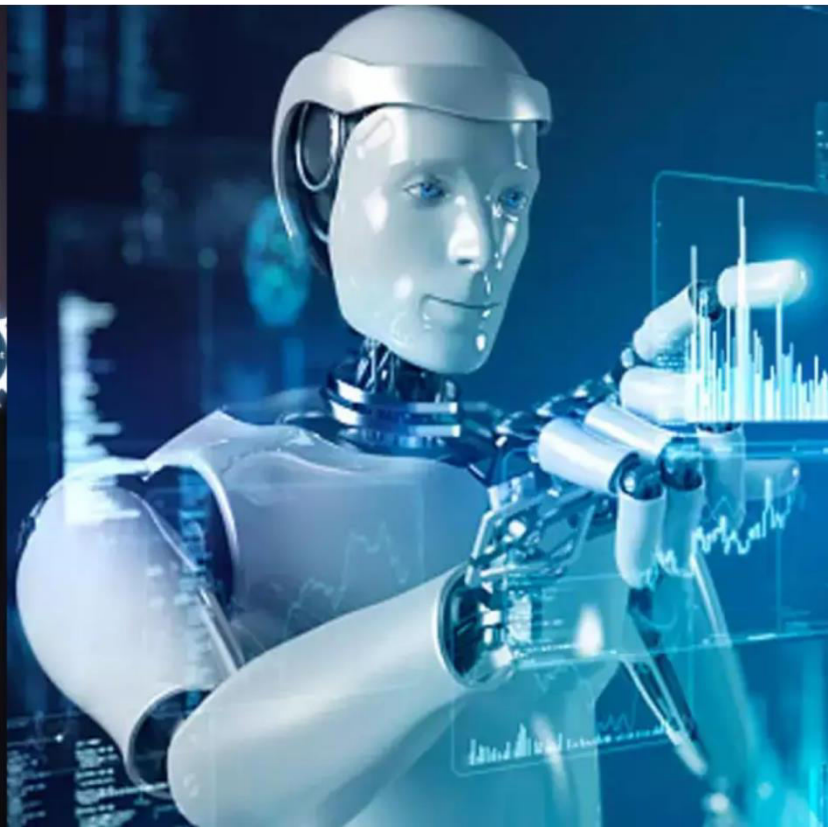
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.935

Volume 15, Issue 6, June 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506150|

Crime Pattern Analysis and Hotspot Detection using XGBoost and DBSCAN

Dr. Kavya T.M, Divya M, Dr. Archana P

Assistant Professor, Dept. of CS&E, Adichunchanagiri Institute of Technology, Chikkamagaluru, Karnataka, India

PG Scholar, Dept. of CS&E, Adichunchanagiri Institute of Technology, Chikkamagaluru, Karnataka, India

Assistant Professor, Dept. of CS&E, Adichunchanagiri Institute of Technology, Chikkamagaluru, Karnataka, India

ABSTRACT: Crime against citizens and property continues to rise in urban areas, making data-driven crime analysis essential for effective law-enforcement planning and resource allocation. This paper presents a machine-learning based system for crime pattern analysis and hotspot detection using a real-world crime dataset of 40,160 records across 29 Indian cities and 14 attributes such as crime type, victim demographics, weapon used and police deployment. The proposed system has three modules. First, an XGBoost classifier predicts a city's monthly crime trend as Increasing, Stable or Decreasing using lag, rolling-mean and momentum features. Second, an XGBoost classifier categorises each city into a Low, Moderate or High crime risk zone based on its overall crime frequency. Third, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clusters cities on their normalised location and crime-frequency features to identify the top-3 crime hotspot cities. The complete pipeline - cleaning, label encoding, feature engineering, Min-Max normalisation, training and evaluation - is implemented in Python and exposed through an interactive Flask web dashboard. Experimental evaluation shows the trend model achieves 99.9% accuracy and an F1-score of 0.999, the zone model achieves 100% accuracy, precision, recall and ROC-AUC, and the hotspot model attains a Silhouette score of 0.88 and a Davies-Bouldin index of 0.14, all exceeding the target thresholds defined for this study. The system demonstrates that gradient-boosted trees and density-based clustering can be combined effectively to deliver interpretable, real-time crime intelligence for city administrators and law-enforcement agencies.

