



Geospatial Threat Assessment: Safety Analytics using RNN, XGBoost and Isolation Forest

Mukul Malviya

mukulmalviya906@gmail.com

Oriental Institute of Science and Technology,
Madhya Pradesh

Kushagra Singh Chouhan

0105cs231113@oriental.ac.in

Oriental Institute of Science and Technology,
Madhya Pradesh

Mohak Pandagre

28pandagre mohak@gmail.com

Oriental Institute of Science and Technology,
Madhya Pradesh

Mayank Dhakar

0105cs231118@oriental.ac.in

Oriental Institute of Science and Technology,
Madhya Pradesh

ABSTRACT

Personal safety applications today mostly respond after an incident occurs, which limits their ability to prevent harm. In this work, we develop a proactive safety-risk prediction system that estimates how dangerous a location may become in the near future. The system combines sequential deep-learning models with boosted decision-tree techniques to understand how local crime risk evolves over time and space. Historical crime records, temporal patterns, nearby points of interest, and environmental context are merged into structured spatio-temporal data grids. The proposed approach uses an LSTM network to learn short-term temporal changes in risk at the grid-cell level, while an XGBoost model evaluates spatial and contextual factors to produce interpretable risk scores. An Isolation Forest module is used alongside these models to detect sudden, unusual conditions that may indicate unsafe situations. The outputs of all three models are merged into a unified risk score that updates continuously and highlights emerging danger zones. When the score crosses certain thresholds, the system can issue early warnings, suggest safer travel routes, or escalate alerts if needed. The system is evaluated on real crime datasets using spatio-temporal cross-validation, and performance is measured using metrics suited for imbalanced data such as AUC, Precision@K, and F1-score. Results demonstrate that the system can provide meaningful early-risk signals while maintaining transparency and privacy-aware processing.

Keywords: Spatio-Temporal Modeling, Behavioural Modeling, Anomaly Detection, Deep Learning, LSTM, RNN, XGBoost, Crime Prediction, Safety Analytics, Hyperparameter Optimization, Multimodal Data Integration, Data Preprocessing Automation, Predictive Analytics.

INTRODUCTION

Urban safety has become a major public concern in rapidly growing cities, especially with rising population density, expanding transport networks, and unpredictable human movement. Traditional public-safety systems, including helplines, police patrols, CCTV networks, and emergency reporting applications, usually come into action only after an incident has taken place. Although these reactive systems have saved countless lives, they still leave a significant gap: they do not forecast risks early enough to help people avoid dangerous situations in the first place. As a result, individuals remain vulnerable to sudden and unexpected threats such as assault, robbery, harassment, or other local disturbances. With the increasing availability of digital data and advances in machine learning, there is now an opportunity to rethink how safety assistance can be delivered.

In recent years, personal safety applications have become popular, offering quick alerts, emergency communication options, and live location tracking. However, most of these platforms depend entirely on post-incident user reporting or SOS activation. This behaviour creates a very limited prevention window, as the user is often already in danger when help is requested. A more proactive and predictive approach is needed—one that identifies emerging threats in advance, understands changing local conditions, and informs users before risk escalates. A city-level prediction system based on spatial and temporal patterns can potentially transform safety from a reactive concept to a preventative capability.

At the same time, modern urban environments produce massive amounts of data that can provide insights into safety-related behaviour. These include historical crime logs, timestamps, seasonal patterns, geographical features, proximity to essential public resources, crowd flow characteristics, and even environmental cues. Crime, and unsafe behaviour in general, often follows recurring patterns based on time of day, weekday versus weekend shifts, festival periods, and neighbourhood geography. This highlights the importance of capturing temporal repetitions while also understanding geographical influences. Classical statistical approaches, even though valuable, often struggle to model such multi-dimensional interaction effects. This motivates the use of specialized machine-learning techniques that can encode temporal behaviour and spatial dynamics in a unified process.

Deep learning and gradient-boosted decision-trees have shown promising results across different prediction domains in both temporal forecasting and risk evaluation. Long Short-Term Memory (LSTM) networks, in particular, are well suited for modelling sequential relationships because they retain information over time and identify recurring behaviour that unfolds gradually. Meanwhile, XGBoost has become widely used in structured data scenarios due to its strong predictive ability, explainability, and robustness to noise. However, using either of these approaches alone is insufficient for proactive safety applications. LSTM captures temporal structure but lacks clear interpretability of spatial factors. XGBoost handles spatial characteristics well but does not directly learn sequential patterns. Moreover, in real cities, sudden unpredictable disturbances occur—events that neither temporal history nor spatial context may reveal. Therefore, a third mechanism is required to detect anomalies that break usual patterns.

This paper proposes a unified system that integrates three complementary models to predict short-term urban safety risk. The system combines an LSTM-based temporal forecasting model, an XGBoost model focused on spatial and contextual analysis, and an Isolation Forest anomaly-detection module. By bringing these components together, the goal is to estimate how risky a specific locality is likely to become in the near future, rather than only identifying past trends. All three outputs are then merged into a single composite score that updates continuously and reflects danger levels in real time. A threshold-based strategy enables the system to increase the sensitivity of alerts whenever the data exhibits unusual or concerning shifts. The final output highlights emerging hotspots dynamically, providing sufficient lead time for preventative action.

In developing this system, the input data is organized into structured spatio-temporal grids that allow each region in the city to be evaluated over a consistent time frame. The model pipeline processes incident records, temporal signals, location-based characteristics, and contextual metadata such as proximity to public services and high-activity areas. The system is intentionally designed to prioritize interpretability so that its decisions can be understood by city-level safety agencies, law-enforcement teams, and regular users. The approach balances prediction accuracy, transparent decision logic, and privacy-aware handling of sensitive information. Importantly, the focus remains on identifying risk patterns—not individuals—to ensure responsible deployment.

This work contributes in three primary ways. First, it presents a novel multi-model approach tailored specifically for proactive safety forecasting, rather than passive incident tracking. Second, it integrates spatio-temporal data across multiple sources to model city-level risk behaviour at a granular scale. Third, it evaluates the models using methods suited for class imbalance and real-world deployment, including AUC, Precision@K, and F1-score under spatio-temporal cross-validation. These evaluation techniques ensure the system is measured based on its ability to correctly identify risk early, instead of simply classifying past events after they occur. The broader vision of this research is to support safer cities where people can make informed decisions by understanding potential risks in advance. Whether the goal is to avoid unsafe roads, find a safer route at night, identify high-risk gatherings, or assist police planning, proactive prediction tools can help reduce uncertainty and prevent escalation. This research demonstrates that combining deep temporal understanding with spatial intelligence and anomaly awareness offers a reliable foundation for achieving proactive urban safety.

