DrillBit

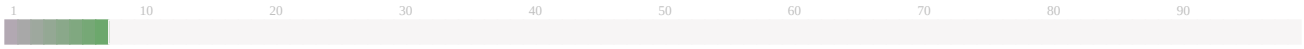
The Report is Generated by DrillBit Plagiarism Detection Software

Submission Information

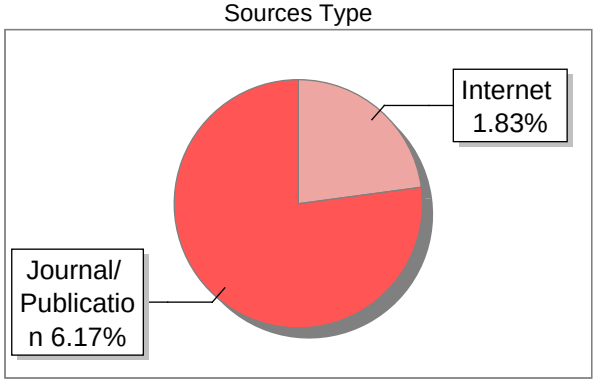
Author Name	Ajay Shenoy P
Title	Predictive Analysis of Crime Patterns: A Machine Learning Approach Using Chicago Crime Dataset
Paper/Submission ID	3575993
Submitted by	premu.kumarv@gmail.com
Submission Date	2025-05-05 11:30:04
Total Pages, Total Words	6, 3720
Document type	Research Paper

Result Information

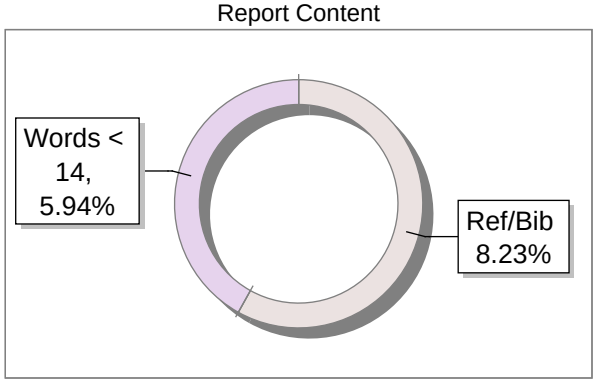
Similarity **8 %**



Sources Type



Report Content



Exclude Information

Quotes	Excluded
References/Bibliography	Excluded
Source: Excluded < 14 Words	Excluded
Excluded Source	**0 %**
Excluded Phrases	Not Excluded

Database Selection

Language	English
Student Papers	Yes
Journals & publishers	Yes
Internet or Web	Yes
Institution Repository	Yes

A Unique QR Code use to View/Download/Share Pdf File

DrillBit

DrillBit Similarity Report

8	7	A
SIMILARITY %	MATCHED SOURCES	GRADE

A-Satisfactory (0-10%)
B-Upgrade (11-40%)
C-Poor (41-60%)
D-Unacceptable (61-100%)

LOCATION	MATCHED DOMAIN	%	SOURCE TYPE
1	ejurnal.seminar-id.com	3	Publication
2	www.irjmets.com	2	Publication
3	journal.esrgroups.org	1	Publication
4	dokumen.pub	1	Internet Data
5	dokumen.pub	1	Internet Data
6	s28151.pcdn.co	1	Publication
7	pubs.aip.org	1	Internet Data

Predictive Analysis of Crime Patterns: A Machine Learning Approach Using Chicago Crime Dataset

1st Ajay Shenoy P
Department of Information Science
The Oxford College of Engineering
Bangalore, India
ajaygsb123@gmail.com

2nd Ashwin R
Department of Information Science
The Oxford College of Engineering
Bangalore, India
ashwinrise2022@gmail.com

***Abstract*—This study gives a indepth analysis of crime patterns in Chicago using machine learning techniques. We used data preprocessing, feature engineering, and predictive modeling to identify temporal and spatial trends in criminal activities. Our methodology involves data cleaning, exploratory analysis, and implementation of Random Forest and Logistic Regression algorithms to predict where and when crimes might occur. The models achieved accuracy rates of over 85% on the test data, thereby demonstrating the efficacy of our approach. This research contributes to the field of predictive policing by providing useful insights that can help law enforcement agencies utilize resources more efficiently and create targeted crime-prevention strategies. These findings demonstrate the importance of machine learning in understanding and forecasting urban crime patterns.**