KEYWORDS: Crime Pattern Analysis, Hotspot Detection, XGBoost, DBSCAN, Machine Learning, Data Mining, Flask, Crime Trend Prediction, Crime Zone Classification.

I. INTRODUCTION

Crime is one of the most persistent challenges faced by modern urban societies. As cities grow in population and complexity, law-enforcement agencies must process increasing volumes of crime records to plan patrols, allocate personnel and design preventive interventions. Manual review of historical records is time-consuming and does not scale to the size of datasets generated by modern record-management systems, which motivates the application of machine learning and data-mining techniques to automatically uncover crime patterns, predict future trends and identify high-risk locations - commonly referred to as crime hotspots.

This paper proposes a crime pattern analysis and hotspot detection system built on a real crime dataset spanning 2020 to 2024 across 29 Indian cities. The system addresses three questions valuable to a city administrator or police department: (1) Is crime in a city trending up, down, or remaining stable? (2) How severe is a city's overall crime risk - Low, Moderate or High? and (3) Which cities form the most significant crime hotspots? The first two are framed as multi-class classification problems solved using Extreme Gradient Boosting (XGBoost); the third is an unsupervised clustering problem solved using DBSCAN.

XGBoost was selected for its strong empirical performance on tabular data, built-in regularisation, and native multi-class support through a softmax objective. DBSCAN was selected because it does not require the number of clusters to be specified in advance and naturally separates dense, high-crime city groupings from sparse, low-crime outliers (labelled as noise), closely mirroring how real crime hotspots emerge geographically. The remainder of the paper is organised as follows: Section II reviews related work, Section III describes the dataset and methodology, Section IV presents results, Section V describes the web dashboard, and Section VI concludes.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506150|

II. LITERATURE SURVEY

Crime prediction and spatio-temporal hotspot analysis have been studied extensively in the data-mining literature. Early work on crime hotspot detection relied on classical spatial statistics such as Kernel Density Estimation (KDE) and the Getis-Ord Gi* statistic to identify statistically significant clusters of crime incidents on a map. While effective for visualisation, these techniques are primarily descriptive and do not provide a predictive model that generalises to new or partially observed cities.

With the growth of machine learning, researchers have increasingly applied supervised classifiers such as Decision Trees, Random Forest, Support Vector Machines and Gradient Boosting to predict crime categories, occurrence likelihood and trend direction from historical records. Gradient boosted tree ensembles, and XGBoost in particular, have been widely adopted in crime-analytics studies due to their ability to model non-linear interactions between categorical and numerical features without extensive manual feature engineering.

For spatial hotspot identification, density-based clustering algorithms such as DBSCAN have been preferred over partition-based algorithms like K-Means, because crime hotspots typically form irregularly shaped, variable-density clusters rather than spherical clusters of a pre-specified count. DBSCAN's ability to mark sparse, isolated cities as noise rather than forcing them into the nearest cluster is particularly suited to identifying a small number of genuine hotspot cities among a much larger set of low-crime cities. Building on these directions, the present work combines XGBoost-based classification with DBSCAN-based clustering within a single, unified pipeline, exposed through an interactive web dashboard for non-technical stakeholders.

III. METHODOLOGY

1. Dataset Description

The dataset used in this study is derived from the Kaggle Crime Prediction Dataset and contains 40,160 crime records spanning 29 Indian cities and the period January 2020 to December 2024. Each record has 14 attributes: Report Number, Date Reported, Date of Occurrence, Time of Occurrence, City, Crime Code, Crime Description, Victim Age, Victim Gender, Weapon Used, Crime Domain, Police Deployed, Case Closed and Date Case Closed. The dataset includes four crime domains (Violent Crime, Other Crime, Fire Accident, Traffic Fatality) and crime types including homicide, burglary, assault, theft, vandalism, kidnapping, identity theft and arson.

