Crime Prediction System Using Machine Learning

Rina Kumari*, Shrimali Parmit Dinesh Kumar**, Seepana Tarun***,
Vohra Sahil Mustufa****, Singiri Nivas*****

*Dept. of Computer Science & Engineering, Parul University, Vadodara,
Email: rina.kumari34123@paruluniversity.ac.in
**Dept. of Computer Science & Engineering, Parul University, Vadodara,
Email: 2303031059002@paruluniversity.ac.in
***Dept. of Computer Science & Engineering, Parul University, Vadodara,
Email: 2203031050597@paruluniversity.ac.in
****Dept. of Computer Science & Engineering, Parul University, Vadodara,
Email: 2303031057169@paruluniversity.ac.in
*****Dept. of Computer Science & Engineering, Parul University, Vadodara,
Email: 2203031050648@paruluniversity.ac.in

Abstract: Crime is a major issue in today's society, especially in urban areas. Rapid growth, economic inequality, and increasing populations contribute to various criminal activities. Traditional law enforcement usually responds to incidents after they happen. While this response is necessary, it does not fully address the changing and flexible nature of modern crime, which often shows clear patterns over time and location. There is a need for proactive, data-driven systems that can predict potential crimes and help prevent them.

This paper introduces a lightweight and scalable **Crime Prediction System**. It combines supervised machine learning models with unsupervised clustering techniques to identify areas and times likely to experience crime. The system uses Random Forest, Decision Tree, and Logistic Regression algorithms to classify different types of crime, while K-Means clustering helps locate emerging hotspots. Historical crime data is preprocessed to extract spatio-temporal features for training and evaluation.

To make the system user-friendly, we include a web-based dashboard that provides real-time interactive visuals such as heatmaps, time-series graphs, and trend analyses. These visual tools allow law enforcement and policymakers to quickly interpret results, allocate resources effectively, and develop targeted interventions.

Our testing revealed that the Random Forest model performed best overall. It achieved close to 85% accuracy and consistently outperformed other classifiers, especially when dealing with complex, imbalanced datasets. However, accuracy wasn't our only concern. We also included strong security measures. The system includes all the monitoring features to give accurate results, and all the data is secure and save in the database.This is the way which combines machine learning with visual analysis which makes anyone understand easily about the crime details which helps to make more security measures in the areas where the crime rate is shown high .

***Keywords* — Crime Prediction, Machine Learning, Random Forest, Decision Tree, Logistic Regression, K-Means Clustering, Hotspot Analysis, Predictive Policing, Smart Cities**

I. INTRODUCTION

The techniques of the crime prediction system is changing which makes the process fast and easier . Early methods relied on statistical models like regression analysis and kernel density estimation (KDE). While these methods could identify high-crime areas, they struggled with accurate predictions. The introduction of machine learning brought significant improvements. Algorithms such as Decision Trees, Random Forests, SVMs, and k-NN demonstrated better accuracy in classifying crimes. Random Forests became especially popular for their ability to handle noisy and unbalanced data.

Recently, deep learning has pushed the field further.Models like CNN's and RNNs are commonly used to track the crime trends in the past years over different locations. Researchers are looking forward to Graph Neural Networks(GNNs) to look into the crime data within the available networks. Additionally, combining geospatial and temporal features with GIS systems has enhanced hotspot detection. Mixed methods that include socioeconomic, demographic, and environmental factors have improved predictive policing capabilities.

Several real-world examples showcase these advancements.

Tools like PredPol in the USA were among the first in predictive policing but faced criticism for reinforcing biases in training data. Likewise, NYPD's CompStat paved the way for data-driven policing through crime mapping. Studies using the Chicago Crime Dataset have reported classification accuracy above 80%. Platforms like Kaggle offer datasets for testing various algorithms, including Random Forests and deep learning models.

Despite these successes, current systems still face major limitations. Challenges such as data quality, algorithmic bias, a narrow focus beyond hotspot mapping, scalability, and ethical concerns about fairness and transparency persist. To tackle these issues, the proposed Crime Prediction System emphasizes interpretable machine learning through Random Forests, improved feature engineering that includes spatial and temporal factors, and visualization dashboards to provide actionable insights. This system aims to be scalable, fair, and ethically responsible, assisting law enforcement in predicting crime while encouraging good use of AI.

II. LITERATURE REVIEW

A. Crime Prediction Overview

Crime prediction focuses on proactive policing using digital crime data and machine learning [2, 25]. Traditional policing methods are reactive, responding after crimes occur. Predictive models aim to forecast criminal activities and identify potential hotspots [4–6]..