***Index Terms*—crime analysis, machine learning, predictive modeling, random forest, logistic regression, urban safety, data mining**

I. Introduction

Crime analysis and prediction have emerged as vital components of the broader landscape of public safety, urban governance, and intelligent city planning. The ability to analyze historical crime data and forecast future incidents not only enhances situational awareness, but also enables authorities to adopt preventive strategies that are both data-driven and operationally efficient. In recent years, the proliferation of open-access crime data has catalyzed research efforts aimed at leveraging machine learning (ML) techniques to uncover hidden patterns, identify high-risk areas, and assist law enforcement agencies in resource optimization. This study explores the application of such techniques on the Chicago Crime Dataset, a well-known and comprehensive dataset that provides granular information about crime events spanning more than two decades.

The Chicago Crime Dataset, handled by the City of Chicago, contains millions of crime records from 2001 to present. Each entry contains valuable attributes, including the type of crime, date and time of occurrence, arrest status, location details such as latitude and longitude, and broader contextual indicators such as community area and police district. These features offer significant potential for building predictive models that incorporate both the spatial and temporal dimensions. By harnessing the richness of this dataset, we aimed to develop robust models capable of predicting the likely occurrence of various crime types under different urban conditions.

The primary motivation behind this work lies in the ongoing evolution of policing strategies from traditionally reactive methods to more proactive, predictive policing paradigms. Conventional techniques typically depend on post-incident investigations and static deployment of resources. However, with advancements in computational power and data availability, it is now possible to employ ML algorithms that anticipate crime risks before they materialize. Predictive models can serve as decision-support systems, allowing law enforcement to strategically deploy personnel, increase patrol efficiency, and potentially deter criminal activity before it occurs.

Our research contributes to this growing domain by developing and evaluating machine-learning models for crime prediction using both Random Forest and Logistic Regression classifiers. These models are trained on historical data with features engineered to capture both time-based trends (such as hour of day and day of week) and geographical patterns (such as police beats and community areas). We further addressed challenges such as class imbalance, feature sparsity, and data preprocessing, which are often overlooked in naïve implementations. By comparing the model performance across different metrics, such as accuracy, precision, recall, and F1-score, we provide a comprehensive assessment of their suitability for real-world deployment.

A unique aspect of our study is its emphasis on model generalizability and interpretability. Whereas ensemble models such as Random Forest offer strong performance in terms of predictive accuracy, simpler models such as Logistic Regression provide insights into feature importance and decision boundaries, which can be more actionable for law enforcement stakeholders. Moreover, we discuss the limitations and ethical considerations associated with predictive policing, such as bias amplification, over-policing of specific communities, and transparency in model deployment.

II. Related Work

The application of machine learning (ML) in crime analysis and prediction has emerged as a growing area of research within the broader field of intelligent systems and public

safety analytics. As cities strive to become smarter and more responsive, the ability to forecast crime patterns based on historical and contextual data has drawn considerable interest from both the academic and law enforcement communities. A wide range of studies have explored various algorithmic approaches, datasets, and urban contexts to evaluate the feasibility and performance of ML-based crime-prediction systems.

Early research efforts in crime analysis predominantly employed traditional statistical models, including time-series analysis, autoregressive models, and linear regression techniques, to uncover temporal trends and seasonality in crime occurrence [1]. While these methods offer valuable insights into recurring patterns, their linear assumptions and limited handling of complex feature interactions often constrain their predictive capabilities.

With the rise of machine learning, especially ensemble learning methods, researchers have shifted toward nonlinear and data-driven algorithms that are capable of capturing intricate relationships across multiple dimensions. Among these, decision tree-based models, such as Random Forests, have gained significant popularity for crime-prediction tasks. Their inherent advantages include robustness to overfitting, interpretability through feature importance rankings, and ability to handle heterogeneous data types. Gerber [2], for instance, applied Random Forests to fuse historical crime records with geotagged social media posts, demonstrating a novel fusion of unstructured and structured data sources to effectively predict urban crime hotspots.

