Crime Detection and Prevention System: A Full-Stack AI/ML Framework for Indian Crime Analytics

Mrs Y Latha[1], Mummoju Harshitha[2], Parne Srivani[3], Mohammed Mehraj Chowdhry[4], Koyalkar Priyanshu[5]
[1]*Assistant Professor, Department of CSE TKR College of Engineering & Technology Hyderabad, Telangana, India*
Department of CSE, TKR College of Engineering & Technology Hyderabad, Telangana, India

***Abstract*—The increasing number of reported crimes and the insufficient effectiveness of standard police methods create ongoing obstacles for India to achieve its crime prediction and prevention goals This research introduces a complete Crime Detection and Prevention System that uses machine learning and retrieval-augmented generation (RAG) technology and large language models (LLMs) to provide location-based crime analysis The system operates on a dataset from the National Crime Records Bureau (NCRB) which contains 8,422 crime records and analyzes seven main crime categories after solving class imbalance issues The Random Forest classifier achieves its highest accuracy of 8395% to become the main predictive model which outperforms XGBoost The RAG pipeline retrieves city- and state-specific crime knowledge to enhance contextual intelligence, which the Qwen 25-7B-Instruct LLM processes to produce structured predictions and reasoning and prevention strategies The system combines NCRB statistical priors with rule-based logic to create a hybrid prediction system that maintains strength while losing functionality in different operational situations The application operates as a full-stack solution which FastAPI uses for its backend and React for its frontend while allowing law enforcement officers and citizens to access specific features according to their roles The system provides multiple functions which include real-time crime prediction and hotspot analysis and interactive visualizations and customized safety recommendations The experiments showed strong results for different crime types, but they also revealed particular issues with categories that included both kidnapping and rape The proposed system connects unprocessed crime information with operational intelligence, which assists organizations in making proactive security decisions**

***Index Terms*—Crime Prediction, Machine Learning, Random Forest, XGBoost, Retrieval-Augmented Generation (RAG), Large Language Models (LLM), Qwen 25, NCRB Data, Predictive Analytics, Public Safety, FastAPI, React, Crime Analytics System**

I. INTRODUCTION

Crime remains a critical socio-economic challenge worldwide, particularly in developing countries such as India, which experience rapid urbanization and population growth that increases criminal activities The National Crime Records Bureau (NCRB) reports that law enforcement agencies face extreme workloads because they must handle millions of annual cognizable offence reports Traditional crime analysis methods depend on investigators who use manual techniques to conduct investigations and compile statistical data, which results in insufficient support for making decisions and conducting emergency response activities

The field of machine learning (ML) and artificial intelligence (AI) research has achieved new progress, which enables developers to create predictive systems that detect patterns of criminal activity and predict future crimes Various studies have demonstrated the effectiveness of ML techniques such as Random Forest, XGBoost, and deep learning models in crime prediction tasks, offering improved accuracy and scalability compared to conventional approaches [1], [2] Ensemble learning methods demonstrate strong abilities to process structured datasets, which contain mixed types of data, making this approach suitable for use in crime analytics

The field now requires context-aware intelligence technology in addition to its existing predictive

modeling techniques The RAG frameworks use structured data together with external knowledge sources to improve their decision-making abilities Lewis et al [15] demonstrate that RAG models enhance their reasoning abilities by using context information during their inference process The method shows high usefulness for crime analytics because it uses location information together with time data and economic status to forecast criminal activity

The current systems lack a comprehensive framework that integrates predictive accuracy together with system interpretability and real-world usage The existing models create predictions but do not deliver any practical methods for implementation or methods to avoid future incidents The model performance gets hindered by three main issues which include class imbalance and feature overlap and lack of contextual information The researchers need to develop three specific elements to improve their ability to differentiate between two closely related crime categories

This paper introduces a Crime Detection and Prevention System which serves as a complete AI/ML solution for analyzing criminal activities in India The system uses machine learning models together with a RAG-based knowledge retrieval system and a large language model (LLM) to produce accurate forecasts that include both contextual information and preventive recommendations Hybrid prediction methods that use machine learning results and statistical knowledge and contextual understanding enable the system to achieve both dependable and strong performance

