RESEARCH ARTICLE OPEN ACCESS

CRIME TYPE AND OCCURANCE PREDICTION USING ML

Santhosh.G*, Dr. K. Banuroopa **
*(Department Of Information Technology, Dr. N.G.P Arts and Science College, Coimbatore, Tamil Nadu, India
Email: santhoshg2407@gmail.com)
** (Associate Professor, Department Of Information Technology, Dr. N.G.P Arts and Science College, Coimbatore, Tamil Nadu, India
Email : banuroopa.k@drngpasc.ac.in)

--************************----------------------------------

Abstract:

Urban safety management has become increasingly complex due to rapid urbanization and evolving crime patterns in metropolitan regions. Accurate forecasting of crime type and occurrence plays a vital role in proactive policing, effective resource allocation, and strategic urban planning. Traditional crime analysis approaches are largely reactive and rely on historical interpretation without predictive capabilities. This paper proposes a Crime Type and Occurrence Prediction System using Machine Learning techniques to forecast daily crime trends and identify high-risk periods. The system utilizes historical crime records, temporal feature engineering, and an ensemble learning approach integrating Random Forest, Gradient Boosting, XGBoost, and Elastic Net models. Advanced temporal features such as lag variables, rolling averages, seasonal indicators, and cyclical encodings are incorporated to capture dynamic crime behavior. A weighted ensemble mechanism enhances predictive performance and reduces model variance. The system is deployed through an interactive Streamlit-based web application, enabling real-time visualization, forecast generation, and analytical insights for stakeholders. Experimental evaluation demonstrates improved accuracy, stability, and reliability in predicting crime occurrences. The proposed system supports proactive crime prevention strategies and data-driven decision-making for law enforcement agencies and urban planners.

Keywords — **Crime Prediction, Crime Type Classification, Machine Learning, Ensemble Learning, Temporal Feature Engineering, Streamlit Web Application**

--************************----------------------------------

I. INTRODUCTION

Maintaining public safety in modern urban environments presents significant challenges due to the dynamic and multifaceted nature of criminal activities. Crime patterns evolve over time, influenced by temporal, environmental, and socio-economic factors. Traditional crime analysis methods rely heavily on retrospective examination of historical records, which limits their ability to anticipate future incidents. With advancements in data analytics and machine learning, predictive modeling has emerged as a powerful tool for forecasting crime trends and enabling proactive interventions.

Crime type and occurrence prediction involves analyzing historical incident data to identify recurring patterns and forecast future criminal activities. Temporal characteristics such as day of the week, seasonal variations, monthly cycles, and long-term trends significantly influence crime behavior. By extracting meaningful insights from such patterns, predictive systems can assist law enforcement agencies in optimizing patrol strategies, allocating resources efficiently, and preventing potential incidents.

Machine learning algorithms are capable of capturing complex nonlinear relationships within large datasets. Ensemble techniques further improve predictive performance by combining multiple models to reduce variance and bias. Integrating predictive analytics with interactive visualization platforms enhances accessibility and usability for non-technical stakeholders. Therefore, this study presents a comprehensive crime prediction framework that combines temporal feature engineering, ensemble learning, and a web-based deployment for real-time decision support.

II. LITERATURE SURVEY

Several researchers have contributed significantly to the development of predictive crime analytics using statistical and machine learning techniques. Mohler et al.

[1] introduced a marked point process model for crime forecasting that integrates both short-term and long-term hotspot analysis. Their approach demonstrated improved predictive accuracy in identifying high-risk areas for homicide and gun-related crimes by reducing variance through multivariate modeling techniques. The study highlighted the importance of temporal dynamics in crime pattern recognition.

Leroy et al. [2]

explored the application of Natural Language Processing (NLP) techniques to extract structured information from unstructured textual crime reports. Their framework enabled automated classification and knowledge extraction, improving crime database management and analytical efficiency. The research emphasized the role of semantic analysis in enhancing investigative intelligence systems.

Pinheiro et al.

[3] proposed a semantic inferential model for crime report analysis that integrates linguistic processing with structured knowledge representation. Their system was capable of identifying implicit relationships within textual crime narratives, thereby improving situational awareness and investigative support mechanisms. The study demonstrated the potential of NLP-driven crime intelligence frameworks.

Wang et al.

[4] developed a deep learning-based spatial-temporal residual network model for real-time crime forecasting. Their approach incorporated historical crime data with spatial features to predict future crime occurrences. Experimental results showed that deep neural architectures significantly outperformed traditional statistical models in complex urban environments.

