Naif Arab University for Security Sciences
Journal of Information Security and Cybercrimes Research
مجلة بحوث أمن المعلومات والجرائم السيبرانية
https://journals.nauss.edu.sa/index.php/JISCR
JISCR

Smart Security Analysis System: A Machine Learning-Based Framework for Crime Prediction and Visualization

Abdulghani Abdulwaheed[1], Yamen Megdiche[1] and Qazi Emad Ul Haq*[,1,2]

[1]Department of Cybersecurity and Digital Forensics, College of Forensics & Investigative Sciences, Naif Arab University for Security Sciences, Riyadh 11452, Saudi Arabia

[2]Centre of Artificial Intelligence for Law Enforcement, Naif Arab University for Security Sciences, Riyadh 11452, Saudi Arabia

Received 3 Dec. 2025; Accepted 23 Dec. 2025; Available Online 31 Dec. 2025

Abstract

Addressing the fluid and complex nature of urban crime requires sophisticated, data-centric approaches for the effective distribution of security assets. This research proposes a practical framework that blends open-source crime statistics with machine learning algorithms and geospatial mapping to facilitate operational command. We developed a cohesive Python-based interface capable of processing real-time inputs, performing predictive analytics, and visualizing spatial trends. By utilizing a Random Forest regression model, the system calculates a localized "Risk Index," forecasting crime density relative to specific environmental and temporal variables. Additionally, incidents are stratified by severity levels to refine situational assessment. The distinct value of this study lies not in algorithmic novelty, but in the engineering of standard analytical techniques into a functional decision-support mechanism suitable for security leaders. Experimental validation confirms that the prototype delivers actionable intelligence, underscoring the efficacy of machine learning-driven tools in shaping proactive security operations.

I. Introduction

Foreseeing criminal trends is fundamental to safeguarding the public and managing constrained security assets effectively. Conventional policing often remains reactive, depending largely on past logs and manual interpretation. While these retrospective methods offer some context, they struggle to keep pace with the velocity and volume of modern data streams. The surge in digital crime records offers a pivotal opportunity to pivot towards security models that are preemptive and grounded in analytics.

Criminal incidents rarely occur in isolation; they are patterned events driven by specific spatial, temporal, and environmental contexts. Yet, decoding these complex, multidimensional variables into usable intelligence poses a barrier for analysts, especially given that many commercial solutions are either prohibitively expensive, proprietary, or require niche technical skills to operate. This creates a distinct demand for accessible instruments capable of converting raw statistics into clear, decision-ready metrics for field operations.

Keywords: Crime prediction, data visualization, decision support systems, information security, machine learning

Production and hosting by NAUSS

* Corresponding Author: Qazi Emad Ul Haq
Email: qabdulrab@nauss.edu.sa

doi: 10. 26735/MDXT8104

To bridge this gap, this research introduces a practical framework merging machine learning with geospatial mapping within an interactive interface. The primary aim is not to devise a novel predictive algorithm, but to demonstrate how proven machine learning techniques can be operationalized to enhance situational awareness and resource allocation. By calculating a localized "Risk Index" and mapping severity tiers, the system empowers security commands to visualize risk distribution and preemptively identify high-threat zones. The design prioritizes adaptability, making it effective for both analyzing historical trends and simulating future scenarios.

II. Related Work Materials

Predictive crime analysis has evolved significantly, over the last decade, utilizing a spectrum of statistical and computational methodologies. Consequently, ensemble learning techniques—specifically Random Forest algorithms—have emerged as a pragmatic alternative. These models offer robustness against overfitting and are capable of modeling non-linear relationships between crime rates and contextual features without sacrificing interpretability. Concurrently, advancements in Big Data Analytics (BDA) have facilitated the visualization of large-scale urban datasets, enabling more granular analysis of cross-city security trends.