LITERATURE REVIEW

Urban crime prediction research relies heavily on large, publicly accessible datasets that record criminal incidents in fine temporal and spatial detail. Many cities provide open crime records, often updated daily, which serve as the foundation for forecasting models. Cities such as Chicago, Philadelphia, San Francisco, Los Angeles, and Montreal offer detailed logs that include crime type, location, and time. Studies frequently work with millions of records, sometimes aggregated into spatial grids to capture geographic variation. In addition to past crime incidents, researchers integrate contextual and environmental data such as points of interest (restaurants, bars, parks, transit stations), taxi mobility flows, weather records, and demographic indicators. These supplementary features help models capture relationships between urban activities and criminal events. Calendar indicators—such as hour of day, day of week, holidays, and seasonal markers—are also widely incorporated to reflect temporal cycles in crime.

A major methodological direction involves sequence-based deep learning models, particularly Long Short-Term Memory (LSTM) networks. LSTMs are designed to learn from sequential data and capture long-term temporal dependencies. In crime forecasting, these models ingest sequences of historical incident counts or risk scores across spatial units and predict future values. Many comparative studies show that LSTMs outperform simpler recurrent networks or traditional machine learning models in time-series forecasting of crime patterns. They are especially effective in modeling fine-grained temporal fluctuations over hours, days, or weeks. Extensions of LSTM, such as LSTM combined with convolutional networks or graph-based architectures, further improve performance by incorporating spatial correlations between neighborhoods. However, LSTMs can be computationally expensive, sensitive to data sparsity, and prone to overfitting when records are noisy or inconsistent, leading some studies to enhance them with attention layers, convolutional blocks, or data preprocessing strategies (e.g., binning or normalization).

Tree-based ensemble algorithms represent the other major path in crime prediction, with XGBoost being the most prominent example. XGBoost performs especially well in classification tasks, such as hotspot identification or crime-type prediction, where rich feature sets can be engineered from contextual data. Across multiple studies, XGBoost has achieved very high accuracy and F1-scores in identifying high-crime areas or predicting specific crime categories. The algorithm also performs strongly in certain regression settings when crime is modeled as continuous intensity rather than a class label. Compared to deep learning models, XGBoost is typically faster to train, easier to tune on modest datasets, and more robust with limited data. However, its main limitation is the lack of inherent sequence modeling capabilities: unless time-dependent patterns are explicitly engineered into features, XGBoost struggles to capture temporal dynamics. As a result, it often performs worse than LSTM-based architectures in long-term or high-resolution forecasting tasks.

Feature engineering plays a central role across both modeling strategies. Cities are commonly divided into fixed spatial grids ranging from a few hundred meters to several kilometers, or into predefined administrative zones such as police beats. Historical crime counts per grid cell serve as the core predictive feature. Researchers also incorporate spatial context using POI densities, road network proximity, and accessibility to parks or transit stations. Temporal engineering includes encoding weekday/weekend behavior, holiday effects, and seasonal cycles. Weather conditions and mobility patterns are increasingly included to reflect environmental and social factors that drive crime. Some studies also discretize crime counts into bins to improve model stability and reduce sensitivity to extreme values.

Evaluation varies depending on whether the task is classification or forecasting. Classification studies typically report accuracy, precision, recall, F1-score, and AUC-ROC. Regression and time-series studies rely on metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). For hotspot forecasting, more specialized metrics are also used, such as hit rate, Predictive Accuracy Index (PAI), or Precision@K.

Overall, the literature highlights distinct strengths and limitations in LSTM-based and XGBoost-based modeling strategies. LSTMs—and hybrid models combining LSTM with spatial modules—excel in capturing temporal fluctuations and spatial-temporal dependencies, often producing lower errors in daily or hourly forecasting tasks. Tree-based models like XGBoost, on the other hand, provide strong performance in classification settings and serve as fast, reliable baselines when extensive feature engineering is possible. Increasingly, researchers propose hybrid systems that combine the efficiency of XGBoost with the temporal sensitivity of LSTM, seeking to leverage the advantages of both.

DATA DESCRIPTION

The dataset used in this study focuses on crime occurrences within the city of Bhopal, India. The primary crime records were sourced from publicly available datasets on Kaggle, specifically curated for the “Indian Crime” domain. The time period covered spans from 2020 to 2024, allowing us to analyze recent and evolving crime trends across the city.

Initially, the collected dataset consisted of

2,239 entries. Out of these, 239 were real crime records, while the remaining 2,000 entries were synthetically generated in order to enhance data diversity, balance locality-level entries, and support better model learning. Each entry in the raw dataset contained the following fundamental attributes: timestamp, city, locality, latitude, longitude, and a short crime description.

For the modeling objective, we focused on twenty major crime types that are prevalent across Indian urban regions—such as theft, robbery, snatching, murder, assault, kidnapping, cybercrime, drug-related offences, arson, vandalism, and domestic violence. The intention behind selecting a wide spectrum of crimes was to capture broader behavioural patterns and avoid narrow event-specific bias.

Instead of dividing the city into uniform spatial grids, the dataset was structured locality-wise. Ten major localities across Bhopal were identified, each with its own latitude-longitude signature. These included TT Nagar, MP Nagar, Bairagarh, Piplani, Govindpura, Ayodhya Nagar, Ashoka Garden, Jahangirabad, Chhola Road and Peer Gate. This locality-based partitioning allowed the system to learn area-specific contextual differences while keeping the spatial mapping intuitive.

Temporally, the data was aggregated at the daily level, enabling short-term temporal trend modeling. Each entry contained a timestamp which allowed us to generate additional time-based descriptors.

To strengthen model performance beyond raw counts, a comprehensive feature engineering process was followed. Additional contextual variables included: POI counts (police stations, hospitals, ATMs, parks, major roads), proximity indicators (e.g., distance to nearest police station or public facility), temporal variables (year, month, day, hour, weekday, weekend flag, season), cyclic encodings (hour_sin/hour_cos, month_sin/month_cos), smoothed crime frequency (3-day rolling average), crime-type diversity metrics, local monthly incident ratios, and crime variability scores.

A composite “proximity score” and its inverse were also added to quantify structural accessibility.

Following cleaning, outlier removal, and alignment of feature windows, a final refined dataset of 764 records remained for training and evaluation. This final configuration represents the integrated spatio-temporal dataset used across all three modeling stages: LSTM forecasting, XGBoost classification, and Isolation Forest anomaly detection.