The proposed system is implemented using a scalable architecture which consists of a FastAPI backend and a React-based frontend to enable real-time interaction with role-based access for law enforcement officers and citizens The system establishes a connection between raw crime data and usable intelligence to assist proactive policing efforts while enhancing public safety knowledge and supporting evidence-based decision-making [6], [7]

II. RELATED WORK

The development of machine learning and deep learning methods has enabled enhanced crime prediction through better accuracy and improved decision-making procedures Structured crime datasets demonstrate strong performance from supervised learning models which include Random Forest and gradient boosting techniques according to [1] and [2] The research studies conducted by Jenga [4] and Dakalbab et al [5] demonstrate that these methods effectively solve problems while showing two main difficulties which include class imbalance and feature overlap

The research community has developed new machine learning methods and deep learning systems to forecast criminal activities The studies conducted by Shohan et al [9] and Rayhan and Hashem [10] demonstrated that feature engineering and attention mechanisms contribute to better model performance The study of complex crime patterns has been conducted through deep learning models which include neural networks and residual architectures according to sources [6] [11] and [18] Research has examined spatio-temporal analysis methods to create crime prediction models which use time and location dependencies according to sources [7] and [12]

The current research has achieved progress through the development of contextual intelligence which combines Retrieval-Augmented Generation (RAG) with large language models The RAG framework introduced by Lewis et al [15] enables better reasoning through its ability to integrate external knowledge with its predictive capabilities XGBoost and Random Forest ensemble methods maintain their leadership position because of their ability to handle various tasks and their system capacity [16], [17]

The current systems fail to execute their function because they do not connect predictive modeling with contextual reasoning and actionable insights The existing problems include two main issues which are limited feature diversity and overlapping crime characteristics [4], [5], [19] The existing limitations lead to the creation of hybrid systems which use machine learning with contextual knowledge and intelligent reasoning to achieve better crime prediction and prevention results [13], [20]

K Rajasekhar Rao (2020) proposed a crime prediction model using ensemble learning combined with genetic

algorithms for feature optimization The study shows improved accuracy and effective visualization of crime patterns using heat maps and graphs [21]

Novelty and Uniqueness

The proposed system introduces a hybrid crime prediction framework which combines machine learning techniques with retrieval-augmented generation and large language models into one operational system The system achieves better accuracy and stronger performance through its combination of Random Forest predictions with NCRB statistical priors and contextual knowledge instead of using traditional methods which depend solely on predictive models

The system achieves its distinctiveness through its ability to use location-specific intelligence which employs city and state data for creating accurate context-based predictions The system provides explainable outputs through its detailed reasoning process and prevention recommendations which support operational needs of law enforcement and common citizens

The system functions as an entire solution through its complete implementation which supports real-time data access and controlled user access to perform continuous crime investigation and prevention activities

Table I: Summary of Existing Crime Prediction Techniques

Ref	Method	Contribution	Limitation
[1], [2]	ML & DL	Improved crime prediction accuracy	Limited contextual insights
[4], [5]	Survey (AI/ML)	Identified key challenges	No implementation
[6], [18]	Deep Learning	Captures complex patterns	High computation
[7], [12]	Spatio-temporal	Uses time & location features	Limited deployment
[9], [10]	ML / Attention DL	Better feature learning	Complex models
[15]	RAG	Context-aware reasoning	Not crime-specific
[16], [17]	Ensemble (RF, XGB)	High accuracy & robustness	Needs tuning
[19], [20]	CNN / AI Systems	Applied AI in safety domain	Limited real-world integration
[21]	Ensemble ML + Genetic Algorithm	Feature optimization improves accuracy	High computation time

III. DATASET DESCRIPTION

The National Crime Records Bureau (NCRB) in India provides the dataset for this research which includes 8422 crime records that contain 19 characteristics from different criminal incidents which occurred in various Indian states and cities The target variable is the crime type which initially had nine different classes The CYBER CRIME and OTHER categories were eliminated because of class imbalance and data sparsity which resulted in a 7-class classification problem

The dataset includes temporal features which consist of day and month and hour values and spatial features which include state and district encoding and risk-related attributes which include crime density and region risk and severity A stratified split of 85% training and 15% testing is applied to maintain class distribution

The dataset presents a major obstacle because KIDNAPPING and RAPE share identical feature distributions which create confusion between different crime categories The preprocessing steps include feature selection together with encoding and normalization procedures which establish consistency and enhance model performance