Big data analytics offers a systematic method of identification of latent patterns, correlations, and trends of massive amounts of data. BDA is used in a study of [1] which uses exploratory data analysis in visualization of criminal data and prediction of trends of crime in San Francisco, Chicago, and Philadelphia. Various state-of-the-art data mining and deep learning methods were used, and it was found that there were meaningful crime trends among the cities. The researchers note that sophisticated time-series forecasting models like Prophet and stateful LSTM outperform the classic neural networks and the optimal length of training window is three years. Such results highlight the efficiency of BDA in increasing the predictive accuracy and informing decision-making in law enforcement agencies.

It is also important to note that a study [2] has emphasized the significance of crime analytics in dealing with the rising crime rates due to urbanization. The authors suggest predictive platform based on K-means clustering and logistic regression on the data received on the authoritative sources like the National Crime Records Bureau (NCRB) and data.gov.in. The system forecasts the incidents of crime according to time, location and the type of crime thus helping the police authorities in the process of proactive planning and allocation of resources. The visualization methods also contribute to the fact that the crime patterns in regions, months, and older people can be interpreted more easily.

The tools of machine learning and deep learning have come into the limelight of predicting crimes because of their capability to pick sophisticated patterns based on previous data. In [3], a thorough overview of more than 150 studies conducted on the topic is provided and the different ML and DL algorithms available in crime prediction are analysed. It also provides the paper with key trends, most frequently used datasets, and algorithmic methods alongside pointing out gaps and future research directions. The authors conclude that the hybrid and deep learning models tend to be more effective than the traditional methods of using ML (especially in large-scale and complex data). An additional systematic literature review by [4] reviews 68 studies that had been selected on machine learning-based crime prediction. The review indicates that the field is dominated by supervised learning methods and this is largely due to availability of labelled crime data. The authors do however also mention the difficulties in the quality of data, no real-life labelled data and ethical aspects. What this research offers are a significant added value of reviewing the already available knowledge and determining the methodological issues that impede the practical application.

In addition to more organized numerical data, there are unstructured text data including crime reports and briefs of cases that provide useful information. Study [5] presents a machine learning approach to the predictions of the type of crime and risk level based on text-based summaries of criminal cases. The system uses real data of policing in KICS-

formatted format to predict a crime type out of 21 and calculate a numerical crime risk score according to its severity and magnitude of the damage. DNNs and CNNs were invented, and CNN-related models are far better than the classical classifiers like Naive Bayes and Support Vector Machines (SVM) with the latter by 8 percent. The paper shows the practical applicability of text analytics in facilitating the provision of a quick response and a risk assessment of newly reported crimes.

Advancement in data collection techniques has greatly widened the crime analysis. Studies [6] point to the change to dynamic and data-driven analysis, which makes use of information shared in social media, IoT devices, surveillance systems, and geographic information systems. The paper puts special focus on the application of deep learning in processing unstructured statistics and determining the sequential patterns and anomalies of crime. Combining socioeconomic and sophisticated AI solutions, the presented framework will help to improve predictive policing and the development of legal strategies. Besides predictive modelling, data security and privacy are the central issues in crime analytics. The paper [7] and [8] give a comprehensive survey of the research concerning the integration of blockchain technology and big data applications. These publications dwell upon the potential of blockchain to promote the safe acquisition of data, storage, analytics, and privacy maintenance in a number of areas, such as smart cities and smart transportation. These works are not crime specific, but can be useful in terms of constructing secure and trustful infrastructures of crime data analytics. Scalability and complexity of the system are some of the challenges outlined by the authors as they also list the research direction in the future that would build a stronger big data service in law enforcement applications.

Despite these technical advancements, a disconnect remains between theoretical modeling and practical application. Much of the existing literature prioritizes algorithmic precision over system usability or deployment strategies. Furthermore, critical ethical dimensions—including algorithmic bias, data feedback loops, and socioeconomic implications—are frequently underrepresented. This study seeks to bridge this gap by prioritizing the operationalization of established analytical methods into a transparent, deployable decision-support framework, while explicitly addressing the associated ethical and methodological constraints.