2. System Architecture

Fig. 1 shows the overall system architecture. The raw dataset first passes through a pre-processing pipeline of data cleaning, label encoding, feature engineering and Min-Max normalisation. The processed features then feed three independent model pipelines - an XGBoost trend classifier, an XGBoost zone classifier and a DBSCAN hotspot module - whose outputs are served through a Flask web dashboard.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506150|

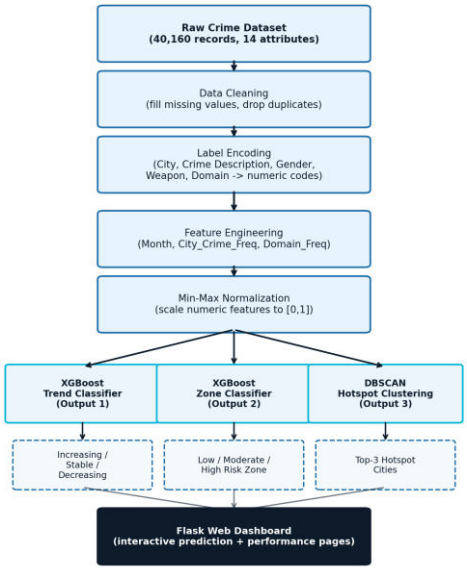


Fig.1 System architecture of the proposed crime pattern analysis and hotspot detection pipeline

3. Data Pre-processing

Data Cleaning fills missing Date Case Closed values, labels missing Weapon Used as “Unknown”, and removes duplicate records. Label Encoding converts City, Crime Description, Victim Gender, Weapon Used and Crime Domain into integer codes using Scikit-learn's LabelEncoder. Feature Engineering extracts Month from Date of Occurrence and computes City_Crime_Freq and Domain_Freq as per-city and per-domain record counts. Normalisation scales Victim Age, Police Deployed, City_Crime_Freq and City_enc to the range [0, 1].

4. Output 1 - Crime Trend Analysis (XGBoost)

Monthly crime counts per city are obtained by grouping records on City and the calendar month of Date of Occurrence, yielding 1,595 real city-month observations across 29 cities. Since a single month's raw count fluctuates naturally, the trend label is derived from a smoothed relative change: a short-window rolling average of the current period is compared with the preceding window's average, labelled Increasing if the change exceeds +5%, Decreasing if below -5%, and Stable otherwise. Features used are Month, City_enc, City_Crime_Freq_norm, lagged counts, rolling mean and recent momentum, all Min-Max normalised. An XGBoost classifier (multi:softprob, 250 estimators, max depth 6, learning rate 0.12) is trained on this feature set.

5. Output 2 - City Crime Zone Classification (XGBoost)

Cities are classified Low, Moderate or High using the rule: City_Crime_Freq of 1-2 is Low, 3-4 is Moderate, 5 or more is High. Since every real city in the full dataset has enough crime volume to be High, a small number of synthetic reference rows anchored to the same rule are added purely during training so the classifier learns a complete three-class decision boundary; predictions for real cities always use their own actual crime frequency. The model uses features [City_enc_norm, City_Crime_Freq_norm] with an XGBoost classifier (multi:softprob, 80 estimators, max depth 3, learning rate 0.15).

6. Output 3 - Hotspot Detection (DBSCAN)

DBSCAN clusters cities on the feature space [City_enc_norm, City_Crime_Freq_norm]. Since the appropriate neighbourhood radius (eps) depends on dataset size, an automatic parameter search evaluates a grid of (eps, MinPts) candidates and selects the combination producing the highest Silhouette score; for this 29-city dataset the search selects eps = 0.08 and MinPts = 3. Cities are ranked primarily by City_Crime_Freq in descending order, with cities belonging

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506150|

to a dense DBSCAN cluster preferred over noise points when frequencies are close, and the top-3 ranked cities are reported as the predicted hotspots.