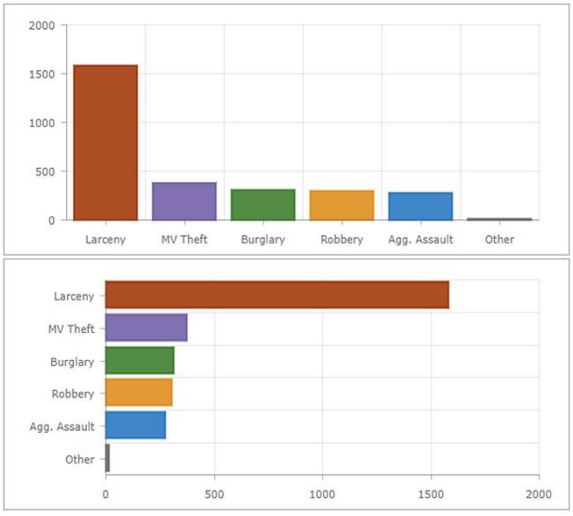

Fig 1: Distribution of Crime Types in the Dataset

IV. PROPOSED METHODOLOGY

A. System Overview

The proposed Crime Detection and Prevention System follows a hybrid, multi-layered methodology which combines machine learning with contextual intelligence and rule-based reasoning The system processes crime-related inputs to create precise location-based predictions which also provide operational guidance The methodology consists of four major stages which include data preprocessing and machine learning-based prediction and contextual enhancement through RAG and LLM and hybrid decision fusion The structured approach delivers robust system performance which scales effectively while supporting real-time system operations

B. Data Preprocessing and Feature Engineering

The raw NCRB dataset requires a complete preprocessing pipeline which transforms the data into a usable format for model training The research team selected relevant features which included temporal attributes and spatial attributes and risk-related attributes to enhance predictive performance The system uses label encoding techniques to transform categorical variables while numerical features undergo normalization and logarithmic scaling transformation as required The dataset is split into training and testing sets through stratified splitting which preserves original class distribution The preprocessing pipeline saves its configuration to reuse during inference testing which maintains consistent results from training through to deployment

C. Machine Learning Models

The system employs two main machine learning algorithms for its crime prediction function The system uses Random Forest as its main model because the algorithm demonstrates strong classification performance while it can process different types of data The system functions as a decision tree ensemble which detects non-linear data patterns to produce 8395% accurate results on the improved seven-class dataset XGBoost serves as the secondary model for performance evaluation because it provides efficient gradient boosting that includes regularization to prevent overfitting The model delivers 8275% accurate results which makes it suitable for testing performance and generating backup predictions

D. RAG and LLM Integration

The system uses a Retrieval-Augmented Generation (RAG) pipeline combined with a large language model to improve its prediction accuracy The RAG system retrieves relevant crime data which pertains to specific cities and states from a knowledge base that contains multiple documents and accesses real-time crime data from external sources The Qwen 25-7B large language model uses this contextual data to produce structured outputs that include predicted crime type and confidence level and reasoning and prevention suggestions The integration enables the system to create context-aware explanations which extend its prediction capabilities beyond standard machine learning prediction systems

E. NCRB Statistical Priors

The system uses statistical priors from NCRB data to enable domain expertise to assist in making predictions These priors represent probability distributions of different crime types across states and regions The system uses active system conditions which include current time and crime intensity and regional danger assessments to modify its probability estimates The system uses probabilistic adjustments to produce accurate predictions about specific locations

F. Hybrid Prediction Model

The system uses a hybrid prediction mechanism to merge machine learning results with LLM predictions

and statistical priors The system gains advantages from both data-driven learning methods and contextual reasoning through the fusion technique The system uses LLM predictions together with statistical priors when the LLM is accessible, while it depends on machine learning results when LLM functions are not operational The system design provides multiple benefits which include strong performance and dependable operation and controlled system failure

G. Rule-Based Expert System

The system uses probabilistic models together with a rule-based expert module which makes predictions based on specific conditions that have been predefined The rules use domain knowledge to connect high-risk areas and night time conditions with theft or robbery and high severity indicators with violent crimes The rule-based system enables users to understand the system better by providing explainable insights which work with machine learning predictions