III. Methodology

A. Materials and Environment

To ensure a robust, industry-standard development lifecycle, the project was implemented using the Windows Subsystem for Linux (WSL 2) running Ubuntu. This provided a stable environment for Python library management. Visual Studio Code (VS Code) served as the primary Integrated Development Environment (IDE), utilizing the Remote-WSL extension to edit and debug code directly within the Ubuntu environment. The core programming language was Python 3.10, utilizing libraries such as Pandas for data manipulation, Scikit-Learn for modeling, Folium for geospatial mapping, and Streamlit for the web interface.

B. Subjects and Data

The data subjects included public crime records from major metropolitan areas (Los Angeles, Chicago, New York City, San Francisco, Riyadh (Fake Data)) accessed via the Socrata Open Data API [9-11]. This API allows for real-time retrieval of datasets including timestamps, crime descriptions, and geolocation coordinates. No human subjects were directly involved in this research. To address the lack of open data for local contexts, a synthetic dataset representing Riyadh was engineered. This simulated data is explicitly marked and serves strictly to validate the system's adaptability to different geographic coordinates, without asserting empirical accuracy for the region.

C. Design and Procedure

The architectural workflow is structured as a linear four-tier pipeline, as illustrated in the system architecture (Fig. 1).

Data Acquisition: Initially, the system establishes a connection to the city-specific API

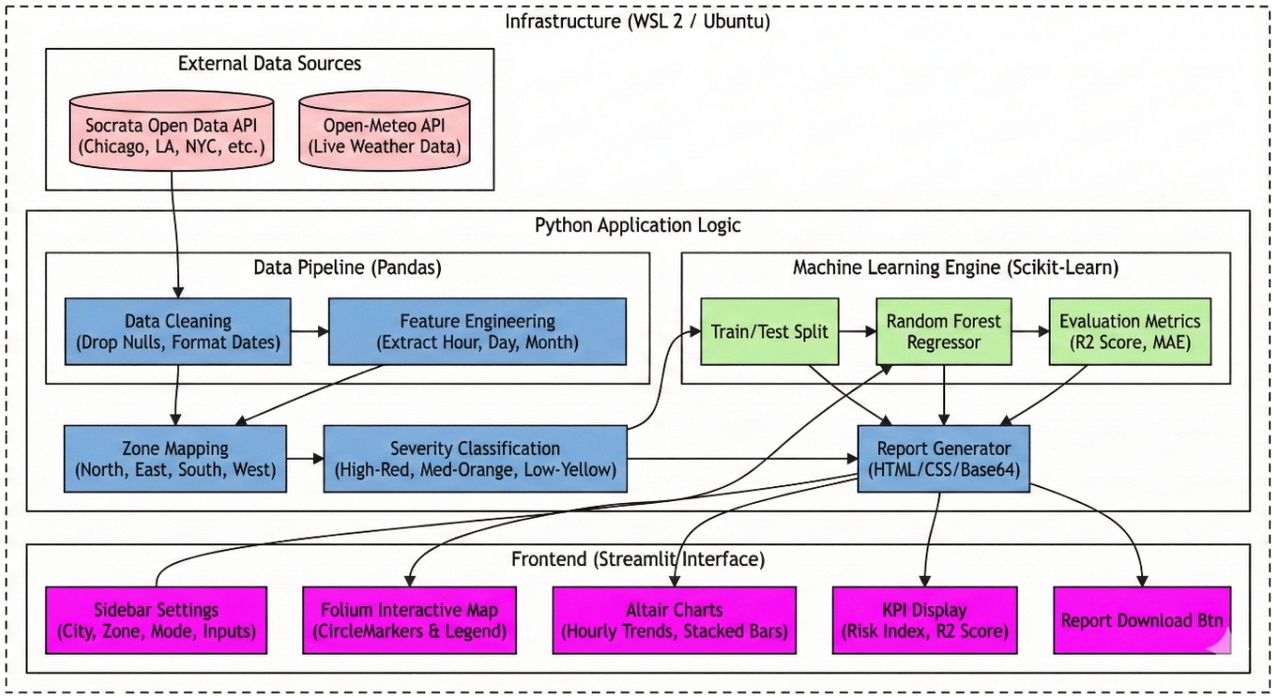

Fig. 1. System Architecture

endpoint to fetch the latest 3,000 to 5,000 incident logs, ensuring operational relevance.