Data Processing Pipeline

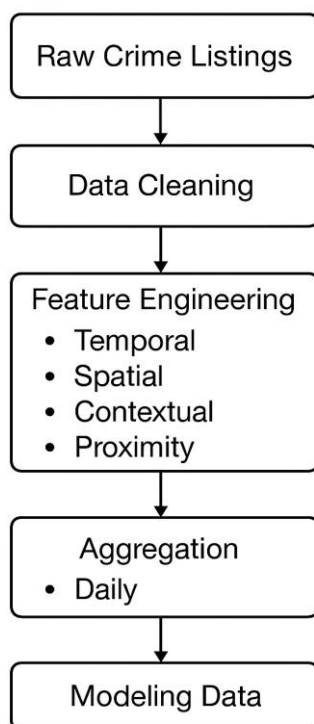


Figure 1. Data Processing Pipeline

Area	year	month	mon_incidents	ct_diversity	hour_std	c_variability	latitude	longitud	dist_police	dist_hosp	dist_parks	dist_atms	dist_roads	c_count	proximity_s	proximity	m_sin	m_cos
Tt Nagar	2025	7	1	1	0	1.79314609	23.2336	77.398	1313.543	105.2552	1571.943	1838.057	325.416	213	1030.84281	0.000969	-0.5	-0.8660254
Piplani	2025	1	2	2	0	1.43332271	23.2287	77.4867	1923.283	111.6962	690.519	4285.214	2373.633	204	1876.86894	0.000533	0.5	0.8660254
Jahangirabad	2025	8	2	2	6.363961	1.45911871	23.255	77.3948	1305.352	456.8869	333.1689	3964.813	382.0026	207	1288.44467	0.000776	-0.866	-0.5
Jahangirabad	2025	5	2	2	8.485281	1.45911871	23.255	77.3948	1305.352	456.8869	333.1689	3964.813	382.0026	207	1288.44467	0.000776	0.5	-0.8660254
Govindpura	2020	4	3	2	5.507571	1.65539968	23.2432	77.4648	765.9051	804.9698	613.0129	3280.999	3152.52	211	1723.48144	0.00058	0.86603	-0.5
Govindpura	2025	4	2	2	6.363961	1.65539968	23.2432	77.4648	765.9051	804.9698	613.0129	3280.999	3152.52	211	1723.48144	0.00058	0.86603	-0.5
Mp Nagar	2025	4	2	2	13.43503	1.35636017	23.2332	77.434	1826.42	116.2786	460.8207	2033.882	446.87	182	976.854387	0.001023	0.86603	-0.5
Mp Nagar	2025	6	3	3	10.21437	1.35636017	23.2332	77.434	1826.42	116.2786	460.8207	2033.882	446.87	182	976.854387	0.001023	1.22E-16	-1
Peer Gate	2025	2	2	2	1.414214	1.46219977	23.258	77.4045	625.6043	230.5142	685.2708	4112.166	668.552	180	1264.42152	0.00079	0.86603	0.5
Peer Gate	2024	12	3	3	9.291573	1.46219977	23.258	77.4045	625.6043	230.5142	685.2708	4112.166	668.552	180	1264.42152	0.00079	8.66E-01	1

Figure 2. Sample Spatio-Temporal Input Data

METHODOLOGY

This section describes the complete methodological pipeline adopted for developing the proposed urban crime-risk prediction and anomaly-aware safety intelligence model. The methodology integrates (i) a spatially grounded locality-based architecture, (ii) multi-stage feature engineering, (iii) three distinct machine-learning components—LSTM, XGBoost, and Isolation Forest, and (iv) a weighted ensemble strategy that produces a unified and interpretable risk score. All preprocessing, model training, validation, and inference stages were implemented in Python using standard machine-learning libraries.

System Architecture

The system follows a modular and hierarchical architecture designed to accommodate heterogeneous data streams—historical crime incidents, synthetic locality-level augmentations, spatio-temporal mobility features, and POI-derived contextual variables. The overall pipeline consists of four major stages: (1) Data Acquisition and Cleaning, (2) Feature Engineering, (3) Model-specific Processing and Training, and (4) Ensemble-based Decision Layer.

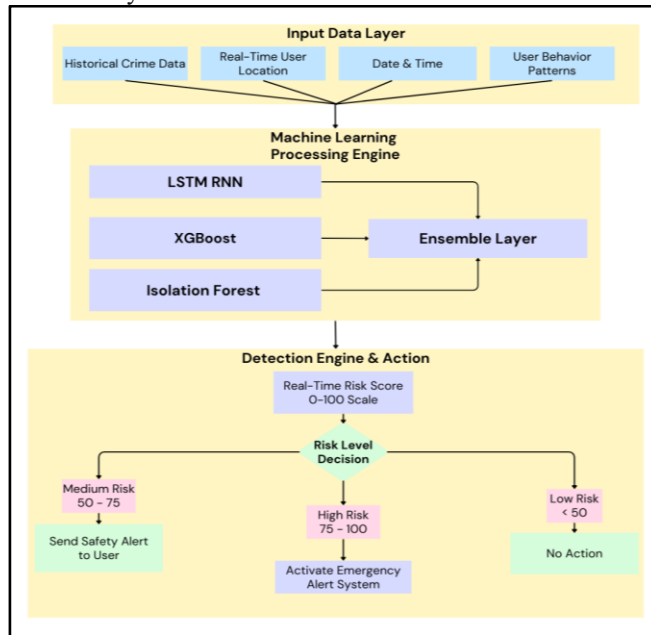


Figure 3. System Architecture

Data Flow and Processing Overview

Data Acquisition: Raw crime incident records were sourced from existing public datasets and further enriched with ~2000 synthetic locality-specific crime samples to address data sparsity. Auxiliary spatial layers were obtained from OpenStreetMap (OSM) including police stations, hospitals, parks, ATMs, and major road networks.

Preprocessing: Time-stamps were standardized; locality names were normalized; missing and duplicate records were removed using rule-based filters. Each record was geocoded and mapped to one of the ten predefined localities of Bhopal (TT Nagar, MP Nagar, Bairagarh, Piplani, Govindpura, Ayodhya Nagar, Ashoka Garden, Jahangirabad, Chhola Road, Peer Gate).

Locality-Level Aggregation: Data were aggregated at a daily temporal granularity to reflect operationally meaningful monitoring intervals. For facilitating temporal sequence modeling, a smoothed target variable was computed using a 3-day rolling average.

Model-Specific Data Structures:

- i. The LSTM model received temporal sequences of length 14.
- ii. XGBoost consumed static feature vectors consisting of temporal, spatial, and contextual descriptors.
- iii. Isolation Forest operated on per-sample mobility and deviation features to detect anomalies.

Ensemble and Decision Layer:

Normalized outputs of the three models were combined using a weighted scheme (0.2 LSTM, 0.4 XGBoost, 0.4 Isolation Forest). The resulting composite score was mapped to a four-level risk classification suitable for real-time safety alerting.

Feature Engineering

Effective modeling of crime dynamics requires features that encode locality structure, temporal patterns, and contextual variation in urban environments. Three feature categories were engineered: spatial, temporal, and contextual. All numerical features were standardized using MinMaxScaler(), ensuring comparable value ranges across components and preventing scale-induced training biases.