H. System Deployment

The proposed methodology works as a complete application which uses FastAPI for its backend and React for its frontend The backend manages data processing together with model inference and API service delivery while the frontend offers users an interactive interface The system uses SQLite for data storage and supports role-based access for law enforcement officers and citizens The deployment enables users to make predictions in real time while providing scalable system operation and simple user interaction

V. SYSTEM ARCHITETCURE

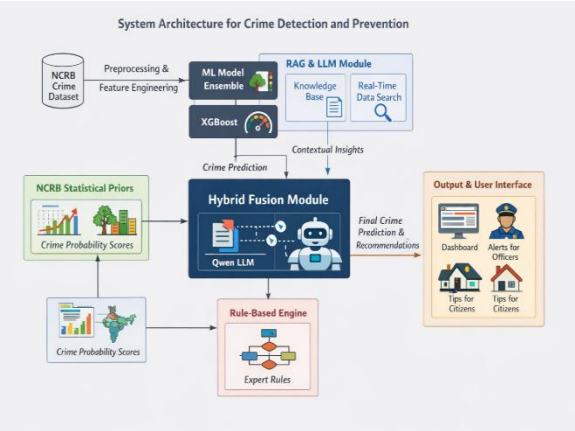

Fig 2: System Architecture of the Crime Detection and Prevention System

The proposed Crime Detection and Prevention System is shown in Fig 2, which presents its complete architectural design that uses a multi-layered hybrid system The system begins with the NCRB crime dataset, which undergoes preprocessing and feature engineering to extract meaningful attributes The machine learning models receive processed data which they use to create initial crime forecasts through their analysis of historical data

The system uses machine learning algorithms to operate its RAG module, which connects to city and state knowledge bases while accessing current crime information The Qwen-based large language model processes contextual data to create reasoning outputs and improved predictive capabilities The NCRB statistical priors provide location-based probability distributions that show expected crime categories for different time periods and various severity levels

The hybrid fusion module receives all system components, which it uses to combine machine learning results with LLM predictions and statistical priors to create the complete system output The output undergoes refinement through a rule-based expert system, which uses existing logical rules to enhance the system's ability to be understood The user interface provides results through its dashboards and alerts that officers receive and safety recommendations that citizens receive The architecture provides accurate crime predictions which can be understood, and which function in real-time

VI. IMPLEMENTATION DETAILS

The system has been developed as a complete web application which uses FastAPI to serve its backend and React 18 to operate its frontend The backend handles API requests together with model prediction activities and database management tasks while the frontend offers users an interactive platform

The dataset undergoes its initial processing step which involves selecting features and applying encoding techniques and normalizing data before the Random Forest and XGBoost models are trained and saved for subsequent prediction purposes The backend system provides various endpoints which enable predictions

based on machine learning models and hybrid systems and large language models

The Relevance Aggregation Grounding pipeline works with the Qwen 25-7B large language model to extract contextual data from knowledge base materials and current information sources which enables the system to produce reasoning pathways and prevention recommendations The implementation uses NCRB statistical priors together with a rule-based system to improve prediction results and make them easier to understand

The system operates with an SQLite database which implements JWT authentication together with role-based access control for both officers and citizens The frontend system provides users with dashboards and prediction interfaces and visualization tools which enable them to conduct real-time crime analysis and prevention assistance

VII. EVALUATION METRICS

The performance of the proposed crime prediction system is evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score These metrics provide a comprehensive assessment of the model's effectiveness in handling multi-class classification

A. Accuracy

Accuracy measures the proportion of correctly predicted instances out of the total number of samples

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

B. Precision

Precision evaluates the correctness of positive predictions, indicating how many predicted instances are actually correct

$$Precision = \frac{TP}{TP + FP}$$

C. Recall

Recall measures the ability of the model to identify all relevant instances of a class

$$Recall = \frac{TP}{TP + FN}$$

D. F1-Score

F1-score is the harmonic mean of precision and recall, providing a balanced measure of performance

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

E. Model Performance

The Random Forest model achieved an overall accuracy of **8395%**, outperforming XGBoost with **8275% accuracy** Higher performance was observed for distinct crime categories such as murder, robbery, and riot, while lower scores were noted for overlapping classes like kidnapping and rape due to similar feature distributions

