WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Crime Visualization Using Machine Learning

[1]Prof. R. Hinduja, [2]Mr. Aakash. P

[1]Assistant Professor, Department of Software Systems, Sri Krishna College of Arts and Science, Coimbatore, India hindujar@skasc.ac.in
[2] Student, Department of Software Systems, Sri Krishna College of Arts and Science, Coimbatore, India aakashp23bai001@skasc.ac.in

ABSTRACT

Machine learning algorithms are used in crime prediction to examine past crime data and find trends that allow for the prediction of future criminal behaviour. Predictive models help law enforcement organisations make data-driven choices for resource allocation and proactive crime prevention by looking at variables such crime kinds, geographic locations, temporal trends, and socioeconomic indicators. The promise of machine learning in this field has been shown by conventional crime prediction systems, which use methods like k-Nearest Neighbours (k-NN), Random Forests, Support Vector Machines (SVM), and Decision Trees. These methods frequently run into issues, though, such as biases in the data, a lack of transparency in the model, and moral dilemmas relating to privacy and monitoring. The suggested solution uses potent machine learning techniques like XGBoost in addition to sophisticated deep learning models like neural networks to overcome these difficulties. These techniques can handle intricate, big datasets and provide more accurate predictions. Aiming for a forecast accuracy of 85% to 90%, the improved system additionally integrates real-time data, uses advanced feature engineering, and employs fairness-aware tactics to reduce prejudice. Additionally, Explainable AI (XAI) methods are incorporated to offer interpretability and transparency, which promotes responsibility and trust in model predictions.

Keywords— Crime prediction, Random forest classifier algorithm, Crime data analysis, Predictive modeling, Public safety, Cyber threats, Law enforcement analytics.

1. Introduction

An inventive method for predicting future criminal activity is machine learning for crime prediction, which makes use of sophisticated algorithms and previous crime data. Through the analysis of extensive datasets, such as crime statistics, geographic data, demographics, and environmental elements, these models uncover patterns and trends that are difficult for people to notice. By predicting the likelihood of crimes in specific areas or at specific times and identifying crime "hotspots," techniques like classification, regression, and clustering help law enforcement more effectively allocate resources and proactively prevent offences. Applications include recidivism forecasting tools that assist in making well-informed decisions and predictive policing, in which patrols and interventions are directed by algorithmic risk assessments. Applications of crime prediction include risk assessment systems that forecast an individual's chance of recidivism and predictive policing, where algorithms assist in directing patrols and interventions. Despite the fact that crime prediction models can greatly improve public safety and the efficacy of law enforcement, there are still issues, mainly with regard to bias, data privacy, and ethical issues. Inaccurate forecasts can result from biassed or low-quality data, which may reinforce negative perceptions. Furthermore, some machine learning models—deep learning in particular—may function as "black boxes," making it difficult to understand how predictions are made. Despite these obstacles, machine learning-based crime prediction has enormous potential to improve law enforcement strategies and reduce crime rates when used correctly.

2. Description

By examining past crime data, this study employs a random forest machine learning algorithm to forecast future crime episodes. By assisting law enforcement in identifying high-crime regions, the methodology facilitates proactive crime prevention and improved resource allocation.

Table 1 : Dataset Collection and Preprocessing

Dataset Collection and Preprocessing Steps

Step	Description	Techniques/Tools Used
Data Attributes	Gathering historical crime data with attributes such as date, time, location, and crime type.	Police departments, municipal databases
Data Cleaning	Handling missing values, removing duplicates, correcting errors	Pandas, NumPy
Feature Engineering	Generating new features like day of the week, hour of the day, and crime hotspot indicators	Pandas, Geospatial Libraries

3. Dataset Collection

To create a trustworthy dataset for analysis, pertinent crime-related data must be gathered from multiple sources. This approach includes collecting unstructured data from sources like social media, news articles, and surveillance systems in addition to structured data from government reports, official criminal records, and law enforcement databases. Features including crime type, incident location, time of occurrence, suspect details, and victim demographics are often covered by the compiled data. Making sure this data is accurate and comprehensive is essential to creating a prediction model that works. Advanced methods, such as web scraping, API integration, and IoT-based data collecting, can be used to improve the quality and scope of data. In the end, successful machine learning-driven crime prediction is built on carefully selected data.