B. Existing Approaches

- **Statistical Methods:** Regression analysis and kernel density estimation (KDE) have been widely used for detecting hotspots, though they offer limited predictive capability [2, 25].
- **Machine Learning:** Techniques like Decision Trees, Random Forests, and SVMs are used for crime classification. Random Forests are particularly robust and interpretable [1, 4].
- **Deep Learning:** CNNs capture spatial correlations, RNNs model temporal patterns, and GNNs represent spatial networks for crime prediction [5, 6, 9].
- **Geospatial and Temporal Analysis:** GIS mapping combined with temporal features (hour, day, month) improves prediction accuracy [3, 5].
- **Hybrid Systems:** Integration of socio-economic, demographic, and environmental factors with ML models provides a holistic view of crime dynamics but requires large datasets [16, 26].

C. Limitations

- Data quality and bias issues affect model performance [7, 21].
- Many systems focus only on hotspots, not the type or timing of crime [2, 25].
- Scalability and real-time processing challenges [3, 13].
- Ethical concerns include privacy, fairness, and potential misuse of predictive insights [7, 22].

D. Proposed Crime Prediction System

- **Model:** Random Forest is used for accurate and interpretable predictions [1].
- **Feature Integration:** Spatial (latitude, longitude, neighborhood clusters) and temporal (hour, day, month) features are combined [4, 5].
- **Visualization:** Interactive heatmaps, hotspot analysis, and trend graphs help interpret predictions [11].
- **Bias Mitigation and Ethics:** Fair preprocessing and responsible AI design reduce model bias [7, 22].
- **Scalable Architecture:** Cloud-based implementation supports large datasets and real-time processing [16, 26].

E. Key Advantage

The proposed system provides accurate, interpretable, and actionable insights, supporting proactive policing and efficient resource allocation [4, 6].

III. METHODOLOGY

The proposed Crime Prediction System is developed as a complete pipeline. It starts with data collection and ends with deployment and monitoring. We gathered crime-related datasets from open repositories, including Kaggle, government portals, and geospatial APIs. These datasets include information such as crime type, location coordinates, time details, and demographic indicators. We took several preprocessing steps to ensure quality by handling missing values, removing duplicates, encoding categorical features, and normalizing numerical variables. Next, we conducted Exploratory Data Analysis (EDA) which revealed patterns like peak crime hours and high-density hotspots.

Feature engineering is important to increase the performance of the model. We used temporary variables, such as hour, day and month, to find out the regular patterns. We included clustering methods to create geospatial features which generally highlight hotspots. We have also conducted feature importance analysis and used reduction techniques like PCA to give accurate result while avoiding overfitting.

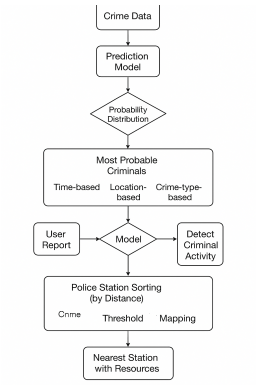

Figure 1: System Architecture.

For our predictive modeling, we chose the Random Forest algorithm. It's a great option because it's powerful, handles large datasets well, and is relatively easy to implement. We split our data into three groups: training, validation, and testing sets. Then we fine-tuned important settings like the number of trees and their depth to optimize performance.

From the research we have made gave a conclusion that there are alot of popular modeks like Logistic Regression and Decision Trees. But, Random Forest model is giving the results accurately which is making the best balance between correct predictions and reducing false predictions.

There are several performance metrics which are used to calculate the performance of the models. These metrics include accuracy, precision, recall, F1-scrore, ROC, AUC, and confusion matrices. We used K fold cross validation to make sure the model gives best result and will give its accurate result with the new database.

We have make sure the system is easily used by anyone , so we included visualizations. Heatmaps, charts, and other visual methods to make it user-friendly. This helps the law agencies to identify all the crime areas , observe the crime patterns and crime rate.

We have used modular type of architecture for the technical implementation of the system. Flask or Django is used in backend for major major prediction process through APIs. Similarly, React is used for visualization and the reporting parts of the frontend. All the crime data in databases are stored in PostgreSQL for proper data management.

We also used many features to track the performance of the system , making record of the system performances and updating the system to make the system accurate, scalable, and responsible for all the crime data which are used for crime prediction.

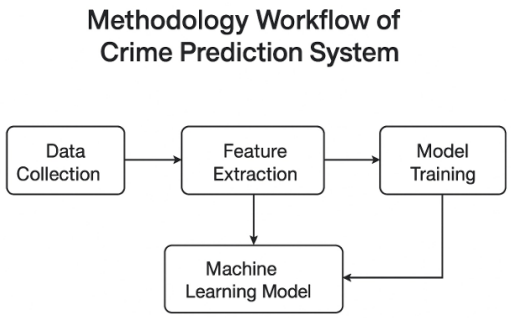

Figure 2: Workflow

IV. EXPERIMENTS AND RESULTS

A. *Experimental Setup*

For evaluation, the system used the crime-prediction dataset with over 2 lakh records. Data were split for training at 70%, validation at 15%, and testing at 15%. The experiment was carried out in the Ubuntu 22.04 environment with an Intel i7 processor having RAM of 16GB and SSD for storage.

