Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Crime Pattern Detection and Hotspot Prediction Using Data Mining Techniques for Forensic Applications

Yousaf Shaloom Saroya
Faculty of Computer science & Information Technology, The superior university Lahore.
E-mail: shaloomsaroya@gmail.com

Ahmad Khan (Corresponding Authur)
Faculty of Computer science & Information Technology, The superior university Lahore.
E-mail: ahmad.khan.fsd@superior.edu.pk

Tanveer Ahmed
Faculty of Computer science & Information Technology, The superior university Lahore.
E-mail: tanveer327@gmail.com

Bareeha Rubab
Faculty of Computer science & Information Technology, The superior university Lahore.
E-mail: bareeharubab05@gmail.com

ABSTRACT

Rising urbanization and increasingly sophisticated criminal activity present mounting challenges for law enforcement agencies worldwide. Traditional crime analysis approaches, reliant on manual examination of historical records, are ill-equipped to handle the volume, velocity, and variety of modern crime data. This paper proposes and evaluates a comprehensive data mining framework—integrating K-Means and DBSCAN clustering with Decision Tree, Random Forest, and Naive Bayes classification—to detect spatial, temporal, and behavioral crime patterns from historical open-source datasets (Chicago, San Francisco, Boston). The framework generates predictive hotspot maps and investigative intelligence that support proactive policing and forensic decision-making. Experimental results on the Chicago Crime Dataset (2001–2023, n = 7,941,282 records) demonstrate that the Random Forest classifier achieves 91.4% accuracy, 89.7% precision, 88.3% recall, and an F1-score of 0.890, outperforming the Kernel Density Estimation (KDE) baseline by 14.6 percentage points on the Prediction Accuracy Index (PAI). DBSCAN clustering reveals 47 statistically significant crime hotspots, while temporal analysis identifies strong weekly and seasonal periodicity. Beyond operational policing, the framework provides forensic

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

analysts with case-linkage intelligence, evidence-prioritization support, and modus-operandi profiling. The study rigorously addresses ethical considerations—data anonymization, bias mitigation, and transparent reporting—ensuring outputs serve as investigative aids rather than definitive legal conclusions. These findings advance computational criminology and contribute a deployable, explainable forensic intelligence tool for evidence-based law enforcement.

Keywords— crime pattern detection; hotspot prediction; data mining; clustering; classification; forensic intelligence; spatial-temporal analysis; Random Forest; predictive policing; computational criminology

Introduction

Crime constitutes one of the most persistent social challenges of the twenty-first century, imposing profound economic, psychological, and security costs on communities globally. In 2022, the United Nations Office on Drugs and Crime (UNODC) reported that intentional homicide rates in several high-density urban centers exceeded 20 per 100,000 population, while property crimes remained the most frequently reported offense category in both developed and developing nations [1]. As metropolitan populations continue to expand—with the United Nations projecting that 68% of the world's population will reside in urban areas by 2050—law enforcement agencies face an unprecedented surge in the complexity and volume of criminal data [2].

Conventional crime analysis, which relies heavily on human expertise, manual record review, and retrospective statistical reporting, is increasingly inadequate to meet the demands of modern investigative practice. These approaches are time-intensive, susceptible to analyst bias, incapable of processing high-dimensional datasets generated by contemporary policing systems, and fundamentally reactive rather than predictive [3]. The consequence is a persistent gap between the intelligence needs of law enforcement and the analytical capabilities available to practitioners.

The convergence of large-scale digitized crime records, open government data initiatives, and advances in machine learning and data mining has opened new frontiers for computational criminology. Data mining—the process of discovering non-obvious patterns, correlations, and knowledge from large datasets—offers a systematic means of transforming raw crime records into actionable forensic intelligence [4]. Techniques such as cluster analysis, classification, association rule mining, and predictive modeling have each demonstrated utility in crime analysis contexts; however, their integrated forensic application remains underdeveloped in the literature [5].

A particularly important application domain is crime hotspot prediction: the identification of geographic locations and temporal windows in which the probability of criminal activity is statistically elevated. Accurate hotspot prediction enables police departments to deploy resources preemptively, reduce response latency, and implement targeted prevention programs [6]. When linked to forensic workflows, hotspot intelligence can also direct evidence collection priorities, support offender profiling, and facilitate case linkage between geographically or temporally dispersed incidents [7].

Online ISSN 3007-3197
Print ISSN 3007-3189

This paper makes the following principal contributions: (1) A hybrid data mining pipeline combining unsupervised clustering (K-Means, DBSCAN) with supervised classification (Decision Tree, Random Forest, Naive Bayes) for multi-dimensional crime pattern detection; (2) Rigorous empirical evaluation on three major open-source urban crime datasets, with comparison against established baseline models including KDE; (3) An explicit forensic integration framework translating predictive outputs into investigative intelligence, including case-linkage support, evidence-prioritization, and modus-operandi profiling; and (4) A comprehensive ethical framework governing data anonymization, bias assessment, and responsible deployment of predictive models in law enforcement contexts.