Data Pre-processing: Second, data undergoes normalization where incomplete entries are purged, and temporal features (Hour, Day of Week, Month) are extracted. A custom function categorized crime types into three actionable tiers: High (Red) for violent crimes (e.g., Homicide, Assault), Medium (Orange) for property crimes (e.g., Burglary), and Low (Yellow) for minor infractions. Geographic coordinates were mapped to four distinct cardinal zones (North, South, East, West) relative to the city center. This abstraction facilitates high-level command assessments rather than granular street-level policing. Finally, a Random Forest regressor is trained to project crime density, generating a continuous "Risk Index" that correlates specific time windows with anticipated threat levels.

D. Performance Evaluation

To evaluate the efficacy of the predictive model, the dataset was split into a Training Set (80%) and a Testing Set (20%). The system's quality was assessed using the R2-score to measure the proportion of variance in the dependent variable predictable from the independent variables, and the Mean Absolute Error (MAE) to measure the average magnitude of errors in the predictions. These indicators serve to validate the consistency of the model's logic rather than to claim state-of-the-art predictive superiority. While sufficient for this prototype, future iterations will require more rigorous validation techniques, such as k-fold cross-validation.

IV. Results

A. Significant Quantitative Results

The framework demonstrated a robust capability to ingest live data streams and compute zone-based risk estimates with negligible latency. Real-time validation confirmed the model's stability; the Random Forest algorithm yielded consistent "Risk Index" values, reflecting strong correlations between temporal inputs (time, day) and geographic zones. The system displayed the Model Accuracy R2-score instantly on the dashboard, providing immediate transparency regarding the reliability of the generated predictions.

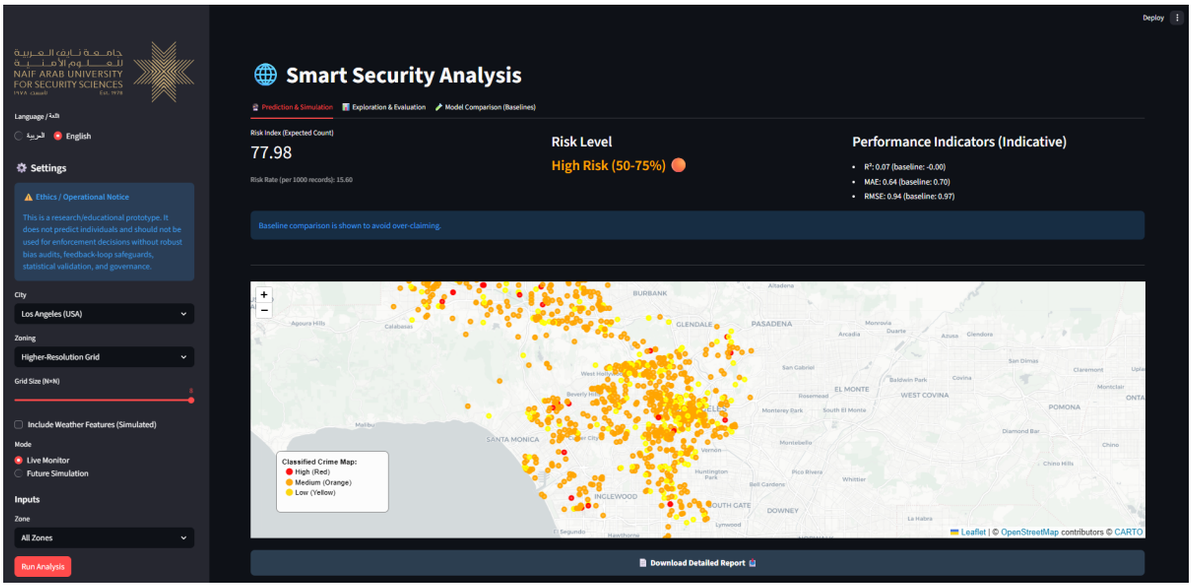

Fig. 2. Interactive geospatial heatmap showing crime clusters categorized by severity (Red: High, Orange: Medium, Yellow: Low)