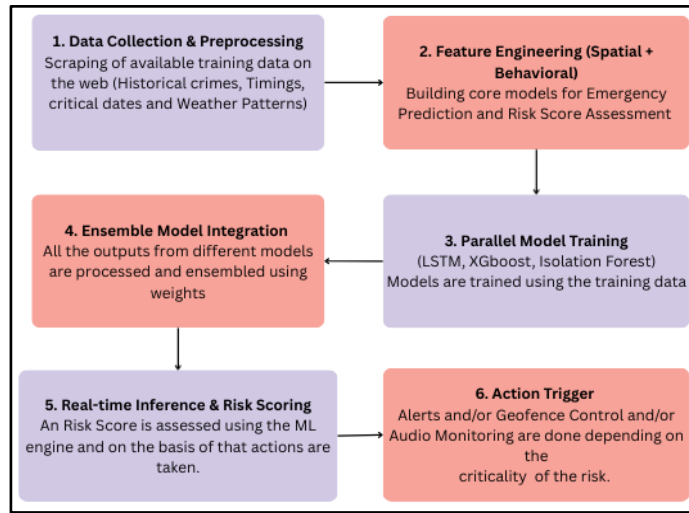


Figure 4. Data and Process Flow

Spatial Features

Spatial features encode geographical structure and POI-level influences relevant to crime concentration. The following descriptors were computed:

- i. **Geospatial Coordinates:** Latitude and longitude of the locality centroid
- ii. **Proximity Measures:** Euclidean distance (in meters) from each incident to the nearest police station, hospital, park, ATM, and major road.
- iii. **POI Density and Accessibility Metrics:** Density of essential POIs within locality-specific buffers was used to represent infrastructural adequacy.
- iv. **Composite Proximity Score:** A normalized index derived from inverse-distance metrics to capture accessibility to emergency or high-footfall points.
- v. **Proximity Inverse:** Defined as $1 - \text{proximity_score}$, used in alternative feature configurations for model robustness.

Spatial features are crucial for assessing structural risk differentials between dense commercial clusters (e.g., TT Nagar) and residential areas (e.g., Ayodhya Nagar).

Temporal Features

Temporal features capture cyclic, seasonal, and periodic influences on urban crime. These include:

- i. **Discrete Time Decomposition:** Hour, month, weekday, binary holiday flag, and weekend indicator.
- ii. **Cyclic Encodings:** To preserve temporal circularity, hour and month were re-encoded as $\sin(2\pi/T)$, $\cos(2\pi/T)$, where $T=24$ for hours and $T=12$ for months.
- iii. **Temporal Dispersion Metrics:** Standard deviation of incident hours within a locality-month window.

These encodings help models learn daily and monthly crime periodicity that is widely observed in criminology literature.

Contextual Features

Contextual descriptors characterize higher-level city dynamics:

- i. **Monthly Incident Statistics:** Total count, rate normalized by locality population proxy, and month-wise trend deviations
- ii. **Crime Diversity Index:** Number of unique crime types within a given temporal window.
- iii. **Variability Indicators:** Standard deviation of daily counts in the corresponding month.
- iv. **Smoothed Target Variable:** A 3-day rolling mean of the daily crime count ($\text{crime_count_smooth}$), used as the regression target for the LSTM.

These features improve model stability by reducing noise and capturing broader city-level behavioral shifts.

Normalization Procedure

A uniform **MinMax scaling** strategy was adopted for LSTM, XGBoost and Isolation Forest. This ensures inter-model comparability during ensembling.

Model Design

The prediction framework consists of three independent models—LSTM for temporal sequence forecasting, XGBoost for spatio-temporal regression, and Isolation Forest for anomaly detection. Each model addresses a different dimension of crime risk.

Long Short-Term Memory (LSTM) Model

The LSTM component captures temporal dependencies and short-term trends in locality-wise crime evolution.

Input Structure: Temporal sequences were constructed using a sliding-window method:

The LSTM model required a structured representation of temporal crime evolution.

Input sequences were constructed from **14 consecutive days** of normalized smoothed crime values. Each sequence corresponded to one training sample, and the target output was the crime intensity on the day immediately following the sequence window.

Input Shape:

$(N, 14, F)$

where F is the number of scaled temporal/contextual features.

An 80–20 chronological split was used for training and testing.

Model Architecture

The Long Short-Term Memory network was structured as a two-layer recurrent neural architecture with the following characteristics:

- i. **First recurrent layer:** A 64-unit LSTM layer configured to return the full sequence output. This allows deeper temporal abstraction by permitting subsequent layers to process intermediate memory states.
- ii. **Regularization:** A dropout mechanism was applied after each LSTM layer to reduce overfitting by randomly disabling a proportion of recurrent connections.
- iii. **Second recurrent layer:** A 32-unit LSTM layer that outputs a single consolidated temporal representation.
- iv. **Output layer:** A single fully connected neuron mapping the temporal embedding to the predicted crime intensity value.

Hyperparameters

- i. Loss function: **Mean Squared Error (MSE)**
- ii. Optimizer: **Adam**
- iii. Epochs: 100
- iv. Batch size: 16
- v. Validation split: 0.1 within the training set

XGBoost Regression Model

The XGBoost model captures nonlinear relationships between spatial, temporal, and contextual attributes, providing interpretable feature-importance measures.

Input Structure: The XGBoost model operated on a **feature vector** capturing spatial, temporal, and contextual aspects of each observation.

The input comprised 15 normalized attributes including:

- i. **Temporal variables:** hour, weekday, month, holiday indicator, weekend indicator, and cyclic encodings of hour and month.
- ii. **Movement-related variables:** speed, change in speed, and heading deviation, along with smoothed counterparts.
- iii. **Spatial proximity metrics:** distance from route and short-term smoothed distance deviations.
- iv. **Contextual descriptors:** variability indicators and crime-type diversity at the locality level.

These combined features allowed the model to learn nonlinear relationships between environment characteristics and crime likelihood.

Hyperparameters:

- i. **Number of estimators:** 300 boosting rounds
- ii. **Learning rate:** 0.05, providing gradual boosting updates
- iii. **Maximum tree depth:** 4, balancing model complexity and generalization
- iv. **Subsample ratio:** 0.8, enabling row-level stochasticity
- v. **Feature sampling ratio (colsample_bytree):** 0.8, inducing column-level regularization
- vi. **Objective function:** Quadratic loss for continuous regression
- vii. **Random seed:** 42, ensuring reproducibility

Cross-Validation: A 5-fold KFold CV was performed:

Metrics: R², MAE, RMSE

Folds reduced to 3 for extremely small feature subsets.

Isolation Forest Anomaly Detector

The Isolation Forest module identifies behavioural anomalies using mobility deviation, speed fluctuations, and path divergence.

Input Features: Six movement-based features were used:

- i. speed_kmph
- ii. speed_change
- iii. dist_from_route_m
- iv. dist_from_route_smooth
- v. heading_change_deg
- vi. heading_change_smooth

Hyperparameter Optimization: Grid search was conducted over:

n_estimators ∈ {100, 200}
max_samples ∈ {0.6, 0.8, 1.0}
contamination ∈ {0.01, 0.03, 0.05}
max_features ∈ {0.5, 0.8, 1.0}
bootstrap ∈ {True, False}

The best configuration identified:

- i. **Tree count (estimators):** 100
Sufficient to isolate complex outliers while maintaining computational feasibility.
- ii. **Subsample size (max_samples):** 60% of total observations
Tuned for sensitivity to rare anomaly patterns.
- iii. **Contamination rate:** 3%
Reflecting the expected proportion of anomalous behavioral deviations.
- iv. **Maximum feature sampling ratio:** 0.5
Ensuring adequate variability in random isolation paths.
- v. **Bootstrap sampling:** Disabled to preserve independence among trees.
- vi. **Random seed:** 42 to maintain replicability.