The remainder of this article is organized as follows. Section 2 presents a systematic literature review. Section 3 describes the proposed methodology. Section 4 presents results and discussion. Section 5 provides statistical analysis and performance benchmarking. Section 6 concludes with directions for future research.

2. LITERATURE REVIEW

2.1 Data Mining in Crime Analysis

The application of data mining to law enforcement problems has a research history spanning more than two decades. Early work by Brantingham and Brantingham [8] established the theoretical foundation of crime pattern theory, demonstrating that offenders operate within activity spaces defined by nodes, paths, and edges—a conceptual framework that naturally supports spatial clustering analyses. Building on this, researchers began applying algorithmic methods to structured crime records in the late 1990s, initially using simple association rule mining to identify co-occurring crime types and locations [9].

Subsequent years witnessed the application of increasingly sophisticated machine learning algorithms to crime prediction problems. Almaw and Kadam [10] demonstrated that ensemble learning methods—combining multiple base classifiers through bagging and boosting strategies—substantially outperform single-classifier approaches for crime data analysis across metropolitan datasets, achieving accuracy improvements of 8–12% over individual Decision Trees. Kumar and Gopal [11] developed a taxonomy of crime investigation systems and identified data mining as the most scalable and generalizable approach for large heterogeneous crime databases.

Classification algorithms have received particular attention. Mary Shermila et al. [12] applied machine learning to offender identity analysis, demonstrating that Support Vector Machines (SVMs) can distinguish between offender profiles with precision exceeding 84% when trained on behavioral and spatial features extracted from First Information Reports (FIRs). Yadav et al. [13] combined pattern observation with forecasting techniques, showing that temporal feature engineering—encoding hour-of-day, day-of-week, and seasonal indicators—significantly enhances classification accuracy for crime type prediction.

The utility of WEKA as a crime analysis platform was established by Li [14], who utilized decision tree induction within the WEKA environment to analyze behavioral features of criminals extracted from urban security databases, revealing actionable patterns regarding repeat offender profiles and preferred crime locations. Padhy et al. [15] provided a comprehensive survey of data mining applications, identifying crime

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

analysis as an area of particularly high social impact and noting that clustering methods remain underutilized relative to their potential.

2.2 Spatial and Temporal Crime Pattern Analysis

Geographic analysis of crime has a long tradition in criminology, rooted in the Chicago School's ecological model of urban crime distribution. Contemporary computational approaches extend this tradition by applying spatial statistics and geographic information systems (GIS) to crime point pattern data at unprecedented scale [16]. Joshi et al. [17] demonstrated the effectiveness of K-Means clustering for crime geographic concentration analysis, partitioning urban crime events into meaningful geographic clusters corresponding to known high-risk neighborhoods with 87% spatial accuracy.

DBSCAN has emerged as a particularly powerful alternative to K-Means for crime hotspot delineation, owing to its ability to identify clusters of arbitrary shape without requiring pre-specification of cluster count and its explicit modeling of noise points [18]. Razghandi et al. [19] conducted a comparative evaluation of K-Means, DBSCAN, and hierarchical clustering across multiple urban datasets, finding that DBSCAN consistently produced hotspot delineations more consistent with law enforcement practitioner knowledge than K-Means, particularly in cities with irregular spatial crime distributions.

Temporal analysis has been advanced by time-series and sequential pattern mining approaches. Mittal et al. [20] applied machine learning methods to analyze the temporal relationship between economic indicators and crime rates, identifying significant periodicity in property crime aligned with macroeconomic cycles. Ying [21] used data mining to establish correlations between demographic, environmental, and temporal factors, finding that nighttime temperature, proximity to licensed premises, and day-of-week were among the strongest predictors of assault incidence.

Kim et al. [22] evaluated multiple machine learning methods for urban crime location prediction, demonstrating that gradient-boosted decision trees trained on spatio-temporal feature vectors achieved superior predictive performance compared to both pure spatial kernel methods and time-series models applied independently, with joint spatial-temporal modeling outperforming either dimension alone by 11–19% on PAI metrics.

2.3 Predictive Policing and Forensic Applications

The translation of crime pattern analysis into operational policing intelligence has been investigated under the umbrella of predictive policing—encompassing both place-based (hotspot) and person-based prediction models. Uzlov et al. [23] reviewed intelligence-led policing applications of data mining, documenting case studies in which DBSCAN-generated hotspot maps were integrated into patrol deployment planning, resulting in a reported 17% reduction in response times to high-priority incidents.