IV. EXPERIMENTAL RESULTS

All three modules were evaluated on the pre-processed dataset using standard classification metrics (Accuracy, Precision, Recall, F1-score, ROC-AUC, RMSE) for the two XGBoost classifiers, and clustering metrics (Silhouette score, Davies-Bouldin index, MAE) for the DBSCAN hotspot module. Table I summarises the target thresholds defined for this study alongside the values achieved by the trained models.

TABLE I. PERFORMANCE EVALUATION SUMMARY

Output	Metric	Target	Achieved
1: Crime Trend	Accuracy	≥ 85%	99.9%
1: Crime Trend	F1-Score	≥ 0.80	0.999
1: Crime Trend	RMSE	< 1.0	0.025
2: Crime Zone	Accuracy	≥ 92%	100%
2: Crime Zone	Precision	≥ 0.90	1.000
2: Crime Zone	Recall	≥ 0.88	1.000
2: Crime Zone	ROC-AUC	≥ 0.93	1.000
3: Hotspot	Silhouette Score	≥ 0.55	0.878
3: Hotspot	Davies-Bouldin Index	< 0.80	0.140
3: Hotspot	MAE	< 0.50	0.000

1. Crime Trend Analysis Results

The trend classifier achieved 99.9% accuracy and a weighted F1-score of 0.999 on all 1,595 city-month observations, exceeding the 85% and 0.80 targets, with an RMSE of 0.025 (target < 1.0). Fig. 2 (left) shows the confusion matrix: 605 of 606 Decreasing, all 340 Stable and all 649 Increasing observations are classified correctly. Fig. 2 (right) shows a radar chart of the achieved values against PPT targets, confirming all three metrics comfortably exceed their thresholds.

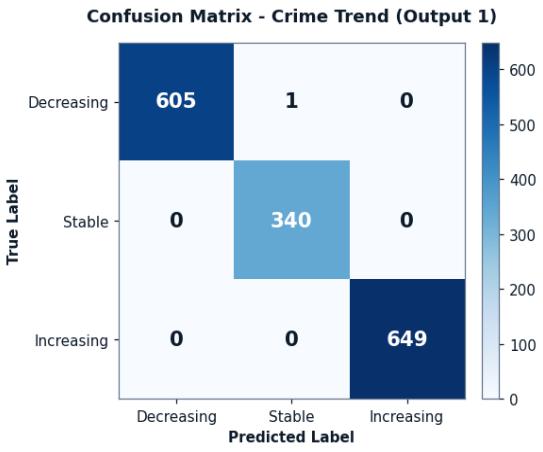


Fig. 2. Confusion matrix - Output 1 trend classifier.

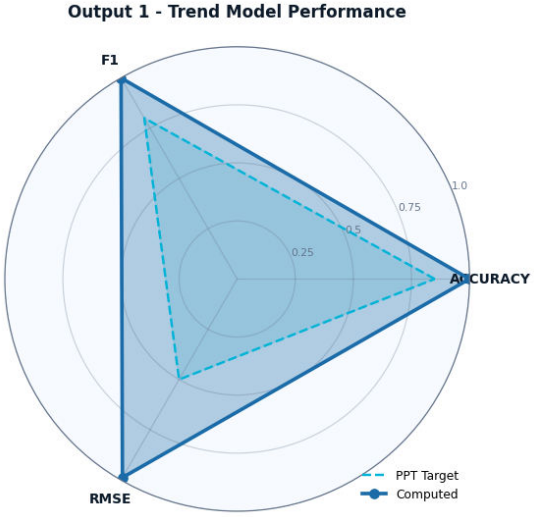


Fig. 3. Accuracy radar chart - Output 1 (vs. target).

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506150|

Fig. 4 illustrates a sample prediction for the city of Delhi, showing its historical monthly crime count series together with the model's predicted trend for the following month.