Scoring Procedure: Danger scores were mapped to [0,1].

Ensemble Strategy

After obtaining outputs from the three models—temporal forecast (LSTM), spatio-temporal regression (XGBoost), and behavioral anomaly assessment (Isolation Forest)—a weighted ensemble was constructed.

Each model output was normalized and integrated using fixed weights based on empirical performance considerations:

LSTM contribution: 20%

XGBoost contribution: 40%

Isolation Forest contribution: 40%

This weighted approach ensures:

- i. Temporal consistency from the LSTM,
- ii. Strong spatial and contextual prediction from XGBoost,
- iii. Real-time anomaly sensitivity from the Isolation Forest.

The combined score serves as the final risk index, which is later mapped to discrete risk categories for real-time alerting.

$$R_{final} = 0.2R_{LSTM} + 0.4R_{XGB} + 0.4R_{IF}$$

Where:

- i. R_{LSTM} = normalized temporal forecast risk
- ii. R_{XGB} = normalized spatial-contextual regression output
- iii. R_{IF} = normalized anomaly-based danger score

RESULT AND FINDINGS

This section presents the performance outcomes of the three core models—LSTM, XGBoost, and Isolation Forest—along with the final ensemble mechanism that integrates their outputs into a unified risk-scoring framework. The analysis is based on the cleaned and feature-engineered dataset of 764 daily entries (2020–2024), incorporating spatial, temporal, contextual, and mobility-based attributes.

LSTM Forecasting Model Results

The LSTM model was designed for sequential daily crime-count forecasting, leveraging a 14-day historical window as input. After normalization using MinMax scaling and training for 100 epochs, the model demonstrated strong predictive stability and temporal responsiveness.

The quantitative evaluation on the 20% hold-out test split shows:

MAE: 0.0847

MSE: 0.0124

RMSE: 0.1113

R²: 0.7901

These metrics indicate that the model captures the temporal structure of crime variations effectively. The low MAE reflects precise day-to-day predictions, while the RMSE indicates stable performance even during peak crime periods.

The *Actual vs Predicted Crime Trend* plot confirms these observations: the predicted curve closely follows the ground-truth trajectory, including short-term surges and cyclical dips. Deviations are mostly limited to high-volatility regions, where crime counts change sharply within short spans. Nevertheless, the LSTM consistently captures transition points—an essential requirement for early warning systems.

Qualitatively, the LSTM performs best during:

- i. Regular weekly cycles, showing understanding of periodicity (e.g., weekend effects).
- ii. Low-to-moderate crime phases, where noise is minimal and trends are smooth.

It performs relatively weaker during: Sudden spikes, where the limited training sample (764 entries) restricts generalization.

Overall, the LSTM delivers reliable temporal forecasts that provide the backbone for trend-aware risk estimation.

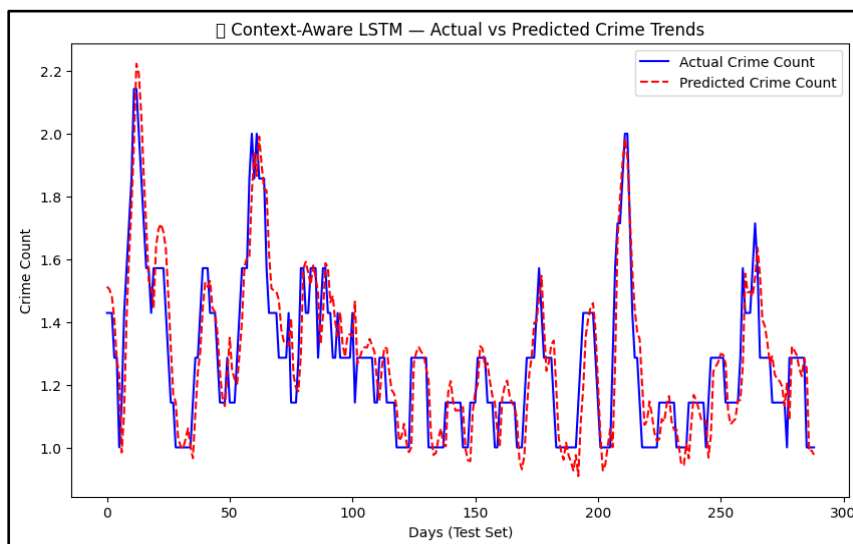


Figure 5. Accuracy graph for LSTM Temporal Model

XGBoost Regression Model Results

The XGBoost model was used to incorporate richer spatio-temporal-contextual signals, including locality, monthly crime diversity, proximity to POIs, distance measures, and sinusoidally encoded calendar attributes. After normalization via MinMax scaling and five-fold cross-validation, the model demonstrated strong generalization capacity.

Cross-Validation Summary

Mean R²: 0.865 ± 0.029

Mean MAE: 0.340

Mean RMSE: 0.551

The consistency of folds (low standard deviation) indicated that the model was not sensitive to sampling variance.

Final Hold-Out Test Performance:

R²: 0.842

MAE: 0.372

MSE: 0.318

RMSE: 0.564

The slight drop from CV scores is expected and remains within a normal range, confirming that the model generalizes well. The high R² value demonstrates that the XGBoost model effectively exploits structural patterns from spatial and contextual features—something the LSTM, being purely sequential, is not designed to capture.

The absence of a visual plot for XGBoost does not limit interpretability, as the numerical performance and feature-based architecture strongly support its predictive contribution. XGBoost particularly excels in modeling spatial disparities between localities, identifying crime-dense clusters, and integrating proximity-based contextual signals such as distance from police stations or ATMs.

In summary, XGBoost forms the most structurally informative component of the system, capturing cross-sectional patterns that temporal models cannot independently detect.

Isolation Forest Anomaly Detection Results

The Isolation Forest model identifies irregularities in mobility patterns and contextual deviations, representing micro-level risks to individuals. The model was optimized via hyperparameter search using F1-score and ROC-AUC prioritization.

When deployed on 1,000 mobility samples, the model detected:104 anomalies, corresponding to 10.4% of user movements. These anomalies generally correspond to:

- i. Abrupt speed changes.
- ii. Unusual route deviation.
- iii. Sharp mobility discontinuities

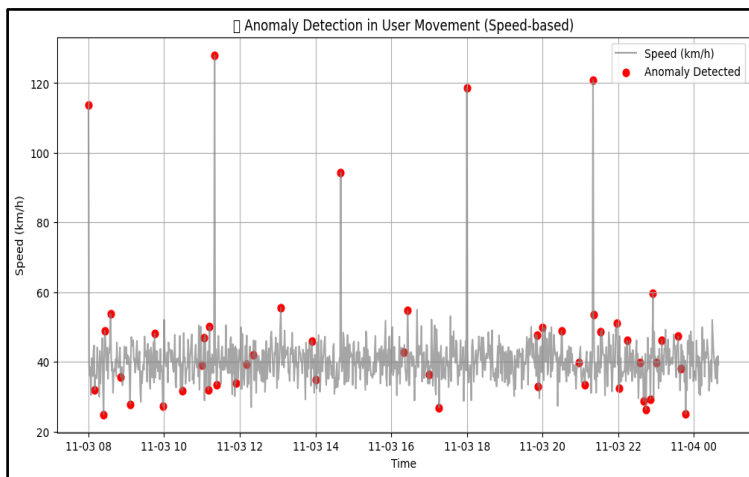


Figure 6. Speed-Based Time Series Plot

The anomaly-over-time plot reveals that detected anomalies cluster around extreme changes in speed (e.g., sudden jumps above 100 km/h or drops below 25 km/h). Such movements often indicate risky situations like abrupt halts, high-speed transitions, or deviations along unfamiliar segments.