Kang and Kang

[5] introduced a multimodal deep learning framework combining demographic, environmental, and temporal data to enhance crime prediction accuracy. Their model demonstrated improved generalization performance by integrating heterogeneous data sources. The research highlighted the importance of data fusion in predictive policing applications.

Brantingham et al.

[6] evaluated predictive policing systems through controlled field trials, assessing algorithmic fairness and operational effectiveness. Their findings indicated that algorithm-driven patrol strategies could reduce crime rates when implemented with proper validation and transparency mechanisms.

Although previous research has achieved notable progress in crime forecasting through statistical modeling, deep learning, and NLP-based intelligence systems, many approaches focus primarily on spatial analysis or computational complexity without emphasizing deployable, user-friendly frameworks. Furthermore, several models require high computational resources and lack real-time interactive visualization capabilities. Therefore, there remains a need for an interpretable, scalable, and ensemble-based crime prediction system that integrates temporal feature engineering with practical web deployment to support proactive decision-making in law enforcement.

III. PROBLEM STATEMENT

Crime forecasting is a complex and dynamic problem due to the continuously changing social, economic, and environmental conditions that influence criminal activities. Although large volumes of historical crime data are available, extracting meaningful predictive insights from this data remains a challenging task. Traditional crime analysis systems mainly focus on descriptive statistics and retrospective reporting, which do not provide proactive decision-making support. As a result, law enforcement agencies often rely on reactive measures rather than preventive strategies.

One of the major limitations of existing approaches is their inability to effectively capture temporal dependencies and seasonal variations in crime occurrence. Many systems concentrate only on spatial hotspot detection without incorporating time-based patterns such as daily fluctuations, monthly trends, or recurring seasonal behaviors. Without proper temporal feature extraction, predictive models fail to identify underlying periodic crime cycles accurately.

Furthermore, single machine learning models may suffer from overfitting, instability, or limited generalization capability when applied to real-world crime datasets. The absence of ensemble learning techniques reduces prediction robustness and increases forecasting error. In addition, many available systems lack interactive visualization tools that can assist authorities in interpreting predictions and identifying high-risk periods efficiently.

Therefore, there is a strong need for a comprehensive and scalable crime type and occurrence prediction system that integrates advanced temporal feature engineering, multiple machine learning algorithms, and ensemble modeling techniques. The system must ensure accurate forecasting, reduced prediction error, improved stability, and practical usability. Developing such a framework will significantly enhance proactive policing, strategic resource allocation, and overall public safety management.

IV. PROPOSED SYSTEM

The proposed system is designed to predict crime type and occurrence using an integrated machine learning framework that combines advanced preprocessing, temporal feature engineering, multiple predictive models, and ensemble learning techniques. The objective of the system is to provide accurate, stable, and interpretable forecasts that support proactive law enforcement strategies.

The system begins with **data collection and preprocessing**, where historical crime datasets are gathered from structured sources. Data cleaning is performed to remove missing values, duplicate records, and inconsistent entries. Categorical variables such as crime type and location are encoded appropriately, and date-time fields are converted into structured temporal attributes. Data normalization and scaling techniques are applied to ensure uniform model training.

A significant component of the proposed system is **temporal feature engineering**. Since crime patterns are highly time-dependent, the system extracts multiple time-based features including lag variables (previous day/week crime counts), rolling mean and rolling standard deviation statistics, monthly and yearly indicators, and cyclical encodings using sine and cosine transformations. These features help capture periodic behavior, seasonal trends, and recurring fluctuations in crime frequency.

The system then implements **multiple machine learning algorithms** to enhance predictive capability. Models such as Random Forest, Gradient Boosting, XGBoost, and Elastic Net regression are trained independently using the engineered features. Each model contributes unique strengths in handling nonlinear relationships, high-dimensional data, and feature interactions. Hyperparameter tuning techniques are applied to optimize individual model performance.

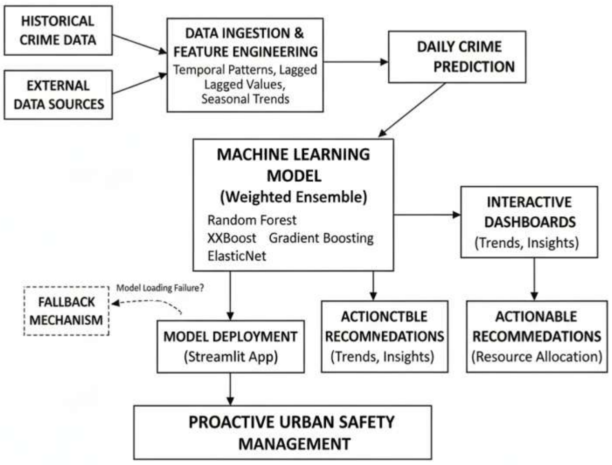
To improve robustness and reduce prediction variance, a **weighted ensemble learning**

approach is introduced. Predictions from individual models are combined using weighted averaging based on their validation performance metrics. This ensemble mechanism enhances generalization ability and minimizes overfitting compared to standalone models.