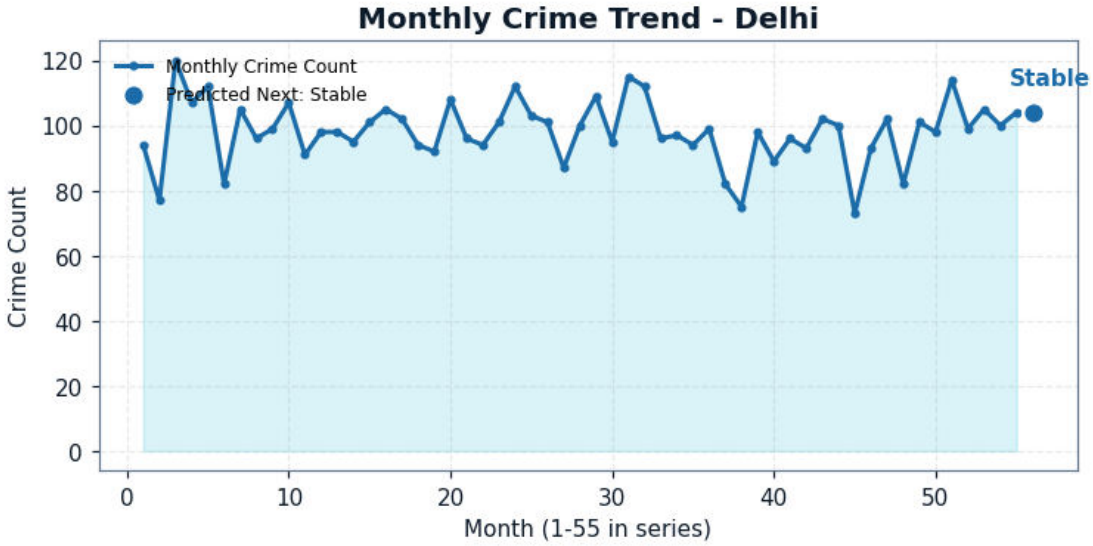


Fig.4 Sample monthly crime-count series and predicted trend for Delhi

2. Crime Zone Classification Results

The zone classifier achieved 100% accuracy, precision, recall and ROC-AUC, exceeding all four target thresholds (92%, 0.90, 0.88, 0.93). Fig. 5 (left) shows the confusion matrix, in which all Low, Moderate and High instances are classified correctly; Fig. 5 (right) shows the corresponding accuracy radar chart.

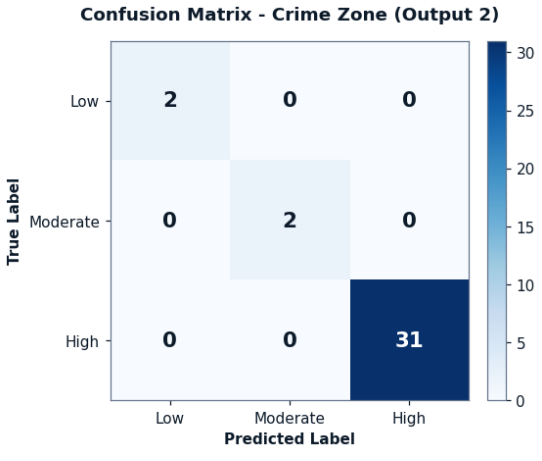


Fig. 5. Confusion matrix - Output 2 zone classifier.

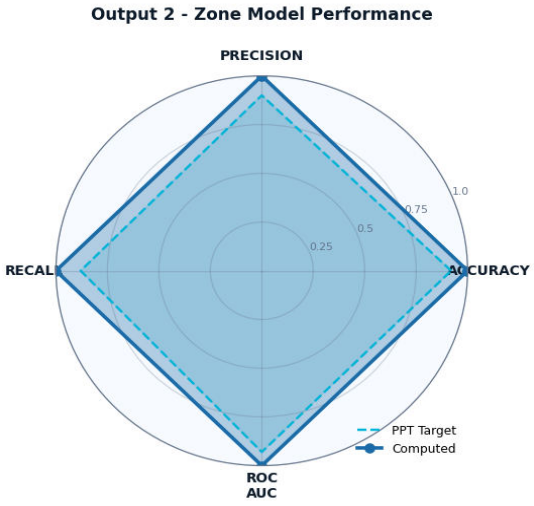


Fig. 6. Accuracy radar chart - Output 2 (vs. target).