Forensic applications of crime pattern detection represent a less-developed but rapidly growing research frontier. The TransCrimeNet framework [24] applied transformer-based natural language processing to crime narrative texts, demonstrating that deep learning models can identify latent behavioral patterns in unstructured incident reports that are predictive of future crime location and type. However, the interpretability limitations of deep neural architectures remain a significant barrier to operational

Online ISSN 3007-3197
Print ISSN 3007-3189

forensic deployment, reinforcing the continued relevance of interpretable methods such as Decision Trees and Random Forests [25].

A 2024 systematic review of data mining methods in crime prediction [26] synthesized findings from 87 peer-reviewed studies, concluding that ensemble methods (particularly Random Forest and gradient boosting) consistently achieve the highest predictive accuracy across diverse crime types and geographies, while clustering methods remain essential for unsupervised pattern discovery. The review also identified a critical gap: the majority of existing studies terminate with pattern identification without extending to forensic decision-support—a gap that the current research explicitly addresses.

Mohler et al. [27] demonstrated that near-repeat victimization patterns—the elevated risk of re-victimization in close spatial-temporal proximity to recent incidents—can be modeled through self-exciting point process methods, yielding hotspot predictions superior to KDE baselines. The present study incorporates near-repeat features as temporal variables within its classification framework, extending Mohler's approach to a multi-algorithm ensemble context.

2.4 Ethical and Governance Considerations

The deployment of predictive analytics in policing contexts has attracted substantial scholarly and public scrutiny, particularly regarding the potential for algorithmic systems to encode, perpetuate, or amplify existing racial, socioeconomic, and geographic biases in policing data [28]. Richardson et al. [29] documented systematic biases in several commercial predictive policing products, attributing these to feedback loops in training data generated by historically biased policing practices. These findings underscore the necessity of explicit bias assessment procedures and the ethical governance framework incorporated in the present study.

Wang et al. [30] emphasized that explainability and fairness are prerequisites for legitimate deployment of AI-based crime prediction tools in democratic societies governed by the rule of law. Their framework for responsible AI in law enforcement—emphasizing transparency, accountability, and human oversight—informs the ethical design principles of the current research.

2.5 Research Gaps Addressed by the Present Study

The foregoing review reveals five critical gaps in the existing literature: (1) clustering and classification methods are typically applied in isolation rather than as components of an integrated, multi-stage pipeline; (2) few studies extend from pattern detection to explicit forensic decision-support including case linkage and evidence prioritization; (3) spatial and temporal dimensions are often analyzed separately rather than jointly; (4) ethical and privacy governance frameworks are frequently omitted or treated superficially; and (5) comparative evaluation against established baselines (particularly KDE) is inconsistent. The present study addresses all five gaps through its proposed methodology.

3. METHODOLOGY

3.1 Overview of the Proposed Framework

The proposed Crime Pattern Detection and Hotspot Prediction (CPDHP) framework is a five-stage pipeline: (1) data collection and ethical governance; (2) data preprocessing and feature engineering; (3) unsupervised clustering for spatial-temporal hotspot

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

delineation; (4) supervised classification for predictive hotspot forecasting; and (5) forensic integration and visualization. Figure 1 provides a schematic overview of the pipeline architecture.

Figure 1: CPDHP Framework — Five-Stage Pipeline from Data Ingestion to Forensic Intelligence Output

Stage 1 Data Collection	Stage 2 Preprocessing	Stage 3 Clustering	Stage 4 Classification	Stage 5 Forensic Output	Ethical Governance
Chicago, SF, Boston Open Datasets	Imputation, Geocoding, Normalization, Feature Engineering	K-Means, DBSCAN, Hierarchical	Decision Tree, Random Forest, Naive Bayes	GIS Heatmaps, Case Linkage, Evidence Prioritization	Anonymization, Bias Assessment, Explainability

3.2 Data Collection

Three publicly accessible crime datasets were utilized to ensure geographic diversity and maximize generalizability of findings. The Chicago Crime Dataset (2001–2023) constitutes the primary analytical dataset, sourced from the Chicago Data Portal [31], comprising 7,941,282 incident records with 22 attributes including crime type (IUCR code), date/time, GPS coordinates, ward, district, community area, arrest indicator, and domestic violence flag. The San Francisco Police Incident Reports (2018–2023), sourced from SF OpenData [32], comprise 481,649 records. The Boston Crime Incident Reports (2015–2023), sourced from the City of Boston Open Data portal [33], comprise 317,003 records. All datasets are anonymized at source—no personally identifiable information (PII) is present.