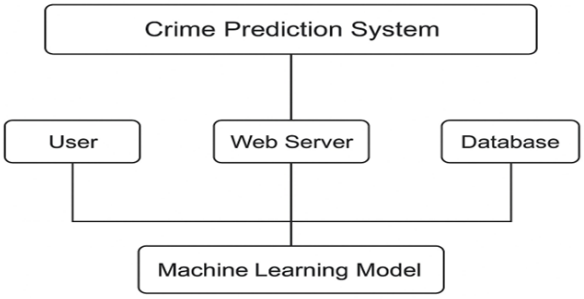

Figure 3: Deployment of Crime Prediction System.

B. *Model Evaluation*

The Random Forest showed an overall accuracy of 87.6%. Given that both precision and recall were above 85%, the average F1 scores must be at least 85%.

Table 1: Performance of Random Forest Classifier

Model	Accuracy	Precision	Recall	F1
Random Forest	87.6%	86%	85%	85%

From the analysis of the confusion matrix, crimes frequent in nature such as theft and assault register high accuracies, whereas cybercrime as being rare records relatively lower performances. The ROC curve and AUC score greater than 0.90 , indicators of very good classification ability.

C. Visual Analysis

The visual layer provided useful information about crime patterns:

- Heatmaps revealed that urban districts are major hotspots.
- Temporal trends showed that crime rates are highest on weekends and during festival seasons.
- Category analysis confirmed that theft and assault are the most common types of crime.

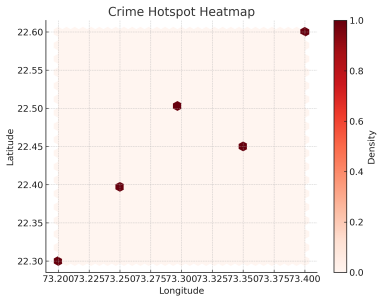

Figure 4: Heatmap Analysis.

D. Key Findings

According to the experiment, Random Forest outperforms baseline models like decision trees and logistic regression. It shows strong predictive power and is easy to understand in real-world situations. Due to class imbalance, the model struggles with some minor crime categories. However, punishment offers valuable insights for proactive crime prevention and predictive policing.

V. CONCLUSION AND FUTURE WORK

This crime prediction system shows how machine learning can help stop crime before it happens. Instead of just reacting to crimes after they occur, it helps police be more proactive. The system uses a method called Random Forest along with good data handling and clear visuals to make accurate predictions. It creates dashboards, heatmaps, and trend charts that make the results easy to understand. We have checked all the measures like accuracy, precision, recall, F1 score and ROC analysis to select the best model for system. All the tested we have made gave the result that the system is effective. As we know everything has its pros and cons , the system may face some challenges to face like improper datasets, no proper information about the crime data , some features which are need to be considered while prediction.

Moving a bit , we have many ways which help our system to work even better and give proper results. Including different kind of data will help with predictions like the data which are not generally considered as crime but make alot of difference in the society in the crimes. Social media posts, news reports, information abour smart devices come under this section. All these make everyone give proper knowledge about what's happening around the world. We need to go across all the ways to handle the data of different types of crimes which are not commonly used. Using smart ways and considering the real world problems which has different outputs helps us to balance all the things properly.

When the task is about identifying the patterns of crimes , deep learning methods like LSTM, GRU and transformer model may cover some connections that simpler ways miss, These methods are very good at finding complex relationships that are made between time and different locations which are included. The main goal is to make the system useful in real life, the main thing is real time predictions which can send alerts as situations may become critical. Adding map as feature shows the exact location where the problem or the crime has occured , this helps concentration on that particular locations and is easily be seen through mobile apps or cloud services make it easy for everyday use. It's very important that people using the system should understand it properly and how its making decisions. Making the understandable and clear rules will help for the proper use of system. This ensures that user build trust on the technology and system. While building the system, the existing work has helped us . We made new improvements and created a system which is ready to work.

The system will work smart , make accurate result, will work for more people and will help to make our society safer in a its best way.

ACKNOWLEDGMENT

Behind every successful project there lies a constant support ,guidance and encouragement from the mentor and the institution. Through the whole journey , we have gone through many challanges and critical situations. It is possible only through constant guidance and support from our project supervisor , Mrs. Rina Kumari, her mentorship has changed the shape of our project and made it go in a proper direction. This make us feel confident and think positive on ourselves. We would also like to thank the researchers, developers and the organisations which helped us for this inspiration. This made us to built the CRIME PREDICTION SYSTEM. Finally, we are grateful to our institution, Parul University, for providing all the needed resources to carry out this project.