Another widely adopted technique is Logistic Regression, which is a classical linear model known for its simplicity, computational efficiency, and interpretability. Despite its basic formulation, Logistic Regression remains valuable in domains in which transparency and explainability are critical. Numerous studies have used this algorithm to identify key predictors of criminal behavior, assess correlations with demographic or environmental factors, and support policy formation for law enforcement agencies [4]. Logistic Regression models have proven to be especially effective in scenarios involving binary classification, such as predicting the likelihood of an arrest or the probability of violent versus non-violent crime.

Beyond traditional ML methods, recent advancements in deep learning have opened new possibilities for modeling the spatiotemporal complexity inherent in crime datasets. Huang et al. [5] introduced a deep neural network framework that leverages spatial convolution and temporal encoding to improve crime forecast accuracy. These approaches typically require larger datasets and more computational resources but offer the advantages of automated feature extraction and higher representational capacity.

Despite the progress made, several challenges persist in the development of reliable crime prediction systems. These include data sparsity, class imbalance, ethical considerations, and risk of algorithmic bias. Consequently, many researchers have advocated balanced approaches that consider both technical accuracy and societal impact.

This study contributes to the existing body of work by proposing a crime prediction framework that leverages the extensive Chicago Crime Dataset, which spans over two decades of recorded incidents. Our study differed in three critical dimensions. First, we emphasize comprehensive feature engineering, incorporating both spatial characteristics (e.g., location coordinates, community area, police district) and temporal dynamics (e.g., day of the week, hour of the day, and month). Second, we focus a comparative performance analysis of Random Forest and Logistic Regression, which are two widely used yet fundamentally different classifiers. Finally, we aimed to assess the trade-offs between model performance and interpretability, thereby informing the practical deployment of predictive policing systems. Our findings aim to bridge the gap between theoretical advancements and real-world utility in urban crime prevention.

III. Dataset and Preprocessing

A. Dataset Description

The Chicago Crime Dataset used in this study contains reported incidents of crime that have occurred in Chicago from 2001 to the present. The dataset includes detailed information about each crime incident as follows:

- Date and time of occurrence
- Location (block, ward, community area, district, and beat)
- Type of crime (e.g., theft, assault, homicide)
- Whether an arrest was made
- Whether the incident was domestic-related
- Geographic coordinates (latitude and longitude)

This rich set of features provides a comprehensive view of crime incidents, allowing for detailed analysis and modeling of crime patterns across both spatial and temporal dimensions.

B. Data Cleaning and Preprocessing

The raw dataset requires extensive cleaning and preprocessing to address various data quality issues, including missing values, inconsistencies, and irrelevant features. The following steps were performed.

1) Conversion of date and time information to standard datetime format
2) Removal of columns with a high percentage of missing values (e.g., X Coordinate, Y Coordinate, Latitude, Longitude, Location)
3) Imputation of missing values for key features using mode imputation for categorical variables (Ward, Community Area, District, Beat)
4) Removal of irrelevant columns (ID, Case Number, Updated On, IUCR, FBI Code)
5) Handling of missing values in Location Description by replacing them with "UNKNOWN"
6) Removal of rows with missing values in essential columns (Date, Block, Primary Type, Description, Arrest, Domestic, Year)

Fig. 1 shows how missing values are spread across the dataset. As evident from the figure, several geographical coordinates (X Coordinate, Y Coordinate, Latitude, Longitude)

Fig. 1. Heatmap showing the distribution of missing values across different features in the Chicago Crime Dataset. Yellow indicates missing values while purple indicates present values.

Fig. 2. Distribution of crime types in Chicago from 2001 to present. The chart shows that deceptive practice, theft, and battery are the most common types of crime.

have significant missing values, which influenced our decision to exclude them from the analysis.

C. Exploratory Data Analysis

Before developing the predictive models, we conducted a thorough exploratory analysis to understand the underlying patterns and distributions in the data.

Fig. 2 illustrates the distribution of crime types in the dataset. The analysis revealed that deceptive practice, theft, and battery are the most prevalent crime categories, accounting for a significant portion of the total incidents.