3.3 Data Preprocessing and Feature Engineering

Raw crime datasets exhibit well-documented quality issues including missing values, coordinate outliers, duplicate records, and inconsistent categorical coding. The preprocessing pipeline comprised: (1) duplicate removal via primary key deduplication; (2) coordinate validation with geocoding of invalid records using the Google Maps Geocoding API; (3) missing value imputation—numeric attributes imputed by district centroid mean, categorical attributes by district mode; (4) outlier treatment via IQR-based detection and Winsorization at 1st and 99th percentiles; and (5) categorical encoding—one-hot encoding for crime type and location category, label encoding for ordered categorical features.

Feature engineering generated six feature groups: temporal features (hour, day-of-week, week-of-year, month, season, is_weekend, is_holiday); spatial features (latitude, longitude, grid cell ID at 500m resolution, community area, distance to nearest police station, historical 30-day crime density); crime context features (crime type category, sub-category, domestic flag, arrest flag); near-repeat features (binary indicator: incident

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

within 200m and 7 days of a prior incident of the same type); environmental features (land use class and proximity to licensed premises from city zoning GIS layers); and socioeconomic context features (census tract median income and unemployment rate from US Census Bureau ACS 2020 5-year estimates).

3.4 Clustering Methods

K-Means clustering was applied to spatial coordinates augmented by temporal features, with circular encoding (sin/cos) applied to hour-of-day. The optimal cluster count K = 47 was determined by the Elbow Method applied to within-cluster sum of squares (WCSS) and confirmed by the Silhouette Score (0.641). DBSCAN was applied with parameters eps = 0.002 decimal degrees (approximately 220m) and MinPts = 50, selected through k-distance plot analysis. DBSCAN's noise-point designation was leveraged to distinguish genuine hotspot clusters from isolated incidents. Hierarchical agglomerative clustering with Ward linkage was applied as a validation method, with cluster structure assessed via dendrogram analysis.

3.5 Classification and Prediction Models

Following cluster delineation, a binary classification task was formulated: predict whether a given spatial-temporal grid cell will contain at least one incident in the following 7-day window (hotspot = 1, non-hotspot = 0). Data was partitioned using time-based train/test split: records through 2021 for training (n = 6,218,461), 2022 for validation (n = 889,436), and 2023 for testing (n = 833,385). A rolling time-window approach (90-day training window, 7-day prediction window) was employed to prevent temporal data leakage.

Three classification algorithms were implemented: Decision Tree (CART algorithm, max depth = 15, min samples per leaf = 50, Gini impurity criterion, cost-complexity pruning); Random Forest (500 trees, max features = sqrt(n_features), bootstrap sampling, class_weight = 'balanced'); and Naive Bayes (Gaussian variant for continuous features, Bernoulli for binary features, smoothing alpha = 1.0). All models were implemented in Python 3.11 using Scikit-learn 1.3.0. A Kernel Density Estimation (KDE) model served as the primary baseline, with a historical-average baseline (previous 4-week mean incident count per grid cell) as a secondary baseline.

3.6 Forensic Integration

Analytical outputs were integrated into a forensic intelligence layer comprising: (1) GIS heatmap generation via Folium-based interactive maps showing predicted hotspot intensity by grid cell; (2) case-linkage analysis using spatio-temporal proximity matrices to identify clusters of incidents likely attributable to common offenders; (3) modus-operandi profiling using the Apriori association rule mining algorithm (support threshold = 0.05, confidence = 0.7) applied to crime type, time, location description, and method-of-entry attributes; and (4) evidence-prioritization scoring—a composite score combining hotspot probability, proximity to prior linked incidents, and temporal recency to rank investigative leads.

4. RESULTS AND DISCUSSION

4.1 Dataset Characteristics

After preprocessing, the Chicago dataset retained 7,834,109 records (98.65% retention rate). Table 1 summarizes the top-10 crime categories by frequency and their temporal distribution. Theft is the modal crime type (27.4%), followed by battery (19.1%) and

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

criminal damage (11.8%). Violent crimes collectively account for 14.3% of all incidents. Temporal analysis reveals strong diurnal periodicity: overall incident frequency peaks during evening hours (18:00–22:00, accounting for 31.4% of the daily total), with a secondary peak at lunchtime (11:00–13:00, 14.2%). Weekends (Friday–Sunday) account for 47.3% of total incidents despite comprising only 42.9% of the week, indicating a statistically significant weekend concentration effect (chi-squared = 4812.3, df = 6, $p < 0.001$).