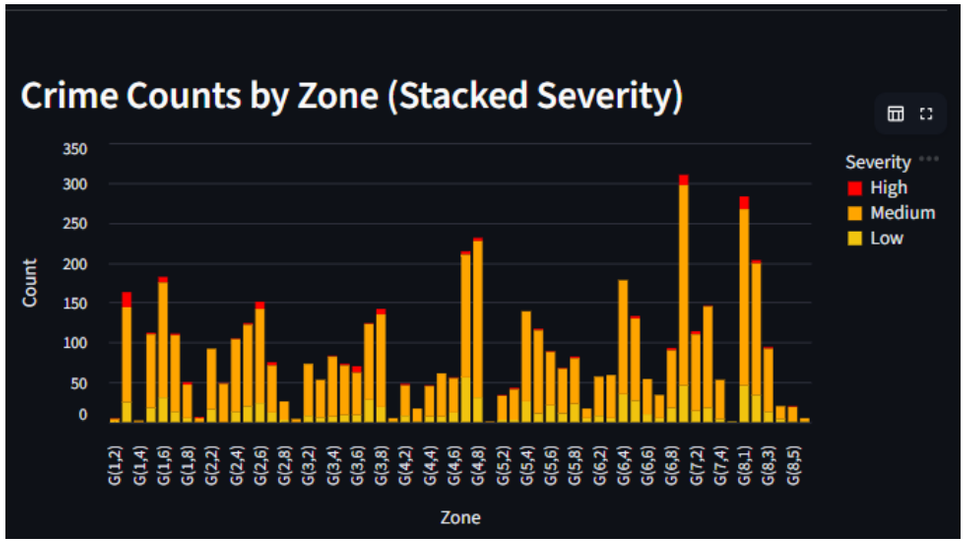

Fig. 3. Crime Counts by District, categorized by severity (Red: High, Orange: Medium, Yellow: Low)

B. Visualizations and Spatial Analysis

Geospatial mapping (Fig. 2). unveiled distinct clustering phenomena, confirming that crime distribution is non-uniform. The interactive heatmap highlighted that "High Severity" events tend to localize in specific quadrants, distinctly separate from the broader dispersion of "Low Severity" infractions.

Further quantitative decomposition (Fig. 3) reveals critical nuances in regional security profiles:

- Volume vs. Severity: The West zone exhibits the highest aggregate volume of incidents, approaching 1,900 recorded events, driven primarily by a substantial base of Low (Yellow) and Medium (Orange) severity offenses.

Risk Density: Conversely, while the South zone registers fewer total incidents than the West, it displays a disproportionately high ratio of High-severity (Red) crimes relative to its total volume.