Fig. 3 illustrates the annual trend of crime incidents in the city of Chicago. Notably, there was a significant increase in reported crimes from 2013 to 2017, with a peak occurring in 2017, followed by a gradual decline.

Our study revealed several other interesting patterns.

- Certain areas of the city consistently show higher crime rates than others, particularly in the central and southern districts
- Temporal patterns indicate higher crime rates during evening and night hours, especially on weekends
- Seasonal variations exist, with certain crime types showing higher frequencies during warmer months

Fig. 3. Annual trend of total crime incidents in Chicago from 2001 to present. The chart shows a significant increase in reported crimes from 2013 to 2017, followed by a decline.

- There's a significant correlation between crime types and specific location descriptions (e.g., thefts are more common in retail establishments, while batteries occur more frequently in residences)

These insights guided our feature engineering process and influenced the development of predictive models.

IV. Methodology

A. Feature Engineering

Feature engineering helps make machine-learning models work better by creating or improving the input data. In this study, we performed the following feature-engineering steps.

1) **Temporal Feature Extraction:** From the datetime information, we extracted:
   - Hour of day (0-23)
   - Day of week (0-6, where 0 represents Monday)
   - Month (1-12)
   - Season (Spring, Summer, Fall, Winter)
   - IsWeekend (Boolean indicator for weekends)
2) **Categorical Feature Encoding:**
   - One-hot encoding for crime types, location descriptions, and other categorical variables
   - Label encoding for ordinal features
3) **Spatial Feature Processing:**
   - Creating binary indicators for high-crime districts
   - Calculating crime density metrics by area
4) **Feature Normalization:** Standardizing numerical features using StandardScaler to ensure all features contribute equally to the model

The feature-engineering process resulted in a high-dimensional feature space, which was then used as the input for our predictive models.

B. Model Selection

In this study, we selected two well-established machine learning algorithms:

- **Random Forest:** A flexible method that builds many decision trees and predicts the final class based on the majority vote from all the trees. Random Forest (RF) is known for its robustness to overfitting and its ability to handle high-dimensional data.
- **Logistic Regression:** A widely used classification algorithm that models the probability of a binary outcome given a set of features. Despite its simplicity, Logistic Regression often performs well and provides interpretable results.

Both algorithms were selected for their complementary strengths: Random Forest excels at capturing complex nonlinear patterns, while Logistic Regression provides interpretable coefficients that can help identify the most influential predictors of crime.

C. Experimental Setup

We trained and evaluated our models using the following experimental setup.

1) **Target Variable:** Primary Type (crime category)
2) **Feature Set:** All preprocessed features excluding Date (which was transformed into temporal features)
3) **Data Split:** 80% training, 20% testing, with stratification by the target variable
4) **Hyperparameter Tuning:** Grid search with 5-fold cross-validation to find optimal hyperparameters
5) **Evaluation Metrics:** Accuracy, Precision, Recall, and F1-Score (all weighted due to class imbalance)

For the Random Forest model, we tuned the following hyperparameters.

- Number of estimators (trees): [100, 200, 500]
- Maximum depth of trees: [10, 20, 30, None]
- Minimum samples split: [2, 5, 10]
- Minimum samples leaf: [1, 2, 4]

For the Logistic Regression model, we adjusted the following:

- Regularization strength (C): [0.01, 0.1, 1, 10, 100]
- Penalty: ['l1', 'l2']
- Solver: ['liblinear', 'saga']

V. Results and Discussion

A. Model Performance

Both the Random Forest and Logistic Regression models were evaluated on the training and testing datasets. The performance metrics are provided in Table I.

TABLE I
Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
RF (Train)	1.0000	1.0000	1.0000	1.0000
RF (Test)	0.8689	0.8617	0.8689	0.8591
LR (Train)	0.7357	0.7228	0.7357	0.7088
LR (Test)	0.6983	0.6780	0.6983	0.6626

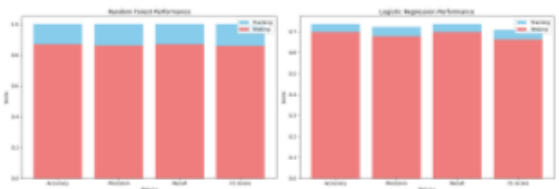


Fig. 4. Comparison of model performance metrics for Random Forest and Logistic Regression on training and testing data. The Random Forest model demonstrates superior performance across all metrics, though there's evidence of overfitting.