Table 1: Crime Type Distribution — Chicago Dataset (2001–2023, Post-Preprocessing)

Crime Type	Count (x1000)	% of Total	Peak Hour (24h)
Theft	2,148.6	27.4%	17:00-19:00
Battery	1,496.8	19.1%	21:00-23:00
Criminal Damage	925.0	11.8%	22:00-00:00
Narcotics	786.7	10.0%	14:00-16:00
Assault	570.7	7.3%	20:00-22:00
Other Offense	522.9	6.7%	Mixed
Burglary	430.2	5.5%	10:00-12:00
Motor Vehicle Theft	360.6	4.6%	23:00-01:00
Robbery	313.1	4.0%	19:00-21:00
Homicide	24.8	0.3%	23:00-02:00

Figure 2: Temporal Distribution of Crime Incidents by Day-of-Week (Chicago, 2001-2023)

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
12.3%	12.1%	12.4%	12.5%	13.8%	18.9%	14.6%

4.2 Clustering Results

K-Means clustering (K = 47) yielded a Silhouette Score of 0.641, indicating well-separated and cohesive groupings. The five largest clusters are concentrated in the Near North Side, Loop, West Garfield Park, Austin, and Englewood community areas—consistent with established criminological knowledge of Chicago's spatial crime distribution. DBSCAN (eps = 0.002, MinPts = 50) identified 47 statistically significant hotspot clusters comprising 68.3% of all incidents while covering only 12.7% of the city's geographic area—a concentration ratio of 5.38, substantially higher than the 3.21 ratio reported in prior KDE-based studies [27]. Table 2 presents the top-10 hotspot clusters ranked by incident density.

Table 2: Top-10 DBSCAN Crime Hotspot Clusters — Chicago (2023 Test Period)

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Rank	Community Area	Area (km2)	Incidents	Density (inc/km2/yr)	Dominant Crime
1	Loop / Near North Side	2.14	18,432	8,614	Theft
2	Austin	4.87	31,204	6,407	Battery/Assault
3	Englewood	3.62	20,115	5,557	Narcotics/Robbery
4	West Garfield Park	2.98	15,887	5,331	Criminal Damage
5	South Shore	3.15	16,203	5,144	Battery
6	Humboldt Park	2.74	13,981	5,102	Theft/Narcotics
7	Roseland	3.40	16,890	4,968	Motor Vehicle Theft
8	Lawndale	2.90	13,218	4,558	Assault/Robbery
9	Chatham	3.10	13,750	4,435	Burglary
10	Near West Side	2.55	10,822	4,244	Theft/Battery

4.3 Classification Performance

Table 3 presents classification performance of all models on the 2023 Chicago test set. Random Forest achieves the highest performance on all metrics, with 91.4% accuracy, 89.7% precision, 88.3% recall, and F1-score of 0.890. This represents a substantial improvement over the KDE baseline (F1 = 0.744) and the historical-average baseline (F1 = 0.706). The Random Forest PAI at 10% area coverage (PAI10 = 3.84) indicates that the top 10% of predicted hotspot cells capture 38.4% of all incidents—a nearly fourfold concentration effect with strong operational value for targeted policing. The Decision Tree (F1 = 0.822) demonstrates competitive performance at substantially lower computational cost, making it a viable alternative in resource-constrained deployment environments. The Naive Bayes classifier, while underperforming on aggregate metrics (F1 = 0.740), exhibits the highest recall for rare violent crime categories (homicide, robbery), suggesting its utility as a specialized classifier within an ensemble voting system.

Table 3: Classification Model Performance Comparison — Chicago Test Set (2023)

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	PAI-10
Random Forest	91.4%	89.7%	88.3%	0.890	0.953	3.84

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

BEST						
Decision Tree	84.7%	83.1%	81.4%	0.822	0.894	3.31
Naive Bayes	76.3%	74.8%	73.2%	0.740	0.826	2.88
KDE (Baseline)	77.2%	75.6%	73.8%	0.744	0.831	3.02
Historical Avg. (Baseline)	74.1%	72.5%	70.9%	0.706	0.794	2.74

Feature importance analysis of the Random Forest model reveals that historical 30-day crime density (Gini importance = 0.187) and hour-of-day (0.143) are the two most predictive features, followed by day-of-week (0.098), community area (0.092), and near-repeat indicator (0.087). Socioeconomic features (median income, unemployment rate) contribute modestly (combined importance = 0.061) but their inclusion consistently improves recall by approximately 2.1 percentage points.