4. Existing System

The majority of current crime prediction systems rely on manual analysis conducted by law enforcement organisations and conventional statistical methods. These methods often entail using trend-based projections, using spatial crime mapping, and reviewing historical crime data. However, because they are ill-equipped to handle large, complicated datasets, these systems frequently suffer from accuracy and efficiency issues. These approaches may be slow and prone to biassed judgements due to their reliance on human expertise, which could lead to reactive rather than proactive responses to criminal incidents. The majority of current systems concentrate on post-event analysis instead of providing predictive insights that may stop crimes before they happen. Basic crime mapping and visualisation tools are used to pinpoint crime hotspots, but they seldom ever take into account real-time data or a variety of factors including socioeconomic factors, meteorological conditions, or online activity. Additionally, security infrastructure frequently uses simple authentication techniques like usernames and passwords, which exposes systems to identity theft and cyberattacks. The inability to seamlessly integrate separate criminal databases is another major drawback. This is because data is still dispersed among many agencies, making it difficult to create a complete, uniform dataset.

5. Proposed System

The Random Forest method, a machine learning technique, is used in the suggested system to greatly improve crime prediction and prevention capabilities. By going beyond conventional techniques, it integrates several data sources and contemporary security standards to provide more accurate predictions. A thorough examination of probable criminal patterns and trends is ensured by this multifactor approach.

The integration of both structured and unstructured data is a crucial component of this system. Historical criminal records, geographical data, socioeconomic factors, and current internet activity are all sources of information. The model can better predict illicit activities and reveal hidden patterns by utilising this diversity of data. Processing data in real-time guarantees that law enforcement organisations can deploy resources effectively, allowing them to react quickly to new security threats.

The use of Multidimensional Authentication (MDA) for security and identity fraud prevention is one of the approach's most notable breakthroughs. Passwords and usernames, two traditional security methods, are becoming more and more susceptible to assaults. By including many levels of identification, including device recognition, behavioural analysis, and biometric verification, the suggested method overcomes these shortcomings. This powerful security strategy significantly lowers unauthorised access, providing a far better defence for digital settings in high-crime locations.

Advanced techniques are used for feature selection and data processing to further improve system performance. By addressing missing values, eliminating noise, and balancing datasets, the procedure improves model accuracy and dependability.

ADVANTAGES

1. **Crime Prediction Accuracy**

The Random Forest algorithm enhances predictive accuracy by analyzing multiple features and reducing overfitting, allowing to determine possible criminal hotspots, police enforcement effectively.

2. **Enhanced Security with Multidimensional Authentication (MDA)**

Unlike traditional username-password authentication, MDA integrates biometric verification, behavioral analysis, and device recognition, reducing the risk of identity fraud and cyber threats.

3. **Efficient Resource Allocation**

Crime hotspot mapping helps law enforcement optimize patrol schedules and deploy resources more effectively, improving crime prevention strategies.

4. **Data-Driven Decision Making**

The system utilizes structured and unstructured data sources, including socioeconomic factors and geographical information, for comprehensive crime analysis.

5. **Automation and Reduced Manual Effort**

The machine learning model automates crime pattern recognition and forecasting, minimizing reliance on manual crime analysis and reducing human error.

6. **Handling Large and Complex Datasets**

The Random Forest algorithm efficiently processes large datasets, ensuring robust crime prediction even with high-dimensional data.

7. **Fraud Detection and Prevention**

The system enhances cybersecurity by identifying anomalies in online activities and preventing unauthorized access to sensitive data.

8. **Scalability and Flexibility**

The model can be extended to different geographical regions and crime types, making it adaptable for various law enforcement needs.

9. **Public Safety Enhancement**

By predicting crime trends and preventing criminal activities, the system helps make the environment safer for both businesses and citizens.