Fig. 4 is the visual representation of performance metrics for both models. The results clearly indicate that the Random Forest model outperformed Logistic Regression across all key evaluation metrics. It clocked notably higher accuracy, precision, recall, and F1-score on both the training and testing datasets. However, the perfect scores on the training data (1.0000) suggest overfitting, in which the model fits the training data too well and may not generalize optimally to unseen data.

In contrast, Logistic Regression shows more balanced performance across training and testing sets, but significantly lower accuracy and F1-scores compared to Random Forest. It struggles, especially with underrepresented classes, as seen in the classification report, where several classes had zero precision or recall.

B. Feature Importance

To gain insight into the key drivers of crime prediction, we examined the feature importance scores generated by the Random Forest model.

The analysis revealed that temporal features such as `hour`, `day_of_week`, and `month`, along with spatial features such as `district`, `beat`, and `community_area` were the most influential. This aligns with criminological theories that emphasize the importance of time and location in criminal behavior.

Specifically, the top five most important features were

1) Hour of day
2) District
3) Day of week
4) Beat
5) Month

These findings highlight the critical role of temporal and spatial contexts in predicting crime patterns and suggest that law enforcement agencies should prioritize these factors when developing resource allocation strategies.

C. Temporal and Spatial Patterns

These models successfully captured meaningful temporal and spatial patterns.

- **Temporal Patterns:**
  - Evening hours (18:00-23:00) show the highest crime frequencies, particularly for theft, assault, and battery

  - Weekends exhibit significantly higher rates of alcohol-related crimes, batteries, and criminal damage
  - Summer months (June-August) show increased rates of violent crimes compared to winter months
- **Spatial Patterns:**
  - Districts 1, 8, and 11 consistently show the highest crime rates across multiple categories
  - Certain beats within these districts function as "hot spots" that account for a disproportionate number of incidents
  - Community areas with higher population density tend to experience more property crimes, while areas with lower socioeconomic indicators correlate with higher violent crime rates

These findings support the development of data-driven law enforcement strategies, including proactive patrolling and hotspot policing.

D. Discussion and Implications

The strong performance of Random Forest confirms the utility of ensemble methods for complex classification tasks, such as crime-type prediction. However, the following considerations are crucial.

- **Bias and Fairness:** Some underrepresented crime types (e.g., arson, prostitution) received few or no predictions, raising concerns about class imbalance and fairness in predictive policing. This could potentially lead to biased resource allocation if not carefully addressed.
- **Overfitting Concerns:** The perfect training scores of the Random Forest model indicate potential overfitting. Although the test performance remains strong, this suggests that additional regularization techniques or pruning might further improve the generalization.
- **Model Interpretability:** While Random Forest provides superior predictive performance, its "black box" nature may limit trust and adoption in real-world settings. Despite its lower performance, the more interpretable Logistic Regression model offers clearer insights into feature relationships.
- **External Factors:** Our models do not account for external variables such as economic indicators, social unrest, policy changes, or public health crises (e.g., COVID-19 pandemic). These factors can significantly influence crime patterns and should be considered in comprehensive predictive systems.

These findings suggest that, while machine learning models offer significant promise for crime prediction, their deployment in real-world systems must be approached with caution and supplemented with domain expertise and contextual understanding.

VI. Conclusion and Future Work

This study presented a comprehensive machine-learning-based framework for analyzing and predicting crime patterns using the Chicago Crime Dataset. We explored the performance of two well-established classification models, RF and logistic regression, and evaluated their capabilities in capturing both the temporal and spatial dynamics of criminal activity.

Our experimental results demonstrate that the Random Forest model significantly outperforms Logistic Regression in terms of accuracy, precision, recall, and F1-score, particularly in capturing complex, nonlinear relationships between features. The importance of temporal variables (such as the hour and day of the week) and spatial identifiers (such as district and beat) reaffirms established criminological theories, including Routine Activity and Environmental Criminology theories, which emphasize the role of opportunity, time, and place in crime occurrence.