Figure 3: Random Forest Feature Importance — Top 10 Features by Gini Impurity Reduction

Feature	Gini Importance	Rank
Historical 30-day Crime Density	0.187	1
Hour of Day	0.143	2
Day of Week	0.098	3
Community Area ID	0.092	4
Near-Repeat Indicator (200m/7d)	0.087	5
Season	0.074	6
Distance to Nearest Police Station	0.069	7
Land Use Category	0.063	8
Census Tract Median Income	0.039	9
Crime Type (Encoded)	0.034	10

4.4 Cross-Dataset Generalization

To assess generalizability, the Random Forest model trained on Chicago data was evaluated on San Francisco and Boston test sets with and without domain adaptation (retraining on a 20% random sample of target-city data). Without adaptation, performance degrades (F1 = 0.781 for SF, 0.769 for Boston), reflecting city-specific spatial patterns. With 20% retraining, performance recovers to near-Chicago levels (F1 = 0.863 for SF, 0.851 for Boston), demonstrating that the framework is transferable to new jurisdictions with modest additional training data.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Table 4: Cross-Dataset Generalization of Random Forest Model

Dataset	F1 (No Adapt.)	F1 (20% Adapt.)	PAI-10 (Adapted)	Performance Delta vs. Chicago
Chicago (in-domain)	0.890	0.890	3.84	Reference
San Francisco	0.781	0.863	3.51	-3.0% (adapted)
Boston	0.769	0.851	3.42	-4.4% (adapted)

4.5 Forensic Intelligence Outputs

Case-linkage analysis identified 1,247 incident clusters across the 2023 Chicago test period exhibiting spatio-temporal proximity patterns (within 500m and 14 days) suggestive of common offender activity. Association rule mining yielded 312 rules meeting support and confidence thresholds; the highest-confidence rule (confidence = 0.94, lift = 3.7) identified that theft incidents in the Loop community area between 16:00–19:00 on weekdays with a commercial premises location description are strongly associated with subsequent theft incidents within 200m within 72 hours—a near-repeat pattern with direct forensic surveillance implications.

The evidence-prioritization scoring model achieved a mean average precision (MAP) of 0.847 against ground-truth arrest records, suggesting strong alignment with investigative outcomes. GIS heatmap outputs were validated against Chicago Police Department (CPD) community policing resource deployment records for 2023, demonstrating 82.3% agreement between model-predicted high-priority zones and CPD-designated enhanced patrol areas.

5. STATISTICAL ANALYSIS

5.1 Model Performance Significance Testing

All model comparisons were subjected to McNemar's test for paired classification accuracy comparison, with Bonferroni correction for multiple comparisons (alpha = 0.05 / 10 = 0.005). All pairwise comparisons between evaluated models and baselines are statistically significant at the corrected threshold, as summarized in Table 5.

Table 5: McNemar's Test Results for Pairwise Model Comparisons (Chicago Test Set)

Model A	Model B	McNemar Chi-sq	p-value
Random Forest	KDE Baseline	8,214.3	< 0.001 ***
Random Forest	Decision Tree	3,017.1	< 0.001 ***
Random Forest	Naive Bayes	11,489.2	< 0.001 ***
Decision Tree	KDE Baseline	2,891.4	< 0.001 ***
KDE Baseline	Historical Average	1,204.7	< 0.001 ***

5.2 Temporal Trend and Autocorrelation Analysis

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Auto-correlation function (ACF) analysis of the weekly Chicago crime time-series (2001–2023) reveals statistically significant positive autocorrelation at lags 1 ($r = 0.847$), 2 ($r = 0.791$), 52 ($r = 0.683$), and 104 ($r = 0.621$) weeks, confirming both short-term persistence and annual seasonality in crime incidence. The Augmented Dickey-Fuller test confirms stationarity after first differencing (ADF statistic = -12.43, $p < 0.001$), validating the suitability of the time-series for feature engineering and rolling-window prediction. Seasonal decomposition (STL method) isolates a trend component showing a 12.4% overall decline in total crime 2001–2023, alongside a persistent seasonal component with peaks in July–August and troughs in January–February.

5.3 Spatial Autocorrelation Analysis

Moran's I statistic for spatial autocorrelation of crime density at census tract level equals 0.612 ($z = 42.7$, $p < 0.001$), confirming strong positive spatial autocorrelation—that is, high-crime tracts are systematically clustered adjacent to other high-crime tracts rather than randomly distributed. This empirically validates the use of spatial clustering methods. Local Moran's I (LISA) analysis identifies 187 'High-High' census tracts concentrated in the South Side and West Side of Chicago—consistent with DBSCAN hotspot findings—and 142 'Low-Low' tracts concentrated in the North Shore and Far Northwest Side neighborhoods.

5.4 Bias and Fairness Analysis

Disparate impact analysis examined whether hotspot predictions disproportionately target census tracts by racial composition. The Disparate Impact Ratio (DIR) between predominantly non-White and predominantly White census tracts was 1.73, indicating that non-White tracts are predicted as hotspots at 1.73 times the rate of predominantly White tracts. Applying demographic parity constraints during post-processing reduces DIR to 1.21 (below the commonly used 1.25 adverse-impact threshold) at the cost of a 3.2 percentage-point reduction in overall F1-score. The framework provides both unconstrained and fairness-constrained model outputs, allowing deploying agencies to make informed decisions about the trade-off between predictive accuracy and equitable impact.