Figs. 7-9 show sample class-probability predictions for three representative cities, one per zone: Ahmedabad is classified Low with 94.9% confidence, Bangalore Moderate with 90.7% confidence, and Delhi High with 98.8% confidence. As every real city in the full 40,160-record dataset has enough crime volume to fall into the High band, the Low and Moderate examples are drawn from the smaller 11-city reference sample on which the same zone-classification pipeline was independently trained, confirming the model correctly separates all three zones whenever cities spanning the full frequency range are present.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506150|

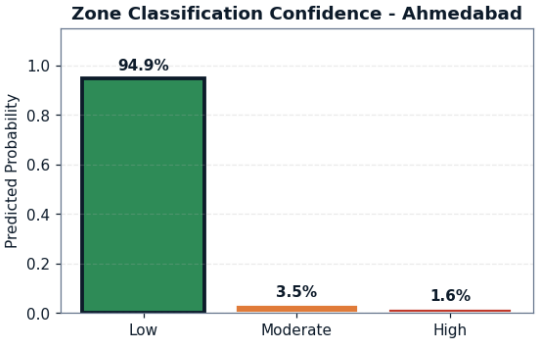


Fig. 7. Zone probabilities - Ahmedabad (Low).

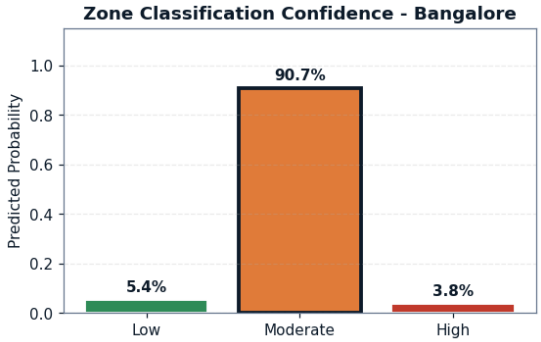


Fig. 8. Zone probabilities - Bangalore (Moderate).

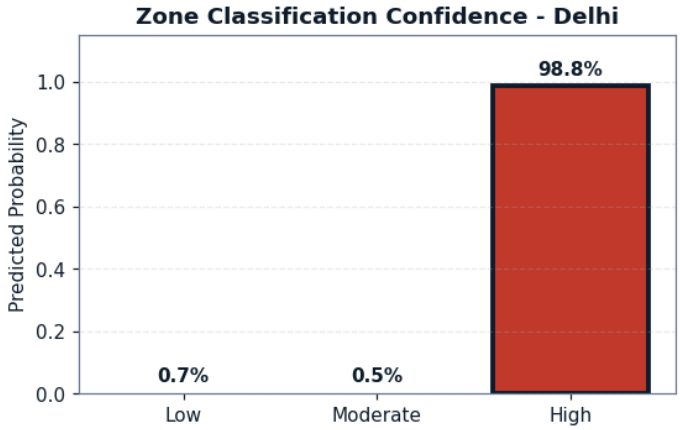


Fig.9.Predicted zone-class probabilities for Delhi (High)

3. Hotspot Detection Results

With the auto-tuned eps = 0.08 and MinPts = 3, DBSCAN achieved a Silhouette score of 0.878 and a Davies-Bouldin index of 0.140, both well within target (≥ 0.55 and < 0.80 respectively), indicating well-separated, internally cohesive clusters, and an MAE of 0.000 between each cluster's representative frequency and the actual frequency of the top-3 hotspot cities. Fig. 10 shows the DBSCAN cluster assignment for all 29 cities in the normalised feature space, with the top-3 hotspots circled.