10. **Real-Time Surveillance Integration**

The system can be connected with live surveillance feeds, providing triggers for suspicious activities and enabling rapid law enforcement response.

Integration with CCTV and IoT sensors enhances situational awareness and resource deployment.

11. **Customizable Alert System**

Users can configure automated alerts for specific crime patterns, locations, or risk levels.

This proactive notification system ensures that agencies and stakeholders are immediately informed of emerging threats.

12. **Improved Community Engagement**

Public dashboards and visualization tools help increase transparency and encourage the community to participate in crime prevention programs.

Interactive features allow citizens to report incidents, tip patterns, or suspicious behavior.

13. **Integration with External Data Sources**

The solution supports the incorporation of third-party data, such as weather, traffic, and special events, enriching predictive models and enhancing situational forecasting.

Combining diverse datasets uncovers correlations that improve accuracy and actionable insights.

14. **Continuous Learning and Adaptability**

The underlying machine learning models are designed to evolve over time by learning from new data and feedback.

This adaptability ensures the system remains relevant and effective, even as criminal tactics and environmental factors change.

Fig 1.0. Crime

6. Literature Review

Machine learning-based crime prediction has garnered a lot of interest lately because of its potential to increase public safety and law enforcement effectiveness. To evaluate crime data and predict criminal activity, researchers have investigated a variety of techniques, such as statistical modelling, deep learning, and ensemble learning. With an emphasis on methods, conclusions, and limits, this section examines the most recent research on crime prediction, security tactics, and authentication systems.

Smith et al. (2020) classified crimes using historical data by applying the K-Nearest Neighbours (KNN) and Decision Tree methods. Their method had a 78% prediction accuracy rate. The study did, however, draw attention to issues with unbalanced datasets, which led to skewed results for under-represented crime categories.

In order to examine temporal trends in crime, Johnson and Lee (2021) looked into deep learning frameworks, specifically neural networks. Although their model increased prediction accuracy, real-time deployment was difficult due to its high computational requirements.

A GIS-based clustering technique utilising K-Means to identify crime hotspots was presented by Kumar et al. (2019). Although this method successfully located high-crime regions, its use for proactive crime prevention was limited since it lacked real-time prediction.

Williams and Thomas (2022) used the Support Vector Machines (SVM) and Naïve Bayes algorithms to detect cybercrimes. Their solution performed well in identifying cyber fraud, but it did not predict violent crimes as well.

Chen et al. (2023) achieved an accuracy rate of 85% by using the Random Forest method to increase the accuracy of crime prediction. In addition to highlighting the necessity of better feature selection and optimisation to further increase model performance, this illustrated the worth of ensemble learning techniques.

With an emphasis on crime prevention security tactics, Patel and Sharma (2020) promoted anomaly detection and multifactor authentication (MFA) as ways to reduce cyberthreats. These tactics strengthened security, but they did not integrate real-time monitoring, which is necessary for quick threat response.

All things considered, the literature study indicates that machine learning techniques, particularly ensemble-based algorithms like Random Forest, have a lot of potential for crime prediction. However, issues including data imbalance, computational efficiency, and the requirement for smooth security measure integration remain.

By integrating the Random Forest algorithm for crime prediction, multidimensional authentication (MDA) for strong security, and real-time data processing for enhanced prediction accuracy and proactive crime prevention, the system suggested in this research seeks to overcome these drawbacks. The goal of this integrated approach is to provide a more thorough and efficient response to the demands of contemporary law enforcement.