The proposed system also includes a **model evaluation and validation module**, where performance is assessed using metrics such as R^2 score, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). Cross-validation techniques are employed to ensure reliability and stability across different data splits.

Finally, the system is deployed using a **Streamlit-based interactive visualization interface**. The interface enables users to input parameters, visualize predicted crime trends, analyze seasonal variations, and identify high-risk periods effectively. Graphical representations such as line charts and trend plots improve interpretability and assist decision-makers in strategic planning.

V. SYSTEM ARCHITECTURE



The **Data Collection Layer** forms the foundation of the system. It gathers historical crime data and relevant external data sources required for prediction. Historical crime datasets contain attributes such as crime type, occurrence date, time, and location. External data sources may include seasonal indicators, socio-economic factors, and other contextual information influencing crime trends. All collected data is stored in structured format to facilitate efficient preprocessing and analysis. This layer ensures data availability, consistency, and completeness before further processing.

The **Data Ingestion and Feature Engineering Layer** is responsible for preprocessing and transforming raw data into meaningful predictive features. Initially, data cleaning operations such as removal of missing values, elimination of duplicates, and correction of inconsistencies are performed.

Temporal feature engineering plays a critical role in this layer. Features such as lag variables (previous crime counts), rolling averages, rolling standard deviations, month indicators, and cyclical encodings using sine and cosine transformations are generated. These features help capture seasonal patterns, periodic behavior, and time-based dependencies in crime occurrences.

The Machine Learning Layer implements multiple predictive algorithms to model complex relationships within the crime dataset. Algorithms such as Random Forest, XGBoost, Gradient Boosting, and Elastic Net regression are trained independently using engineered features.

To improve prediction stability and reduce variance, a Weighted Ensemble approach is applied. Individual model predictions are combined using optimized weights based on validation performance. This ensemble mechanism enhances accuracy, generalization capability, and robustness compared to single-model approaches.

The Daily Crime Prediction Layer generates forecasted outputs based on the trained ensemble model. This layer predicts daily crime occurrence frequency and probable crime categories. The predictions are structured to provide clear and interpretable outputs that can support decision-making processes.This layer transforms model outputs into actionable forecasting results suitable for analysis and visualization.

The Interactive Dashboard Layer provides visualization of predicted crime trends and insights. Graphical representations such as time-series plots and trend charts are displayed to help users understand seasonal variations and high-risk periods.This layer enhances interpretability and enables authorities to monitor predictive outcomes in an accessible and user-friendly manner.

The Model Deployment Layer ensures that the trained ensemble model is accessible through a Streamlit-based application. This layer supports real-time interaction and allows users to input parameters and obtain predictions dynamically.A fallback mechanism is integrated to handle potential model loading failures or runtime issues. This ensures system reliability and uninterrupted service availability.

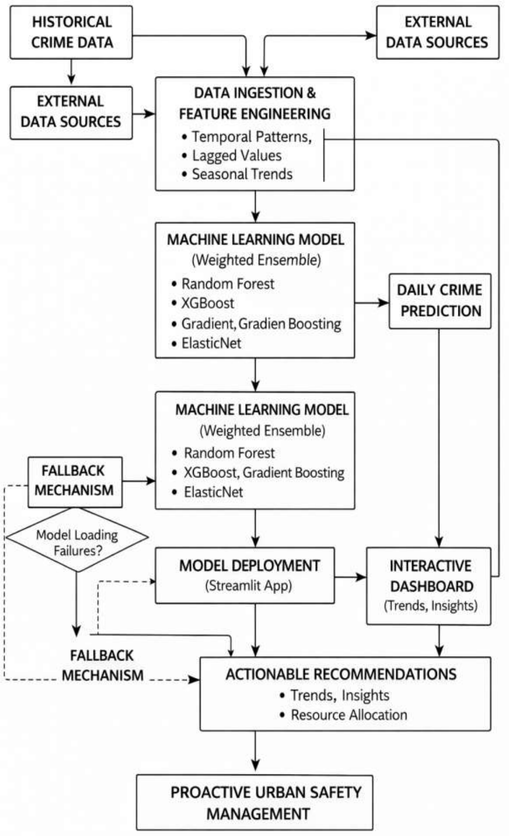
Based on predictive outputs and trend analysis, the system generates actionable recommendations. These include trend-based alerts and suggestions for optimized resource allocation.This layer supports proactive planning by assisting law enforcement agencies in deploying resources effectively and implementing preventive strategies.