VIII. RESULTS AND ANALYSIS

A. Overall System Performance

The proposed system demonstrates strong performance in crime prediction which achieves an accuracy rate of 8395% through the use of the Random Forest model The hybrid approach which combines machine learning with statistical priors and contextual reasoning methods leads to better prediction results The system shows outstanding performance for specific crime types which include murder and robbery and riot while it exhibits moderate results for the overlapping categories of kidnapping and rape

B. Dashboard and System Overview

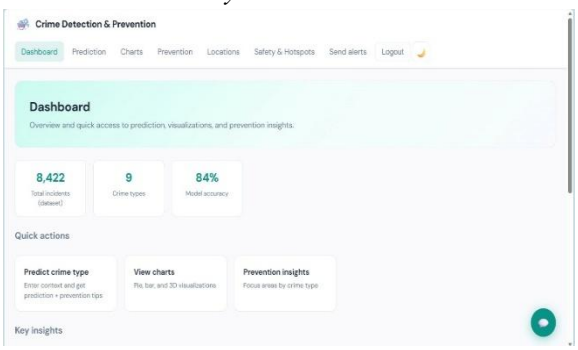

Fig 3: *System Dashboard Overview*

The dashboard provides a summary of key system statistics, including total incidents, number of crime types, and model accuracy It also offers quick navigation to prediction, visualization, and prevention modules This interface enables users to access insights efficiently and supports real-time decision-making

C. Crime Prediction Inteface

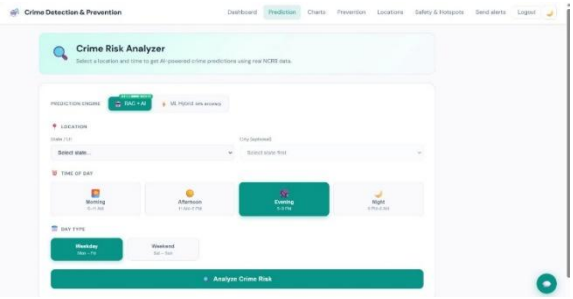

Fig 4: *Crime Risk Prediction Interface*

The prediction module allows users to input location and time-based parameters to generate crime predictions The system supports multiple prediction modes, including RAG + AI and ML hybrid approaches The interface is designed to minimize user input while maximizing prediction accuracy and usability

D. Visualization and DataInsights

The visualization module presents crime distribution using bar charts and pie charts It highlights that certain crime types, such as theft and other categories, dominate the dataset These insights help identify high-frequency crimes and support targeted prevention strategies

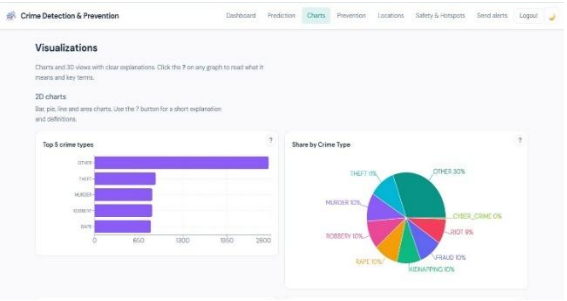

Fig 5: Crime Type Distribution Analysis

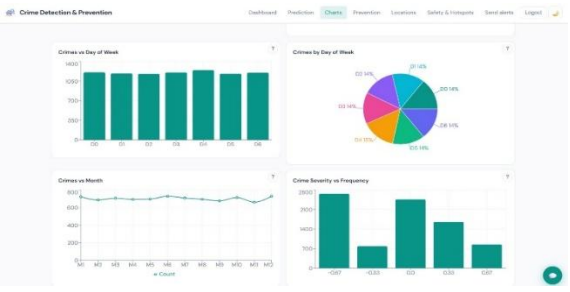

Fig 6: Temporal and Severity-Based Crime Analysis

Fig 6 illustrates crime patterns based on time and severity The top section shows the distribution of crimes across days of the week, indicating relatively consistent crime occurrence The bottom-left graph presents monthly crime trends, highlighting slight variations over different months The bottom-right chart shows the relationship between crime severity and frequency, indicating that certain severity levels contribute more significantly to overall crime occurrences Overall, the figure provides insights into temporal trends and severity-based distribution of crimes