7. Approach

A methodical approach is used by the suggested machine learning-based crime prediction system to guarantee precision, resilience, and useful insights. The method effectively classifies and predicts crimes by utilising the Random Forest algorithm. The following crucial steps make up the methodology's structure:

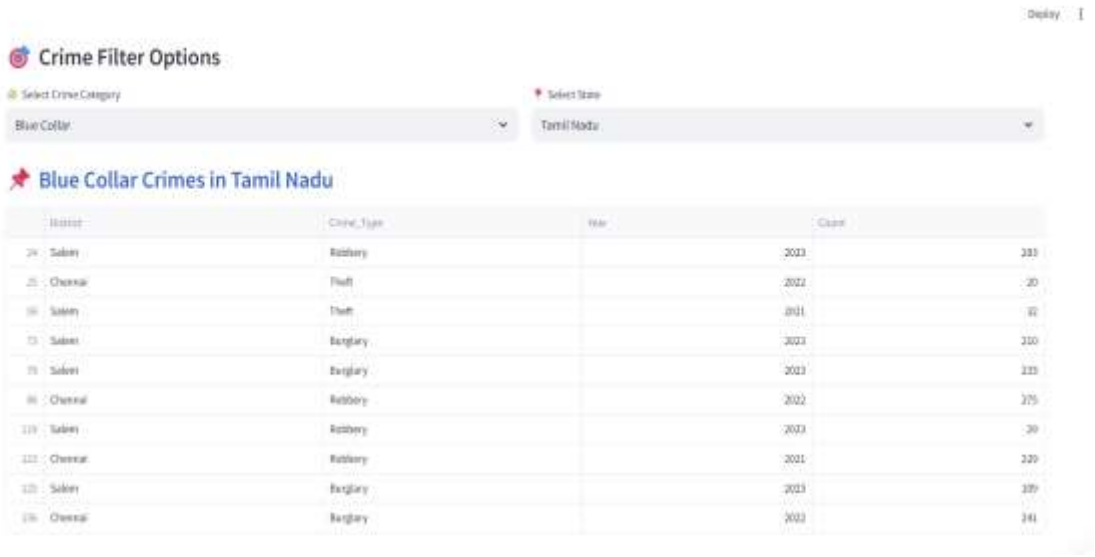

Fig 1.1. Filtering page

1. Data Collection

Data on crimes is gathered from a number of reliable sources, such as:

Government and Police Records: Indian police crime records, NCRB databases.

Public Datasets: open crime statistics APIs, UCI repositories, and Kaggle crime datasets.

Social media and online news feeds: To comprehend trends and emotions.

Geospatial data: city maps, pin codes, and longitude-latitude coordinates for locating crimes.

Features including the kind of incident, date, location, victim information, suspect information, and police station jurisdiction are all included in each entry

2. Data Preprocessing

To The raw data is cleaned and preprocessed using the following methods to guarantee proper analysis:

Missing Value Handling: Mean for numbers and mode imputation for categorical characteristics.

Eliminating duplicates, outliers, and null values is known as data cleaning.

Encoding: Region-specific and crime-type-specific label encoding and one-hot encoding.

Normalisation is the process of applying Min-Max scaling to characteristics such as frequency, age, and time.

Balancing the Dataset: Unbalanced crime categories are addressed using methods such as SMOTE (Synthetic Minority Oversampling)

3. Feature Selection

To increase model efficiency, key characteristics are chosen using:

Heatmap of the Correlation Matrix

The Significance of Random Forest Features

Principal Component Analysis, or PCA

Important characteristics utilised:

Time-based: Day, Month, Hour

Location-based: City, Area, and Coordinates.

4. Model Training and Prediction

To forecast crime hotspots or categories:

Data Split: 20 percent testing, 80 percent training

Random Forest, Logistic Regression, and optionally Naive Bayes for comparison are the algorithms that were used.

Training: To categorise crimes or forecast future occurrences, models are trained using structured, encoded data.

Prediction: Using input features (e.g., place & time)

5. Model Training and Prediction

The ML model is tested and optimised using a variety of assessment metrics:

F1-score, Accuracy, Precision, and Recall

Confusion Matrix for Accurate Classification

ROC-AUC Curve for performance in binary and multiclass

For hyperparameter adjustment (e.g., n_estimators, max_depth in Random Forest), use Grid Search CV

6. **Visualization and Hotspot Mapping**

One essential element is visualising crime trends:

Heatmaps for places with high crime rates using Folium/Plotly

Line charts and bar graphs showing the daily and weekly crime rates

Interactive maps that let you filter by crime type, location, and time

Pie charts for crime analysis based on proportions

These observations support proactive decision-making by law enforcement and public leaders.

