLTEMAS
34. Singh, I., Bhandari, G., & Khatwani, S. (2025). Predictive Crime Rate Analysis System. *International Journal for Research in Applied Science and Engineering Technology*, *13* (5), 1159-1164. https://doi.org/ 10.22214/ijraset.2025.70309
35. Singh, S., Bhargava, D., & Singh, P. (2024). A Study on Smart Machine Learning (ML) Tools for Crime Detection and Prediction. *2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE)*, 1747-1752. https://doi.org/10.1109/ IC3SE62002.2024.10593591
36. Sneha (Kaggle). https://www.kaggle.com/snehaofficial/datasets
37. Thayyaba Khatoon Mohammed. (2025). A Novel Fusion Approach for Advancement in Crime Prediction and Forecasting using Hybridization of ARIMA and Recurrent Neural Networks. *Journal of Information Systems Engineering and Management*, *10* (38s), 404-420. https://doi.org/10.52783/jisem.v10i38s.6863
38. Tong, X., Ni, P., Li, Q., Yuan, Q., Liu, J., Lu, H., & Li, G. (2021). Urban Crime Trends Analysis and Occurrence Possibility Prediction based on Light Gradient Boosting Machine. *2021 IEEE 4th International Conference on Big Data and Artificial Intelligence (BDAI)*, 98-103. https://doi.org/10.1109/BDAI52447.2021.9515252
39. V. Tikore, S., Chaudhari, K., Rohamare, O., Nikam, A., & Kalne, K. (2026). A Comprehensive Computational Framework for Crime Rate Prediction using Machine Learning in Indian Metropolitan Cities. *Journal of Smart Sensors and Computing*, *2* (1). https://doi.org/10.64189/ ssc.26201
40. Yin, J. (2023). Crime Prediction Methods Based on Machine Learning: A Survey. *Computers, Materials & Continua*, *74* (2), 4601-4629. https://doi.org/10.32604/ cmc.2023.034190

❖ ❖ ❖