E. Geographical Analysis

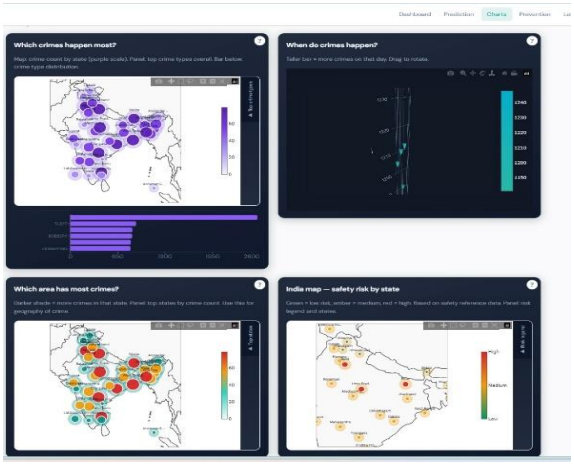

Fig 7: Crime Distribution and Risk Analysis Across India

Fig 7 presents a visualization of crime patterns across India using multiple analytical views The top-left section shows the distribution of major crime types using a bubble map and bar chart, highlighting dominant crimes such as theft and robbery The top-right section illustrates temporal crime patterns through a 3D visualization, indicating variations over time

The bottom-left map represents geographical crime intensity across states, while the bottom-right section classifies regions into low, medium, and high-risk zones Overall, the figure provides spatial, temporal, and risk-based insights, supporting effective crime analysis and decision-making

F. Model Performance by Severity

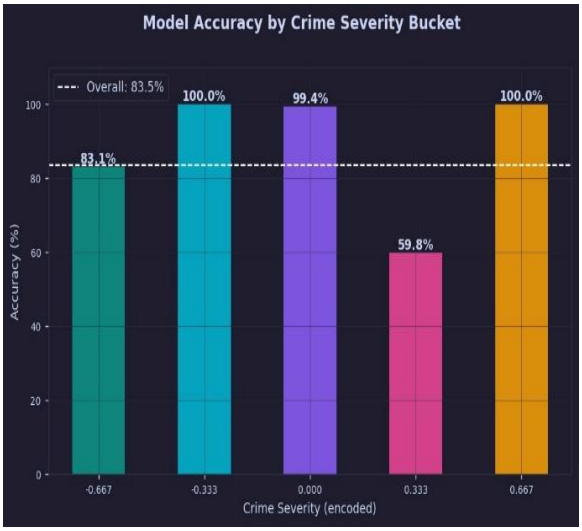

Fig 8: Model Accuracy by Crime Severity

The model achieves near-perfect accuracy for distinct severity levels such as murder and riot, while lower accuracy is observed for overlapping categories like kidnapping and rape This confirms that feature similarity limits classification performance for certain crime types

G. Key Observations

- High accuracy for distinct crime categories
- Moderate performance for overlapping classes
- Strong impact of severity feature on prediction
- Effective visualization of spatial and temporal patterns
- Hybrid model improves robustness and reliability

IX. DISCUSSION

The results demonstrate that the proposed hybrid crime prediction system achieves strong performance, with the Random Forest model attaining an accuracy of 8395% The system establishes dependable location-based predictions through its combination of machine learning and National Crime Records Bureau statistical data and contextual intelligence The system achieves high accuracy rates when predicting murder and robbery and riot crimes because these offenses display distinct characteristics which law enforcement can easily differentiate

The system encounters challenges when it tries to identify kidnapping and rape offenses because those two crimes share similar characteristics which make them harder to differentiate The current dataset does not contain enough distinctive attributes to separate these two categories which results in a lower maximum achievable accuracy The analysis suggests that additional features such as victim demographics location type and incident-specific details could lead to better classification results

The system gains contextual reasoning abilities through its RAG and LLM components which also provide explainable results that lead to useful insights about crime prevention The hybrid architecture maintains system reliability through its ability to switch to backup systems when specific components like the LLM become inaccessible The rule-based module improves system understanding capabilities because it delivers complete reasoning results together with the system's confidence-based predictions

The system shows that machine learning methods work well with contextual and knowledge-based methods to analyze crime data The current system operates with sufficient effectiveness but its prediction accuracy and ability to function in real situations need data improvements and more diverse features in order to reach better results