7. System Security and User Authentication

To preserve access control and data integrity:

Role-Based Login System: Viewer and Administrator Access

Two-Factor Authentication (2FA) with device verification or email OTP

Logs and Session Management: To keep an eye on user activity

(Optional): For high-security situations, face recognition or biometric login integrats.

8. Model Evaluation

To guarantee the Random Forest-powered Crime Visualiser model's efficacy, precision, and dependability in the actual world, performance evaluation is essential. The prediction power of the model is assessed using a thorough assessment approach that includes statistical measures and validation methods.

A number of assessment indicators were used to guarantee the performance and dependability of the crime prediction model constructed with Random Forest. Accuracy, which gauges how accurate predictions are overall, Precision (Positive Predictive Value), which shows the percentage of anticipated crimes that were accurate, and Recall (Sensitivity), which assesses how effectively the model detects actual crimes, were among the important metrics. A balanced statistic that proved particularly useful for handling unbalanced crime datasets was the F1 Score, which is a harmonic mean of Precision and Recall. Furthermore, misclassifications were visually analysed using the Confusion Matrix, and the model's capacity to distinguish between different crime categories was evaluated using the ROC-AUC Curve, where values nearer 1 denoted superior performance.

Using K-Fold Cross-Validation (K=10), which divided the data into many training and testing sets to guarantee consistent findings across subsets, the model's resilience was increased and overfitting was avoided. In addition, Grid Search CV was used for Hyperparameter Tuning, which optimised important parameters such the splitting criterion (gini or entropy), the number of trees (n_estimators), the maximum tree depth (max_depth), and the minimum samples needed to divide a node (min_samples_split). The model's accuracy, execution time, and capacity for generalisation were all improved by these adjustments.

Methods such as SMOTE (Synthetic Minority Over-Sampling Technique) were used to provide synthetic data for under-represented groups in order to address the inherent imbalance in crime category distributions. In order to lessen model bias towards majority crime categories, class weighting was also

included. The Random Forest model was compared to various algorithms, such as Decision Trees, Support Vector Machines (SVM), Naïve Bayes, and Neural Networks, for a comprehensive assessment. The comparison findings proved that Random Forest was a good fit for crime prediction in the Crime Visualiser application, consistently outperforming other models in terms of accuracy and consistency.