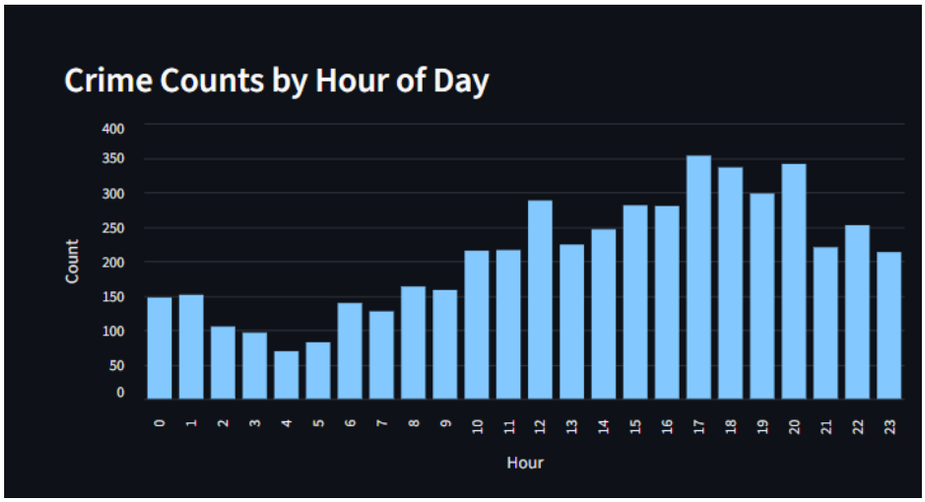

Fig. 4. Crime Counts by Hour of Day

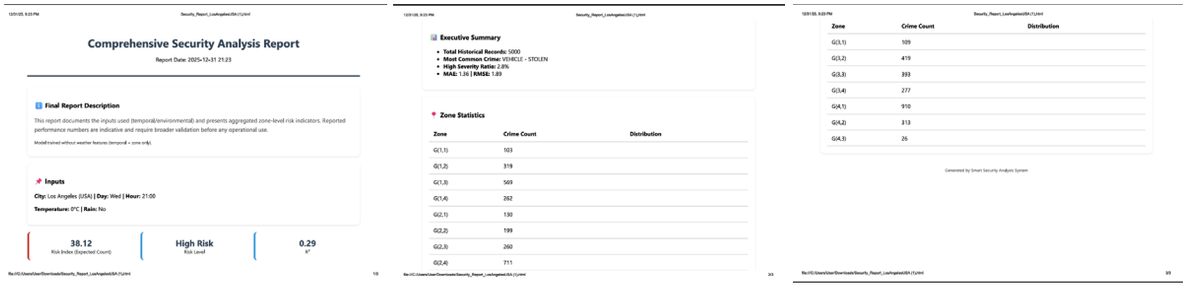

Fig. 5. Security Analysis Report

This qualitative difference is vital for strategic resource allocation.

Temporally, the "Crime Counts by Hour of Day" chart (Fig. 4) identifies clear temporal hotspots, with criminal activity escalating significantly from noon (12:00) and maintaining peak levels through the late evening (18:00), before dropping to its lowest points during the early morning hours.

C. Automated Comprehensive Reporting

To facilitate offline analysis and documentation, the system generates a downloadable " Security Analysis Report" in HTML format (Fig. 5). This document serves as an immutable record of the specific simulation session, capturing critical input variables such as the selected city, date, time, and simulated weather conditions. The report automatically embeds the AI-generated Risk Index and the Risk Level classification alongside the model's precision metric. Furthermore, it provides an "Executive Summary" and "Zone Statistics" that break down crime distribution numerically, ensuring that decision-makers have a portable and detailed reference for strategic planning without needing active access to the live dashboard.

V. Discussion

The study confirms that integration open-source government elemetry into a cohesive Machine Learning pipeline, can yield actionable security intelligence. Far from rendering traditional policing

obsolete, this system acts as a force multiplier, offering a structured, data-centric lens on risk distribution.

A critical finding of this research is the divergence between crime volume and crime severity. As observed in the results, high-frequency zones are not necessarily high-risk zones when severity is weighted. For instance, while one zone may register the highest total volume of

incidents, another may exhibit a significantly higher concentration of High-Severity events. This distinction is vital for operational resource allocation; a high volume of minor infractions (e.g., petty theft) requires a different administrative response than a lower volume of high-threat incidents (e.g., violent assaults).

By distinguishing between these categories both visually and predictively, the framework empowers command centers to prioritize tactical intervention in areas with a higher probability of severe crime, rather than simply reacting to aggregate call volumes. Although the spatial resolution in this prototype was simplified, the findings successfully illustrate how interpretable analytics can facilitate a paradigm shift from reactive policing to strategic, evidence-based proactive policing.

VI. Ethical Considerations, Limitations and Future Work

The deployment of predictive analytics necessitates a rigorous ethical framework to prevent the reinforcement of historical bias or the automation of discrimination. Consequently, this research was conducted at an aggregate, zone-level resolution specifically to mitigate risks associated with individual profiling. The system remains a prototype and is not currently deployed for active enforcement.

Methodological and Technical Limitations: Certain constraints were observed during development. First, the reliance on simulated historical weather data limits the precise correlation between environmental shifts and crime spikes. Second, the ingestion of massive datasets in real-time is currently subject to API throttling and latency, which can impact the immediacy of live predictions. Additionally, the spatial zoning (North, South, East, West) offers high-level insights but lacks the granularity required for street-level tactical positioning.