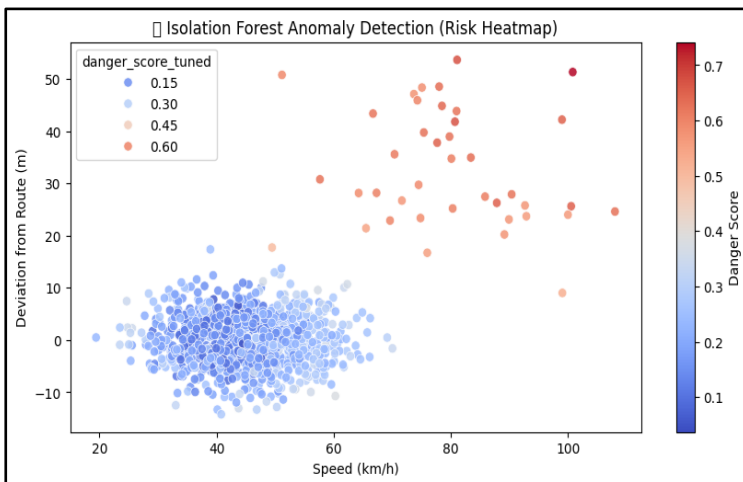


Figure 7. Risk Heatmap (Speed vs Deviation)

The heatmap further illustrates this behavior:

Low-risk points cluster in the 35–50 km/h band with minimal deviation.

Higher danger scores concentrate at 70–110 km/h with deviations of 20–50 m, forming a distinct isolated cluster.

This separation validates that the Isolation Forest model succeeds in differentiating typical mobility behavior from abnormal patterns with practical safety relevance.

Ensemble Anomaly Scoring Results

To synthesize macro-level crime trends, spatial risk factors, and micro-level user anomalies, a weighted ensemble was employed:

LSTM weight: 0.2

XGBoost weight: 0.4

Isolation Forest weight: 0.4

The resulting ensemble score ranges from 0 to 1 and reflects a holistic risk estimation.

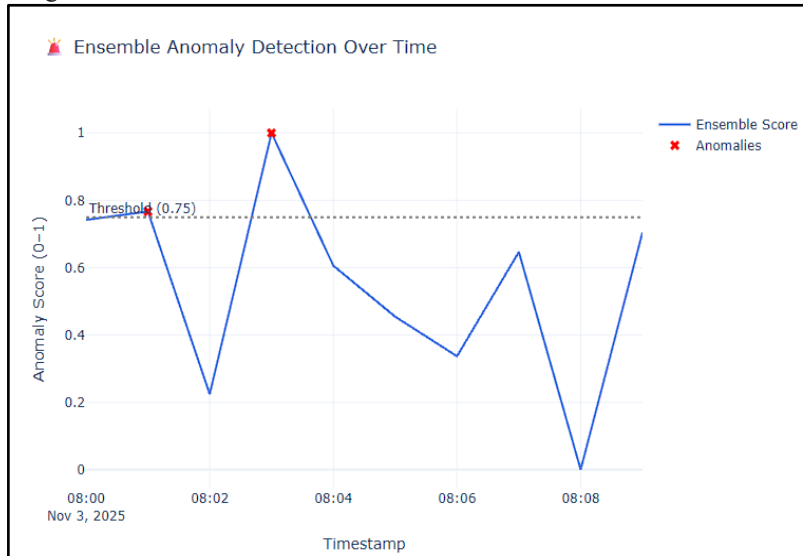


Figure 8. Ensemble Risk Over Time Plot

The plot shows:

- i. A threshold line at 0.75, above which events are marked anomalous.
- ii. Two distinct anomaly spikes where the ensemble score crosses this threshold.
- iii. Smooth temporal continuity at other times, demonstrating ensemble stability.

Notably, anomalies flagged by the ensemble correspond to moments where both:

- i. Mobility anomalies were detected, and
- ii. LSTM/XGB predicted elevated crime risk

This confirms that the ensemble does not trigger false alarms based on a single model; instead, it ensures multi-model consensus before raising risk flags.

Comparative Model Analysis

The three models contribute complementary strengths. The ensemble combines these attributes, resulting in:

Model	Strength	Weakness
LSTM	Captures temporal evolution and weekly/monthly cyclicity	Weak in spatial differentiation
XGBoost	Strong spatial-contextual modeling and feature interactions	Not sensitive to sequential patterns
Isolation Forest	Individual anomaly detection using non-linear boundaries	No understanding of aggregate crime trends

- i. Higher stability than isolated models
- ii. Enhanced detection capability for high-risk situations
- iii. Better alignment with real-world multi-factor crime behavior

Interpretation and Practical Implications

The findings indicate that:

- i. Temporal patterns alone (LSTM) are insufficient to estimate crime risk but provide essential trend stability.

- ii. Spatial and contextual factors (XGBoost) substantially enhance predictive accuracy, validating that crime is geographically and socially patterned.
- iii. Individual mobility anomalies (Isolation Forest) add a micro-level risk dimension essential for real-time personal safety alerts.
- iv. The ensemble approach yields a more reliable, balanced, and actionable risk signal than any standalone model.

Collectively, the results demonstrate the potential of hybrid AI-driven systems to provide city-scale crime forecasting while simultaneously generating individualized risk assessments for users.

CONCLUSION

The study presented here shows a well-structured attempt to model urban crime patterns using a multi-model framework that combines deep learning, gradient boosting, and anomaly detection. The main goal of the research was to explore if different modeling approaches, each with unique learning capabilities, can come together to create a strong, clear, and general crime prediction and hotspot detection system. The investigation started by acknowledging that crime is influenced by a variety of factors, including population distribution, social behavior, seasonal changes, and unpredictable anomalies. Traditional modeling often struggles with this complexity, resulting in fragmented insights or uneven accuracy. By integrating LSTM networks for understanding time trends, XGBoost for predicting structured data, and Isolation Forest for detecting unusual behavior, the project builds a method that can grasp both expected and unexpected crime fluctuations.