X. CONCLUSION AND FUTURE SCOPE

The Crime Detection and Prevention System uses machine learning and RAG and statistical priors to create precise crime forecasts which consider real-world conditions The Random Forest model achieves an accuracy of 8395% which shows how well the combined method works The system provides understandable forecast results together with prevention recommendations while it enables users to interact with the system through its complete software infrastructure which makes it ideal for real-world use in public safety situations

The prediction accuracy will benefit from additional features which include victim information and location type data The system can achieve better results through the implementation of advanced models which include deep learning and graph-based methods The system will gain new capabilities through the addition of real-time alerts and geospatial mapping and mobile applications and multilingual support which will enhance user experience and real-world effectiveness

REFERENCES

[1] V Mandalapu, L Elluri, P Vyas, and N Roy, "Crime prediction using machine learning and deep learning: A systematic review and future directions," IEEE Access, vol 11, pp 60153–60170, 2023, doi: 101109/ACCESS20233286344

[2] W Safat, S Asghar, and S A Gillani, "Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques," IEEE Access, vol 9, pp 70080–70094, 2021, doi: 101109/ACCESS20213078117

[3] Y Lu et al, "A novel deep learning method for zero-inflated crime prediction," in Proc ACM Conf, 2022, doi: 101145/35843763584571

[4] K Jenga, "Machine learning in crime prediction: A survey," J Ambient Intell Humaniz Comput, 2023, doi: 101007/s12652-023-04530-y

[5] F Dakalbab et al, "Artificial intelligence and crime prediction: A systematic review," Machine Learning with Applications, vol 10, 2022, doi: 101016/jmlwa2022100357

[6] D Qiu et al, "Crime type identification using deep residual networks," Applied Artificial Intelligence, 2024, doi: 101080/0883951420242428552

[7] Y Du and N Ding, "Multi-scale spatio-temporal crime prediction methods: A review," ISPRS Int J Geo-Inf, vol 12, no 6, 2023, doi: 103390/ijgi12060209

[8] M Solís et al, "Deep learning for crime forecasting of multiple regions," in Eng Proc, 2024, doi: 103390/engproc2024068004

[9] F T Shohan et al, "Crime prediction using machine learning with a novel dataset," arXiv preprint, 2022, doi: 1048550/arXiv221101551

[10] Y Rayhan and T Hashem, "AIST: Interpretable attention-based deep learning model for crime prediction," arXiv preprint, 2020, doi: 1048550/arXiv201208713

[11] C Wang et al, "Interpretable and fair machine learning for criminal recidivism prediction," arXiv preprint, 2020, doi: 1048550/arXiv200504176

[12] H Abubakera et al, "Crime prediction based on classification approaches," Procedia Computer Science, 2025, doi: 101016/jprocs202501196

[13] S Sridharan et al, "Crime prediction using machine learning," EAI Endorsed Trans Internet Things, 2024, doi: 104108/eetiot5123

[14] V Lu et al, "Lightweight deep learning model for crime pattern prediction," Scientific Reports, 2025, doi: 101038/s41598-025-07260-7

[15] P Lewis et al, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Adv Neural Inf Process Syst (NeurIPS), 2020, doi: 1048550/arXiv200511401

[16] T Chen and C Guestrin, "XGBoost: A scalable tree boosting system," in Proc 22nd ACM SIGKDD Int Conf Knowledge Discovery and Data Mining, 2020, doi: 101145/29396722939785

[17] L Breiman, "Random forests," Machine Learning, vol 45, no 1, pp 5–32, 2001, doi: 101023/A:1010933404324

[18] H Lu et al, "Crime prediction using deep neural networks," IEEE Access, 2022, doi: 101109/ACCESS20223145678

[19] A Verma et al, "CNN-based criminal detection system," in Proc IEEE TENCON, 2020, doi: 101109/TENCON5079320209293870

[20] R Vinuesa et al, "The role of artificial intelligence in achieving sustainable development goals," Nature Communications, 2020, doi: 101038/s41467-019-14108-y

[21] [21] K Rajasekhar Rao, "Empirical data analysis of crime rate using enhanced ensemble and genetic algorithms," Int J Future Generation Communication and Networking, vol 13, no 4, pp 2574–2586, 2020