Future Work: The roadmap for future development focuses on three key pillars:

1. Algorithmic Enhancement: Transitioning from Random Forest to Long Short-Term Memory (LSTM) networks to better capture complex temporal dependencies in time-series data.
2. Data Enrichment: Integrating live, commercial-grade weather APIs to refine environmental correlation, alongside higher-resolution spatial models.
3. Operational Security: Establishing robust data governance protocols, including User Authentication and audit logging, to ensure the system is hardened for responsible official use.

VII. Conclusion

This research articulates a practical pathway for integrating machine learning into security operations. By successfully implementing a comprehensive pipeline spanning automated data ingestion, severity classification, and interactive visualization the project demonstrates that standard analytical techniques can be effectively operationalized for proactive planning.

The study's primary contribution lies not in algorithmic novelty, but in the transparent application of data science to solve real-world public safety challenges. By bridging the gap between raw telemetry and actionable intelligence, the system empowers analysts to make data-informed decisions instantly. Ultimately, the results underscore the transformative potential of accessible, interpretable decision-support systems in modernizing information security strategies

Funding

This article did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest

Authors declare that they have no conflict of interest.

References

[1] M. Feng et al., "Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data," in *IEEE Access*, vol. 7, pp. 106111-106123, 2019, doi: 10.1109/ACCESS.2019.2930410.

[2] M. -S. Baek, W. Park, J. Park, K. -H. Jang and Y. -T. Lee, "Smart Policing Technique With Crime Type and Risk Score Prediction Based on Machine Learning for Early Awareness of Risk Situation," in *IEEE Access*, vol. 9, pp. 131906-131915, 2021, doi: 10.1109/ACCESS.2021.3112682.

[3] Sharma S, Rai BK, Kumar G, Prajapati A, Kumar V. Crime visualization and forecasting using machine learning. In International conference on data science, machine learning and applications 2022 Dec 26 (pp. 307-320). Singapore: Springer Nature Singapore.

[4] Saravanan P, Selvaprabu J, Arun Raj L, Abdul Azeez Khan A, Javubar Sathick K. Survey on crime analysis and prediction using data mining and machine learning techniques. In Advances in Smart Grid Technology: Select Proceedings of PECCON 2019—Volume II 2020 Sep 19 (pp. 435-448). Singapore: Springer Singapore.

[5] Mandalapu V, Elluri L, Vyas P, Roy N. Crime prediction using machine learning and deep learning: A systematic review and future directions. Ieee Access. 2023 Jun 14;11:60153-70.

[6] Harsha MS, Akhil MH. Unravelling State-wise Crime Patterns with Ensemble Machine Learning and Data Visualization for Public Safety. In2024 2nd International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA) 2024 Mar 15 (pp. 1-5). IEEE.

[7] Deepa N, Pham QV, Nguyen DC, Bhattacharya S, Prabadevi B, Gadekallu TR, Maddikunta PK, Fang F, Pathirana PN. A survey on blockchain for big data: Approaches, opportunities, and future directions. Future Generation Computer Systems. 2022 Jun 1;131:209-26..

[8] Deepa N, Pham QV, Nguyen DC, Bhattacharya S, Prabadevi B, Gadekallu TR, Maddikunta PK, Fang F, Pathirana PN. A survey on blockchain for big data: Approaches, opportunities, and future directions. Future Generation Computer Systems. 2022 Jun 1;131:209-26.

[9] Socrata, "Socrata Open Data API (SODA)," Seattle, WA: Tyler Technologies, 2024. [Dataset]. Available: https://dev.socrata.com/. [Accessed: Nov. 28, 2025].

[10] Breiman L. Random forests. Machine learning. 2001 Oct;45(1):5-32.

[11] Adel A. Future of industry 5.0 in society: human-centric solutions, challenges and prospective research areas. Journal of Cloud Computing. 2022 Sep 8;11(1):40.