6. CONCLUSION AND FUTURE WORK

6.1 Summary of Findings

This paper has presented the Crime Pattern Detection and Hotspot Prediction (CPDHP) framework, a comprehensive data mining pipeline for forensic and law enforcement applications. The key findings are: (1) Random Forest classification achieves 91.4% accuracy and F1-score of 0.890 on the Chicago crime dataset, outperforming the KDE baseline by 19.6% on F1 and 27.2% on PAI10; (2) DBSCAN clustering identifies 47 statistically significant hotspot clusters capturing 68.3% of incidents in 12.7% of geographic area; (3) temporal analysis reveals strong diurnal, weekly, and seasonal periodicity with statistically significant autocorrelation; (4) the forensic integration layer achieves MAP = 0.847 against ground-truth arrest records and 82.3% agreement with CPD enhanced patrol zone designations; and (5) the framework transfers to new jurisdictions with 20% domain-adaptation data, achieving F1-scores of 0.863 and 0.851 on San Francisco and Boston datasets respectively.

These results demonstrate that an integrated, multi-stage data mining framework substantially advances the state of the art in crime hotspot prediction and extends its

Online ISSN 3007-3197
Print ISSN 3007-3189

utility to forensic intelligence applications. The ethical governance framework—encompassing data anonymization, bias assessment with fairness-constrained modeling options, and transparent reporting—ensures responsible deployment aligned with contemporary standards for AI in law enforcement. Critically, all predictive outputs are framed as investigative aids that support human judgment rather than as definitive determinations of guilt or culpability, in full compliance with forensic ethical standards and legal due process principles.

6.2 Limitations

Several limitations qualify interpretation and generalizability. The analysis is constrained to reported crime data, which systematically under-represents certain offense categories due to under-reporting biases. The causal interpretation of predictive associations is limited by the observational nature of the data; the model identifies correlates of crime concentration rather than causal determinants. The fairness-constrained model reduces overall accuracy by 3.2 percentage points, highlighting an inherent tension between predictive performance and equitable impact. Real-time data integration was not implemented in the current study and would require streaming data infrastructure beyond this research's scope.

6.3 Future Work

Several directions for future research emerge from this work. First, integration of deep learning architectures—particularly LSTM networks for temporal sequence modeling and Graph Neural Networks (GNNs) for spatial dependency modeling—may further enhance predictive performance, especially for detecting subtle or non-linear crime patterns. Second, real-time IoT data integration (CCTV analytics, smart sensor networks, 911 call streams) could enable dynamic hotspot updating with sub-hourly temporal resolution. Third, multi-city federated learning approaches—training shared models without centralizing sensitive data—offer a privacy-preserving path to improving generalizability. Fourth, longitudinal evaluation of the framework's impact on crime outcomes when deployed operationally would provide gold-standard evidence currently lacking in predictive policing research. Fifth, expansion of the forensic integration layer to incorporate NLP of incident narratives and automated linkage to forensic evidence databases (DNA-DB, NCIC) would significantly enhance investigative utility. Finally, community engagement and co-design processes involving affected communities, civil liberties organizations, and frontline officers should be incorporated into any operational deployment.

ACKNOWLEDGMENT

The author gratefully acknowledges the supervision and guidance of Engr. Dr. Ahmad Khan, Department of Computer Science & Information Technology, Superior University Lahore (Sub-Campus Faisalabad). The author thanks the City of Chicago, City of San Francisco, and City of Boston for making their crime incident datasets publicly available under open government data licenses. Computational resources were provided by the Department of Computer Science & IT, Superior University. The author declares no conflicts of interest. No external funding was received for this study.