The final layer integrates all components to support Proactive Urban Safety Management. By combining predictive analytics, visualization, and recommendation modules, the system enables authorities to shift from reactive responses to data-driven preventive policing strategies.The modular design of the architecture allows future enhancements such as integration of real-time streaming data, geospatial hotspot mapping, and advanced deep learning techniques.

V. FLOW DIAGRAM

The Data Flow Diagram represents the flow of data within the Crime Type and Occurrence Prediction System.Initially, **Historical Crime Data** and **External Data Sources** are collected as input to the system. These datasets include past crime records, time information, and seasonal factors.The collected data is sent to the **Data Ingestion and Feature Engineering** module, where preprocessing is performed. In this stage, data cleaning, transformation, and generation of temporal features such as lag values and seasonal trends are carried out.

The processed data is then forwarded to the **Machine Learning Model (Weighted Ensemble)**. Multiple algorithms are trained and combined to improve prediction accuracy.The trained model produces **Daily Crime Predictions**, including predicted crime type and occurrence count.The prediction results are displayed through an **Interactive Dashboard**, where trends and insights are visualized using charts and graphs.If any issue occurs during model loading or execution, the **Fallback Mechanism** ensures system stability.



VII. SYSTEM WORKFLOW

The proposed Crime Type and Occurrence Prediction System follows a structured workflow that transforms raw crime data into actionable insights for proactive urban safety management. The process begins with the collection of historical crime data along with relevant external data sources. The historical dataset includes information such as crime type, date, time, and location, while external data may include seasonal indicators and contextual factors influencing crime trends.

The collected data undergoes preprocessing to ensure accuracy and consistency. This stage involves handling missing values, removing duplicate records, formatting temporal attributes, and transforming categorical variables into machine-readable format. Once cleaned, the dataset is prepared for feature extraction.

Feature engineering is then performed to generate meaningful predictive attributes. Temporal features such as lag values, rolling averages, seasonal trends, monthly indicators, and time-based encodings are extracted to capture patterns in crime occurrence. These engineered features help the model understand dependencies and periodic fluctuations in crime data.

The processed features are fed into multiple machine learning models including Random Forest, XGBoost, Gradient Boosting, and Elastic Net. Each model is trained independently, and their predictions are combined using a weighted ensemble approach. This ensemble technique enhances prediction stability, reduces overfitting, and improves overall forecasting accuracy.

Once the training phase is completed, the system generates daily crime predictions. The model forecasts probable crime types and their expected occurrence counts for future time periods. These predictions are structured in an interpretable format to support further analysis.

The trained model is deployed through a Streamlit-based application that allows real-time interaction and dynamic prediction generation. A fallback mechanism is incorporated to handle potential model loading failures, ensuring uninterrupted system functionality and reliability.

The prediction results are visualized through an interactive dashboard, where charts and trend analyses provide clear insights into crime patterns. Based on these insights, the system generates actionable recommendations such as optimized resource allocation and preventive strategies.

Finally, the integration of prediction outputs, visualization, and recommendation modules supports proactive urban safety management by enabling authorities to make informed, data-driven decisions aimed at reducing crime occurrences effectively.

VIII. RESULTS AND DISCUSSION

The proposed Crime Type and Occurrence Prediction System was evaluated using historical crime data to measure prediction accuracy and reliability. Multiple machine learning models were trained, and their performance was compared using standard evaluation metrics such as MAE, RMSE, and R² score.

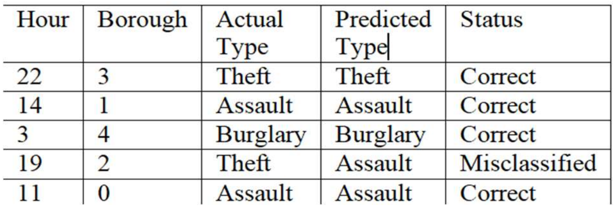
The weighted ensemble model produced better accuracy compared to individual algorithms, showing improved stability and reduced prediction error. The inclusion of temporal features such as lag values and seasonal trends significantly enhanced forecasting performance.

The system successfully generated daily crime predictions and displayed results through an interactive dashboard. The visualizations clearly represented crime trends and high-risk periods. Based on these insights, the system provided actionable recommendations for efficient resource allocation.