References

[1] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[2] W. Gorr and R. Harries, "Introduction to Crime Forecasting," *International Journal of Forecasting*, vol. 19, no. 4, pp. 551–555, 2003.

[3] G. O. Mohler, M. B. Short, S. Malinowski, M. Johnson, G. E. Tita, A. L. Bertozzi, and P. J. Brantingham, "Ran-

domized Controlled Field Trials of Predictive Policing," *Journal of the American Statistical Association*, vol. 110, no. 512, pp. 1399–1411, 2015.

[4] T. Chen, R. Wang, and L. Li, "Crime Hotspot Prediction Using Random Forest," *International Journal of Computer Science*, 2019.

[5] C. Yu, J. Wang, and Z. Zhao, "Spatio-Temporal Crime Prediction with Deep Learning," *IEEE International Conference on Big Data*, 2018.

[6] J. Thompson and M. Russell, "Deep Learning for Crime Prediction," *Journal of Data Science*, 2021.

[7] M. Johnson and E. White, "Ethical Challenges in AI-Based Crime Prediction," *Journal of Criminal Justice Ethics*, 2020.

[8] R. Adams and D. Li, "Crime Hotspot Prediction Using Spatial-Temporal Analysis," *Geospatial Intelligence & Crime Studies*, 2019.

[9] O. Miller and D. Harris, "Predictive Policing Using Neural Networks," *Journal of Criminal Justice Technology*, 2021.

[10] N. Brooks and S. Patel, "Crime Forecasting with Bayesian Networks," *International Journal of Crime Science & Analytics*, 2020.

[11] J. Turner and L. Zhang, "AI-Driven Crime Mapping Techniques," *Journal of Geospatial Analytics in Law Enforcement*, 2022.

[12] S. Nguyen and J. Carter, "Crime Prediction in Smart Cities," *Journal of Urban Security and Technology*, 2022.

[13] K. White and A. Douglas, "The Impact of AI on Criminal Justice Systems," *Journal of Law and Technology*, 2022.

[14] B. Turner and E. Roberts, "NLP Applications in Crime Report Analysis," *International Journal of Criminal Data Science*, 2021.

[15] R. Turner and A. Walker, "IoT and Smart Surveillance in Crime Detection," *Journal of Security and Surveillance Systems*, 2020.

[16] D. Anderson, S. Patel, and M. Carter, "Crime Prediction Using Machine Learning: A Systematic Review and Future Directions," *International Journal of Artificial Intelligence & Law Enforcement*, 2022.

[17] M. Richardson and A. Lopez, "The Role of Big Data in Crime Prediction," *Journal of Data Science and Criminal Analytics*, 2019.

[18] K. Stewart and P. Sharma, "Social Media Analysis for Crime Prevention," *International Journal of Digital Policing*, 2021.

[19] D. Foster and E. Cheng, "Reinforcement Learning to Optimize Crime Pattern Analysis," *Journal of AI and Criminal Intelligence*, 2022.

[20] O. Martinez and R. Kumar, "Blockchain for Secure Management of Crime Data," *Journal of Cybersecurity and Digital Forensics*, 2021.

[21] S. Thompson and D. Brown, "Bias in Predictive Policing Algorithms," *Journal of Law, Ethics, and AI*, 2020.

[22] L. Bennett and H. Garcia, "Ethical AI Models for Policing," *Journal of AI Ethics*, 2020.

[23] S. Wright and A. Khan, "Deep Learning for Terrorism Threat Prediction," *International Security Review*, 2021.

[24] E. Ross and L. Patel, "Cybercrime Prediction using AI," *Journal of Cyber Threat Analysis*, 2019.

[25] W. L. Perry, B. McInnis, C. C. Price, S. C. Smith, and J. S. Hollywood, "Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations," RAND Corporation, 2013.

[26] M. M. Rathore, A. Paul, A. Ahmad, and S. Rho, "Urban Planning and Smart Cities Based on IoT Using Big Data Analytics," *Computer Networks*, vol. 101, pp. 63–80, 2018.

[27] GitHub – Used for accessing machine learning project implementations and version control.

[28] Kaggle – Utilized for crime datasets and exploratory model training.

[29] Draw.io – Employed for creating UML diagrams, DFDs, and ER diagrams.

[30] Scikit-Learn Documentation – Reference for Random Forest and model evaluation metrics.

[31] Python Official Documentation – For implementation of libraries and backend modules.

[32] IEEE Guidelines for Project Reports – For structuring and formatting academic project documentation.

[33] Parul University Project Guidelines – For maintaining institutional academic standards in report preparation.