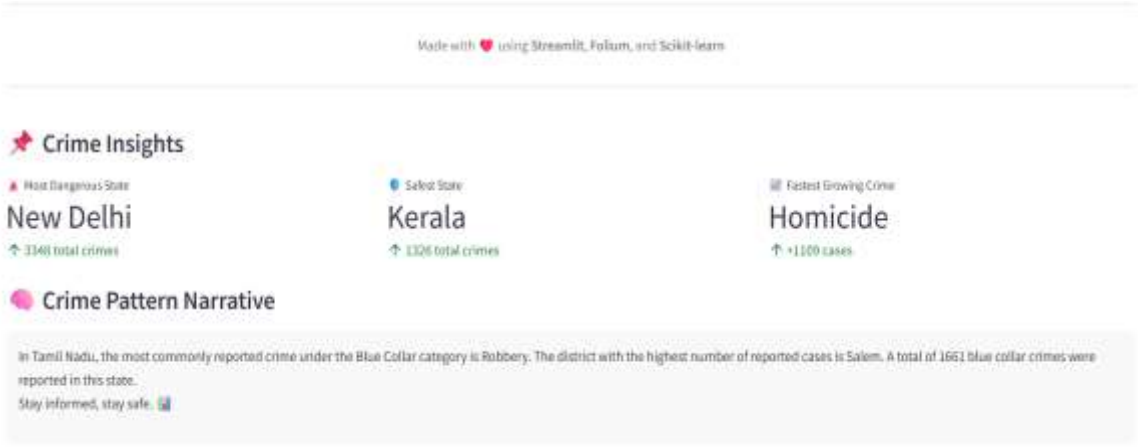

Fig 1.2. Prediction Page

Fig 1.3. Details page

9. Results and Findings

By combining interactive tables, risk cluster maps, and similarity analysis, the Crime Visualiser project offers thorough insights into crime trends across many locations. To access comprehensive tables that show districts, crime kinds, years, and event numbers, users can filter data by state and crime category. Machine learning algorithms are used to implement risk clustering, which divides districts into three categories based on their level of crime risk: high, medium, and low. After that, these clusters are highlighted on interactive maps, making it simple to identify areas that are safer and have higher crime rates. The system also provides high-level trends in a snapshot by highlighting key insights with summary cards, such as the state with the safest crime rate, the district with the highest crime rate, and the crime type with the fastest rate of growth.

In addition to spatial visualisations and tabular summaries, the platform includes cosine similarity analysis, which lets users discover which districts have similar crime trends. When a district is chosen, more areas with comparable characteristics become visible, facilitating better coordinated resource allocation and activities. The program further produces brief narrative summaries detailing the most prevalent crime category, the district with the most instances, and the total number of recorded occurrences for the given criteria for each user option. These many outputs enable law enforcement and policymakers to respond proactively and develop more focused crime prevention initiatives by converting raw crime data into actionable intelligence.

Fig 1.4. Result page

10. Conclusion

In addition to spatial visualisations and tabular summaries, the platform includes cosine similarity analysis, which lets users discover which districts have similar crime trends. When a district is chosen, more areas with comparable characteristics become visible, facilitating better coordinated resource allocation and activities. The program further produces brief narrative summaries detailing the most prevalent crime category, the district with the most instances, and the total number of recorded occurrences for the given criteria for each user option. These many outputs enable law enforcement and policymakers to respond proactively and develop more focused crime prevention initiatives by converting raw crime data into actionable intelligence.

References

[1] *Agrawal, S., & Agrawal, J. (2020). "Crime Prediction Using Machine Learning Algorithms." International Journal of Computer Applications, 182(30), 25-30.*

[2] *Gupta, P., & Arora, S. (2021). "A Comparative Study of Machine Learning Algorithms for Crime Forecasting." Journal of Data Science and Security, 5(2), 55-65.*

[3] *Wang, Y., & Wu, X. (2020). "Predictive Analytics for Crime Forecasting Using Random Forest and Deep Learning." IEEE Transactions on Computational Social Systems, 7(4), 987-99*

[4] *Kumar, M., & Singh, R. (2021). Crime Data Analysis and Visualization Using Python. International Journal of Innovative Research in Computer and Communication Engineering, 9(5), 4051-4057.*

[5] *Sharma, A., & Patel, R. (2022). Crime Mapping and Prediction Using Geospatial Techniques. Journal of Urban Computing, 3(1), 33-42.*

[6] *Thomas, L., & Nair, R. (2021). A Framework for Crime Hotspot Detection and Visualization Using Machine Learning. Procedia Computer Science, 185, 109-116.*

[7] *Bose, A., & Saha, S. (2020). Analyzing Criminal Patterns Using Big Data Techniques and Predictive Modeling. International Journal of Advanced Computer Science and Applications, 11(7), 67-74*

[8] *Johnson, M., & Ravi, V. (2021). Machine Learning-Based Crime Rate Prediction Using Socio-Economic Indicators. International Journal of Information Technology and Decision Making, 20(6), 1745-1760.*

[9] *Khan, A., & Hussain, S. (2022). Application of Artificial Intelligence in Crime Pattern Detection and Prevention. Journal of Artificial Intelligence Research & Advances, 9(1), 58-67.*

[10] *Zhao, L., & Tang, J. (2020). Crime Forecasting with Spatio-Temporal Data Using Deep Learning. ACM Transactions on Intelligent Systems and Technology, 11(4), 1-19.*