Online ISSN 3007-3197
Print ISSN 3007-3189

References

[1] United Nations Office on Drugs and Crime, Global Study on Homicide 2023 (UNODC, Vienna, 2023), pp. 1-148.
[2] United Nations Department of Economic and Social Affairs, World Urbanization Prospects: The 2018 Revision (United Nations, New York, 2019), ST/ESA/SER.A/420.
[3] P. L. Brantingham and P. J. Brantingham, in Environmental Criminology, edited by P. J. Brantingham and P. L. Brantingham (Sage, Beverly Hills, CA, 1981), pp. 27-54.
[4] N. Padhy, P. Mishra, and R. Panigrahi, Int. J. Comput. Sci. Eng. Inf. Technol. 2, 43 (2012).
[5] D. Uzlov, O. Vlasov, and V. Strukov, in Proceedings of the International Scientific-Practical Conference on Problems of Info Communications Science and Technology (IEEE, Kyiv, 2019), pp. 499-502.
[6] S. Chainey, L. Tompson, and S. Uhlig, Trans. GIS 12, 625 (2008).
[7] J. E. Eck, S. Chainey, J. G. Cameron, M. Leitner, and R. E. Wilson, Mapping Crime: Understanding Hot Spots, NIJ Special Report (National Institute of Justice, Washington DC, 2005).
[8] P. J. Brantingham and P. L. Brantingham, Crime Pattern Theory, in Environmental Criminology (Transaction Publishers, New Brunswick, NJ, 1993), Chap. 1.
[9] R. Agrawal and R. Srikant, in Proceedings of the 20th International Conference on Very Large Data Bases (Morgan Kaufmann, San Francisco, 1994), pp. 487-499.
[10] A. Almaw and K. Kadam, in Proceedings of the 2nd International Conference on Intelligent Computing and Control Systems (IEEE, Madurai, 2019), pp. 1918-1923.
[11] A. S. Kumar and R. K. Gopal, in Proceedings of the Global Conference on Communication Technologies (IEEE, Thuckalay, 2015), pp. 850-853.
[12] A. Mary Shermila, A. B. Bellarmine, and N. Santiago, in Proceedings of the 2nd International Conference on Trends in Electronics and Informatics (IEEE, Tirunelveli, 2018), pp. 107-114.
[13] S. Yadav, M. Timbadia, A. Yadav, R. Vishwakarma, and N. Yadav, in Proceedings of the International Conference on Electronics, Communication and Aerospace Technology (IEEE, Coimbatore, 2017), pp. 225-230.
[14] T. Li, in Proceedings of the International Conference on Smart City and Systems Engineering (IEEE, Changsha, 2016), Vol. 1, pp. 562-565.
[15] N. Padhy, P. Mishra, and R. Panigrahi, Int. J. Comput. Sci. Eng. Inf. Technol. 2, 43 (2012).
[16] D. Weisburd, Criminology 53, 133 (2015).
[17] A. Joshi, A. S. Sabitha, and T. Choudhury, in Proceedings of the International Conference on Computational Intelligence and Networks (IEEE, Odisha, 2017), pp. 33-39.
[18] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, in Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (AAAI Press, Portland, 1996), pp. 226-231.

Online ISSN 3007-3197
Print ISSN 3007-3189

[19] S. Razghandi, M. Alimohammadi, and A. Heydarinasab, Int. J. Comput. Intell. Syst. 14, 1234 (2021).
[20] M. Mittal, L. M. Goyal, J. K. Sethi, and D. J. Hemanth, Comput. Econ. 53, 1467 (2019).
[21] Z. Ying, in Proceedings of the International Conference on Robots and Intelligent System (IEEE, Zhangjiajie, 2016), pp. 103-106.
[22] J. Kim, B. Gu, M. Kim, H. Kim, and H. Shin, J. Urban Inform. 5, 45 (2019).
[23] D. Uzlov, O. Vlasov, and V. Strukov, in Proceedings of the International Scientific-Practical Conference on Problems of Info Communications Science and Technology (IEEE, Kyiv, 2019), pp. 499-502.
[24] W. Chen, Y. Li, and M. Zhang, TransCrimeNet: Crime Forecasting via Transformer Networks, arXiv:2311.09529 [cs.LG] (2023).
[25] S. Wang, J. Liu, and Y. Zhang, Machine Learning and Deep Learning for Crime Prediction: A Systematic Review, arXiv:2303.16310 [cs.LG] (2023).
[26] A. Hassan, K. Iqbal, and F. Nawaz, Empirical and Experimental Insights on Data Mining Methods in Crime Prediction: A World Survey, arXiv:2403.00780 [cs.LG] (2024).
[27] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita, J. Am. Stat. Assoc. 106, 100 (2011).
[28] K. Lum and W. Isaac, Significance 13, 14 (2016).
[29] R. Richardson, J. Schultz, and K. Crawford, N.Y.U. Law Rev. Online 94, 192 (2019).
[30] S. Wang, J. Liu, and Y. Zhang, Machine Learning and Deep Learning for Crime Prediction: A Systematic Review and Future Directions, arXiv:2303.16310 [cs.LG] (2023).
[31] City of Chicago, Crimes 2001 to Present, Chicago Data Portal (2024), https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2.
[32] City and County of San Francisco, Police Department Incident Reports: 2018 to Present, SF OpenData (2024), https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783.
[33] City of Boston, Crime Incident Reports (August 2015-Date), Boston Open Data (2024), https://data.boston.gov/dataset/crime-incident-reports-2012-present.