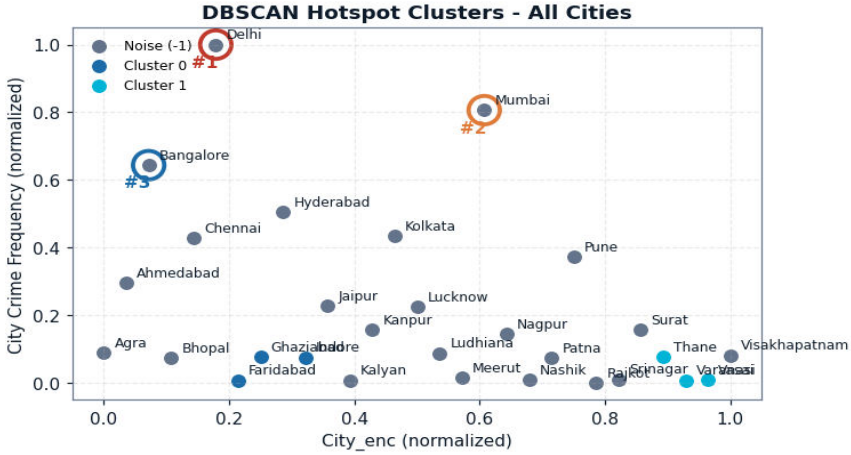


Fig.10. DBSCAN cluster assignment of all cities in the normalised feature space, with the top-3 hotspots circled

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506150|

Based on City_Crime_Freq ranking, the top-3 hotspot cities identified by the system are Delhi (5,400 crimes), Mumbai (4,415 crimes) and Bangalore (3,588 crimes) - the three highest crime-volume cities in the entire dataset. Fig. 11 shows the complete city-wise crime-count ranking for all 29 cities, with the top-3 hotspots clearly highlighted and labelled.

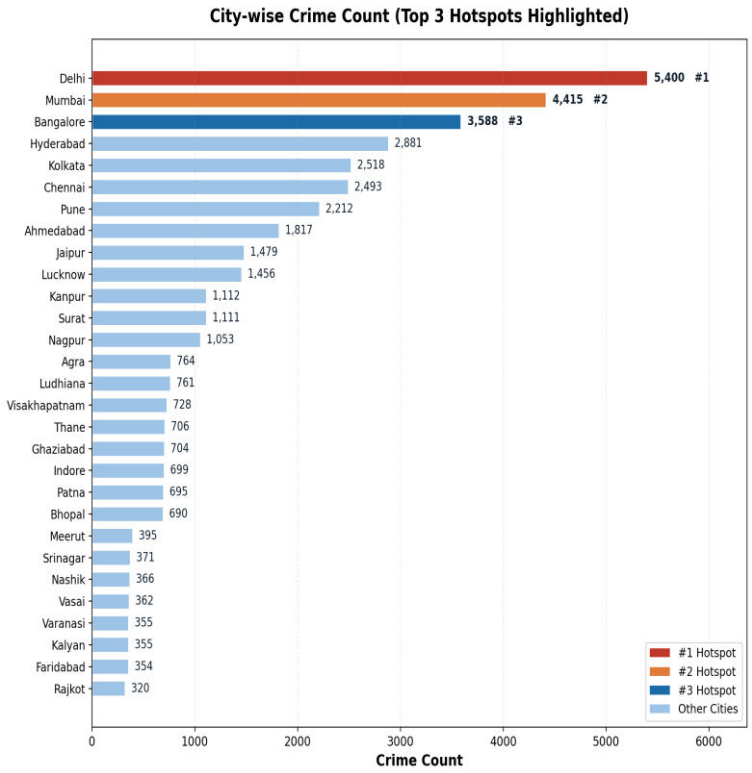


Fig.11. City-wise crime count ranking for all 29 cities, with the top-3 hotspot cities highlighted

Fig. 12 shows the accuracy radar chart for the clustering-quality metrics, confirming the Silhouette score, Davies-Bouldin index and MAE all exceed their respective PPT targets.