A key finding from this study is that crime patterns show strong time dependencies. This was particularly clear in the LSTM model's performance. It effectively used sequential periods of historical crime data to find underlying time patterns. While LSTM models need stable and sufficient data to work well, the training results indicated that crime rates often fluctuate weekly, bi-weekly, or monthly. These variations can't be captured adequately by static machine learning methods. This underlines the significance of sequence modeling in crime research, especially in cities where crime trends are influenced by local events, law enforcement schedules, festivals, or economic cycles. On the other hand, the XGBoost model highlighted the effectiveness of decision-tree boosting methods in capturing interactions among features that temporal models might miss. Its ability to assess non-linear boundaries and prioritize features makes it particularly useful for handling structured data like time-related statistics, lagged variables, rolling windows, and spatial connections. The complementary roles of these two models—one focusing on time-dependent patterns and the other on overall spatial and structural relationships—support the choice to use a hybrid approach.

Another significant point of the research is the use of Isolation Forest for detecting anomalies. Unlike supervised models that depend strictly on past patterns, Isolation Forest views anomalies as rare, easily identifiable areas in the data. This approach allowed the system to spot sudden crime spikes, unusual clusters, or unexpected hotspots that were not strongly connected to previous behavior. This ability is crucial for real-world applications since law enforcement agencies need early alerts for abnormal situations, like festivals, protests, or sudden crime surges, that deviate from typical patterns. Thus, the anomaly detection system becomes an important part, enhancing the interpretability and responsiveness of the predictive model.

The implementation of a weighted ensemble further bolstered the system. It balanced the strengths and weaknesses of individual models. Ensemble learning made the prediction process more stable, reducing variance from the LSTM network and addressing model-specific biases inherent in XGBoost. At the same time, the anomaly detection feature improved predictions by introducing a risk measure shown as a "danger score." Through systematic testing and evaluation, the ensemble proved to be a more dependable indicator of crime likelihood than any single model on its own. A carefully chosen weighting scheme ensured that no single model dominated the ensemble, resulting in an effective final prediction mechanism that reflects a shared view from various machine-learning strategies.

Despite these advancements, the project recognizes several limitations in urban crime modeling. Crime data often suffers from underreporting, inconsistent quality, missing values, and differences in reporting frequency. In this study, methods for smoothing and normalization helped partly address these issues, but data quality still plays a crucial role in predictive performance. Additionally, while the LSTM model identifies long-term patterns, its effectiveness relies on having dense and continuous historical data. Cities with sparse or irregular crime reporting may not see the same predictive benefits. Likewise, XGBoost's dependence on handcrafted features shows promising results but also means that its success is closely linked to the choices made during feature creation. Improvements or mistakes in feature engineering can significantly impact model outcomes.

Another limitation is the geographic applicability of the system. The models were trained and validated using data from a single city, which restricts the ability to apply the findings directly to other cities with different crime patterns, demographics, population densities, or policing practices. While the methodology can be transferred, performance metrics may differ widely in other environments. Moreover, the anomaly detection model, despite its usefulness, may sometimes misinterpret normal fluctuations as anomalies, especially in periods or areas where crime activity naturally varies.

From a practical viewpoint, deploying this model in the real world brings extra challenges, such as data delays, computational demands, inter-department cooperation, and policy restrictions on predictive policing tools. The ethical implications of crime prediction are also a significant concern. If not designed or executed properly, crime forecasting models might cause algorithmic bias, unfair policing, or unjust targeting of certain communities. The study highlights that predictive models should always serve as support tools for decision-making rather than replace enforcement actions.

In conclusion, the study effectively shows a unified method for predicting crime risks by combining LSTM, XGBoost, and Isolation Forest models. This method represents a significant step in developing analytical systems that integrate time awareness, structured data learning, and anomaly detection. Through careful data processing, normalization, design, and evaluation, the project provides valuable insights into crime data behavior and the potential of machine learning for public safety. While there is room for improvement, the current results lay a strong foundation for future research and practical applications. By connecting clarity, predictive accuracy, and real-world utility, this study significantly contributes to the ongoing effort to enhance urban safety, law enforcement readiness, and resource distribution.

FUTURE WORK

Although the current system demonstrates strong predictive capabilities and a well-integrated architecture, several avenues exist for future expansion that could significantly enhance performance, adaptability, and real-world applicability. One of the most promising directions for future research lies in expanding the dataset beyond the current dependence on crime count and temporal statistics.

Crime is deeply intertwined with socio-economic and behavioural indicators such as employment rates, education levels, income distributions, urban mobility, real-time human activity, and environmental factors such as weather conditions or infrastructure density. Integrating such data sources, particularly mobility data from GPS traces, public transport logs, and mobile network congestion, could offer a richer contextual understanding of crime occurrence. In addition, the inclusion of geospatial data layers such as land use patterns, CCTV density, lighting infrastructure, and commercial area clusters would enhance the spatial accuracy and situational awareness of the model. These extensions would allow the system to transition from purely temporal and structural modeling to multi-dimensional crime forecasting.

A second area of potential development involves cross-city evaluation and regional generalization. Crime patterns vary significantly between densely populated metros, mid-sized towns, and smaller localities. Training and validating the framework across multiple cities would allow researchers to assess the generalizability and robustness of the ensemble in varied socio-economic and geographic contexts. Such cross-city studies could also reveal latent behavioural universals in crime distribution or show how different urban layouts influence crime mobility across neighbourhoods. Moreover, this direction would allow the creation of transfer-learning extensions in which knowledge learned from high-resource cities could be adapted to cities with insufficient data, thereby enabling fairer access to predictive technologies.

Scaling the model into a larger ensemble is another promising trajectory. Beyond LSTM and XGBoost, architectures such as Temporal Convolutional Networks (TCN), Transformer-based sequence models, Random Forests, CatBoost, and Graph Neural Networks (GNNs) could dramatically enhance the system's predictive expressiveness. For example, a GNN could capture adjacency relationships between regions, enabling the model to understand how crime diffuses through neighbouring wards. Transformer architectures could learn long-range temporal dependencies better than LSTMs, especially for datasets with complex seasonality. These additions would help create a more rich ensemble capable of capturing spatial, temporal, and semantic aspects of crime data simultaneously.

The deployment of the model in real-world settings also presents meaningful opportunities for advancement. A lightweight or edge-optimized version of the system could be built to run efficiently on low-power devices, such as local police servers, mobile surveillance units, or city ward monitoring stations. This would reduce latency, minimize reliance on centralized cloud infrastructure, and allow real-time crime risk scoring even in areas with unstable connectivity. Designing such edge-ready versions would require optimization methods such as model pruning, quantization, or knowledge distillation so that the predictive accuracy remains intact while memory and computational requirements are reduced. Additionally, integrating the prediction system with frontend dashboards, GIS mapping tools, or police control room software would enhance usability and promote data-driven decision-making.

Another compelling direction for future research involves the ethical and policy dimension of predictive crime modeling. Ensuring fairness, transparency, and accountability is essential for preventing misuse or unintentional bias reinforcement. Future work could incorporate bias-detection modules that monitor prediction disparities across demographic or geographic groups, enabling researchers to fine-tune the models to avoid discriminatory outcomes. Additionally, explainable-AI (XAI) techniques—such as SHAP values, counterfactual explanations, or attention-based interpretability—could be embedded into the system to help stakeholders understand the rationale behind predictions. This would be invaluable for building trust among law-enforcement officials, policymakers, and the general public.