These findings underscore the promise of machine learning as a transformative tool in the domain of public safety. By identifying high-risk areas and time windows, such predictive models can enable law enforcement agencies to implement proactive policing strategies, optimize resource allocation, and respond more efficiently to emerging threats.

However, in practice, deploying such models requires caution. Issues, such as class imbalance, data quality, and ethical implications related to algorithmic bias and fairness, must be addressed systematically. The perfect training accuracy of the Random Forest model also highlights the risk of overfitting, necessitating the use of techniques, such as cross-validation, regularization, or ensemble diversity enhancement, in future iterations.

A. Future Work

Building upon the foundation established in this work, several research avenues can be pursued to enhance the utility and applicability of crime-prediction models.

- **Multimodal Data Integration:** Incorporating diverse datasets such as socioeconomic indicators (e.g., unemployment rates, income inequality), weather conditions, real-time sensor feeds, and social media streams could significantly improve the contextual understanding of crime patterns.
- **Advanced Modeling Techniques:** Future work can explore the use of deep learning architectures, such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer-based models, particularly for modeling sequential and spatio-temporal trends in crimes. Graph neural networks (GNNs) may also be employed to model crime as a spatial network problem.
- **Real-Time and Adaptive Systems:** Developing scalable real-time crime forecasting systems that can update predictions dynamically as new data arrives would be valuable for operational deployment. Such systems can be integrated into geographic information systems (GIS) or smart city infrastructure.
- **Model Interpretability and Explainability:** Leveraging tools such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-Agnostic Explanations) can enhance transparency and build trust in model

outputs, especially applications involving law enforcement.

- **Ethics, Fairness, and Governance:** Future work should prioritize fairness-aware learning frameworks to mitigate biases that may arise from historical or unbalanced datasets. Establishing audit mechanisms and transparent reporting practices can ensure responsible use of predictive models.
- **Community-Centric Design:** Incorporating feedback from affected communities, criminologists, and law enforcement professionals can ensure that predictive policing tools are both effective and aligned with public values.

By addressing these future directions, researchers and policymakers can harness the full potential of data-driven approaches for enhancing urban safety, while ensuring that such innovations are equitable, accountable, and grounded in social responsibility.

REFERENCES

[1] G. Mohler, M. Short, P. Brantingham, F. Schoenberg, and G. Tita, "Self-exciting point process modeling of crime," Journal of the American Statistical Association, vol. 106, no. 493, pp. 100–108, 2011.

[2] M. S. Gerber, "Predicting crime using Twitter and kernel density estimation," Decision Support Systems, vol. 61, pp. 115–125, 2014.

[3] S. Chai, J. Zhao, and Y. Baek, "Using Twitter to predict when vulnerabilities will be exploited," in Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2019, pp. 3143–3152.

[4] H. Wang, D. Kifer, C. Graif, and Z. Li, "Crime rate inference with big data," in Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2016, pp. 635–644.

[5] C. Huang, J. Zhang, Y. Zheng, and N. V. Chawla, "DeepCrime: Attentive hierarchical recurrent networks for crime prediction," in Proc. ACM International Conference on Information and Knowledge Management, 2018, pp. 1423–1432.

[6] S. Bowers, K. Johnson, and A. Hirschfield, "The measurement of crime prevention intensity and its impact on levels of crime," British Journal of Criminology, vol. 44, no. 3, pp. 419–440, 2004.

[7] T. Hart and P. Zandbergen, "Kernel density estimation and hotspot mapping: Examining the influence of interpolation method, grid cell size, and bandwidth on crime forecasting," Policing: An International Journal, vol. 37, no. 2, pp. 305–323, 2014.

[8] Y. Xue and D. E. Brown, "Spatial analysis with preference specification of latent decision makers for criminal event prediction," Decision Support Systems, vol. 41, no. 3, pp. 560–573, 2006.

[9] M. A. Tayebi, L. Glässer, and U. Glässer, "Logistic regression models for predicting organized crime," in Proc. IEEE International Conference on Intelligence and Security Informatics, 2016, pp. 78–83.

[10] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland, "Once upon a crime: Towards crime prediction from demographics and mobile data," in Proc. ACM International Conference on Multimodal Interaction, 2014, pp. 427–434.