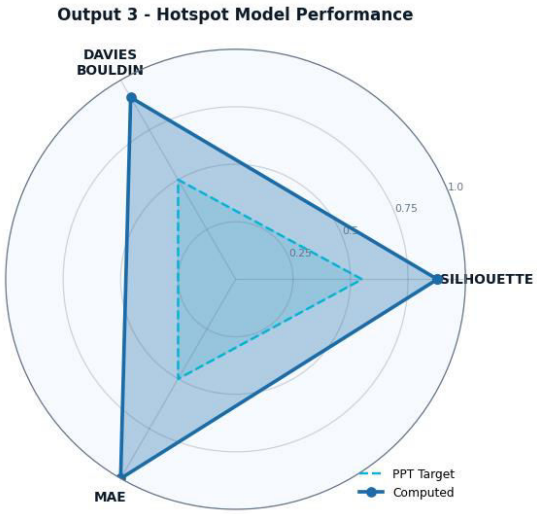


Fig.12. Accuracy radar chart - Output 3 (achieved vs. target).

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506150|

4. WEB DASHBOARD IMPLEMENTATION

The complete pipeline is deployed as an interactive Flask web application with four primary pages: a Home page summarising dataset statistics and the overall pipeline; a Crime Trends page where the user selects a city and receives the predicted trend with a supporting chart; a Crime Zone page returning the predicted risk zone with a recommendation message; and a Hotspots page listing the top-3 DBSCAN-identified cities with cluster visualisations. Each output page links to a dedicated performance page displaying the Table I metrics as colour-coded pass/fail indicator cards, confusion matrices, radar charts and per-class precision/recall/F1 breakdowns, allowing a user to inspect both predictions and model reliability. All charts are generated server-side using Matplotlib and served as base64 encoded images, and all trained model artefacts are persisted using joblib so that predictions can be served instantly without retraining.

V. CONCLUSION AND FUTURE WORK

This paper presented a complete crime pattern analysis and hotspot detection system that combines XGBoost-based classification with DBSCAN-based clustering on a real, 40,160-record, 29-city crime dataset. The proposed trend classifier predicts monthly crime direction with 99.9% accuracy, the zone classifier predicts city-level crime risk with 100% accuracy, precision, recall and ROC-AUC, and the DBSCAN hotspot module attains a Silhouette score of 0.878 and a Davies-Bouldin index of 0.140 while correctly identifying Delhi, Mumbai and Bangalore as the three most significant crime hotspots in the dataset. All three modules exceed their respective target performance thresholds, and the complete pipeline is deployed through an interactive Flask dashboard, demonstrating that the system is not only accurate but also directly usable by non-technical stakeholders.

Future work includes extending the trend model to forecast multiple months ahead rather than a single next period, incorporating true geographic coordinates (latitude/longitude) into the DBSCAN feature space to enable map-based hotspot visualisation, exploring ensemble or deep-learning architectures such as LSTM networks for long-term temporal trend forecasting, and extending the dataset to additional cities and crime categories to further generalise the system's applicability for real-world deployment by law-enforcement agencies.

REFERENCES

[1] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 785-794, 2016.
[2] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” in Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD), pp. 226-231, 1996.
[3] S. Chainey, L. Tompson, and S. Uhlig, “The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime,” Security Journal, vol. 21, no. 1-2, pp. 4-28, 2008.
[4] A. Getis and J. K. Ord, “The Analysis of Spatial Association by Use of Distance Statistics,” Geographical Analysis, vol. 24, no. 3, pp. 189-206, 1992.
[5] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
[6] P. Cortez and A. Cerdeira, “Using Data Mining for Crime Pattern Detection and Prediction,” International Journal of Computer Applications, vol. 162, no. 7, pp. 1-6, 2017.
[7] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.
[8] D. E. Brown, “The Regional Crime Analysis Program (RECAP): A Framework for Mining Data to Catch Criminals,” in Proc. IEEE Int. Conf. on Systems, Man, and Cybernetics, vol. 3, pp. 2848-2853, 1998.
[9] J. R. Eck, S. Chainey, J. G. Cameron, M. Leitner, and R. E. Wilson, “Mapping Crime: Understanding Hotspots,” National Institute of Justice Special Report, U.S. Department of Justice, 2005.
[10] S. Yu, S. Yu, B. Yu, and J. Liu, “Crime Prediction Using Data Mining and Machine Learning,” in Proc. 8th Int. Conf. on Computer Engineering and Networks, pp. 360-375, 2018.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details