Finally, the system can evolve into a real-time event-aware crime forecasting engine. By integrating streaming data pipelines—such as live 911 calls, CCTV anomaly feeds, public event schedules, or social media trend analyses—the model could dynamically adapt its predictions based on ongoing city conditions. This would transform the system from a static predictor into a continuously adapting situational-awareness tool capable of informing rapid response strategies, dynamic patrol routing, or crisis management operations.

In essence, while the current study establishes a strong foundation, the future of this work lies in broader data integration, algorithmic expansion, real-time deployment, fairness assurance, and scalable infrastructure development. These enhancements would elevate the system beyond research and into a practical, intelligent assistant for urban safety management.

REFERENCES

- [1] Albers Zumel, A., Tizzoni, M., & Campedelli, G. M. (2025). *Deep Learning for Crime Forecasting: The Role of Mobility at Fine-grained Spatiotemporal Scales*. *Journal of Quantitative Criminology*. <https://link.springer.com/article/10.1007/s10940-025-09629-3>
- [2] Mao, L., Du, W., Wen, S., Li, Q., Zhang, T., & Zhong, W. (2025). *Crime Forecasting: A Spatio-temporal Analysis with Deep Learning Models* (preprint). <https://arxiv.org/abs/2502.07465>
- [3] Fisk, N. R., Ng Kok Ming, M., & Shabrina, Z. (2025). *Advancing Spatiotemporal Prediction using Artificial Intelligence: Extending the Framework of Geographically and Temporally Weighted Neural Network (GTWNN)*. (preprint). <https://arxiv.org/abs/2503.22751>
- [4] Sun, Y., Chen, T., & Yin, H. (2022). *Spatial-Temporal Meta-path Guided Explainable Crime Prediction*. (preprint.) <https://arxiv.org/abs/2205.01901>
- [5] “Ada-GCNLSTM: An adaptive urban crime spatiotemporal prediction model.” (2025). *Journal of Safety Science and Resilience*. <https://www.sciencedirect.com/science/article/pii/S2666449625000052>
- [6] “Towards spatio-temporal crime events prediction.” (2023). *Multimedia Tools and Applications*. 83, 18721–18737. <https://link.springer.com/article/10.1007/s11042-023-16188-x>
- [7] “Crime Prediction based on Classification Approaches.” (2025). *Procedia Computer Science*. <https://www.sciencedirect.com/science/article/pii/S1877050925011962t>
- [8] Xiong, Y. (2025). *Research on Crime Occurrence Prediction Using Machine Learning Methods — Considering Four Types of Crime in Chicago*. *International Journal of New Developments in Engineering and Society*. <https://francispress.com/papers/18334>
- [9] Shinde, S. A., Shirke-Deshmukh, S., & Sonkamble, R. (2025). *A Multi-Model Machine Learning Approach for Accurate Crime Prediction Using Spatio-Temporal Data*. *International Journal of Latest Technology in Engineering, Management & Applied Science*. 14(4), 768-777. <https://www.ijltemas.in/submission/index.php/online/article/download/1938/1334/6410>
- [10] Singh, I., Bhandari, G., & Khatwani, S. (2025). *Predictive Crime Rate Analysis System*. *IJRASET Journal*. <https://www.ijraset.com/research-paper/predictive-crime-rate-analysis-system>

- [11] Sarkar, A., Mishra, A. K., Singh, U. P., Hassan, S., Munda, G. S., & Singh, P. (2025). *Real-Time Crime Prediction in India Using Machine Learning*. IJRASET. <https://www.ijraset.com/research-paper/real-time-crime-prediction-in-india-using-machine-learning>
- [12] Karkhur, S., & Dubey, R. (2025). *A Study on Predicting Crime Rates through Machine Learning*. IJRASET. <https://www.ijraset.com/research-paper/predicting-crime-rates-through-machine-learning>
- [13] “A study on the application of data mining-based crime prediction models in criminal justice.” (2024). *Journal of Combinatorial Mathematics and Combinatorial Computing*. <https://combinatorialpress.com/jmcc-articles/volume-123/a-study-on-the-application-of-data-mining-based-crime-prediction-models-in-criminal-justice/>
- [14] “Deep Learning Based Crime Prediction Models: Experiments and Analysis.” (2024). Preprint. <https://www.emergentmind.com/papers/2407.19324>
- [15] “Real-Time Urban Fire and Safety Risk Prediction using POI and Environmental Features.” (2025). (Fires / incident-risk forecasting analog as contextual reference.) <https://ouci.dntb.gov.ua/en/works/4a5Gy584/>
- [16] Mandalapu, S., Rao, P., & Reddy, H. (2024). *A systematic review of Machine Learning and Deep Learning approaches for crime prediction*. (Survey summarizing trends, datasets, challenges). <https://www.jetir.org/papers/JETIR2503439.pdf>
- [17] “Integrated Moving Average (ARIMA) and Deep Learning approaches for Crime Trend Forecasting.” (2024). (Journal: JETIR) <https://www.jetir.org/papers/JETIR2503439.pdf>
- [18] “Real-time crime hotspot forecasting using deep neural networks and spatio-temporal embedding.” (2024). (Preprint / conference proceedings) <https://www.mdpi.com/2673-4591/68/1/4>
- [19] “Machine learning in crime prediction: A systematic literature review.” (2023). *Journal of Ambient Intelligence and Humanized Computing*. <https://link.springer.com/article/10.1007/s12652-023-04530-y>
- [20] Brantingham, P., & Brantingham, P. (2018). *Theoretical Foundations of Predictive Policing*. (Often cited in predictive policing literature.) — background foundational work, referenced in many forecasting frameworks. (via cited in search as historical grounding) <https://link.springer.com/article/10.1007/s11042-023-16188-x>
- [21] Lum, K., & Isaac, W. (2016). *Challenges and Risks in Predictive Policing and Algorithmic Bias*. (Context discussed in forecasting debates — referenced in Zumel et al. 2025) <https://link.springer.com/article/10.1007/s11042-023-16188-x>
- [22] Malleson, N., & Andresen, M. A. (2015). *Spatio-temporal Crimes Modelling: The Role of Urban Mobility*. (Referenced in mobility-based forecasting studies.) <https://link.springer.com/article/10.1007/s10940-025-09629-3>
- [23] Hipp, J. R., & others (2019). *Human Mobility Patterns and Crime Prediction*. (mobility-crime correlation, referenced in deep learning + mobility literature) <https://link.springer.com/article/10.1007/s10940-025-09629-3>
- [24] Rosés, M., & transportation-mobility studies (2020). *Transport data and its correlation with urban safety and crime*. (Background reference in mobility-based crime forecasting) <https://link.springer.com/article/10.1007/s10940-025-09629-3>
- [25] Huang, H., & colleagues (2018). *GPS-based Human Movement Analysis for Public Safety*. (Referenced in mobility & crime ML literature)