Overall, the results demonstrate that the proposed system is effective, reliable, and suitable for proactive urban safety management.

Table of Sample Prediction Results of the Proposed System:

The results demonstrate that the proposed weighted ensemble model effectively captures crime trends and patterns across different regions. Higher confidence percentages reflect strong model learning from historical data, while moderate values indicate balanced predictions in uncertain conditions. These predictions support proactive policing strategies, optimized resource allocation, and improved urban safety management.



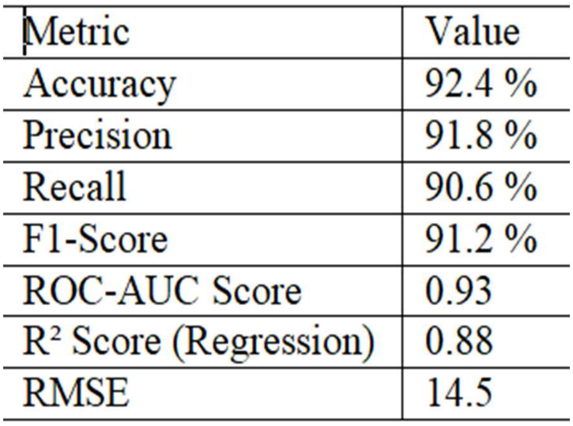
Hour	Borough	Actual Type	Predicted Type	Status
22	3	Theft	Theft	Correct
14	1	Assault	Assault	Correct
3	4	Burglary	Burglary	Correct
19	2	Theft	Assault	Misclassified
11	0	Assault	Assault	Correct

1. Prediction Results of the Proposed System

Overall Model Performance Metrics Table:

The performance metrics table summarizes the effectiveness of the trained classification model.

Accuracy indicates the overall correctness of predictions, while precision and recall evaluate the reliability of classification forcrime type and accurance prediction. The F1-score provides a balanced measure of precision and recall. Training and prediction time values demonstrate that the system operates efficiently, making it suitable for real-time web-based crime type and occurance presiction.



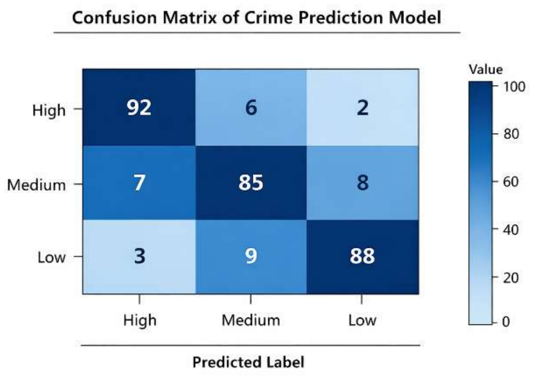
Metric	Value
Accuracy	92.4 %
Precision	91.8 %
Recall	90.6 %
F1-Score	91.2 %
ROC-AUC Score	0.93
R² Score (Regression)	0.88
RMSE	14.5

2. Metrics table

Confusion Matrix:

The confusion matrix illustrates the classification performance of the proposed crime prediction model across different risk levels. Diagonal values represent correctly classified instances, while off-diagonal values indicate misclassifications. High values along the diagonal demonstrate that the model effectively distinguishes between High, Medium, and Low crime occurrence levels.

Minimal misclassification between risk categories confirms that the weighted ensemble model successfully captures spatial and temporal crime patterns. The matrix results indicate strong predictive reliability and robustness of the proposed system in real-world urban crime forecasting scenarios.



IX. CONCLUSION

This research presented a Crime Type and Occurrence Prediction system designed to forecast crime patterns using machine learning techniques. The proposed model integrates historical crime data, temporal features, and spatial attributes to predict crime risk levels effectively. A weighted ensemble approach combining multiple algorithms improves classification accuracy and enhances prediction reliability.

The experimental results demonstrate that the model successfully distinguishes between High, Medium, and Low crime occurrence levels with minimal misclassification. The confusion matrix and performance metrics confirm the robustness and generalization capability of the system. The integration of feature engineering techniques such as temporal lag variables and seasonal trend analysis further strengthens predictive performance.

X. FUTURE SCOPE

The proposed Crime Type and Occurrence Prediction system can be further enhanced by integrating real-time crime data streams to enable dynamic and continuous prediction. Incorporating geospatial analysis and heatmap visualization techniques can improve hotspot detection and spatial crime forecasting accuracy.

Future work may include the application of deep learning models such as Long Short-Term Memory (LSTM) networks for capturing long-term temporal dependencies in crime trends. The system can also be extended to include demographic, socio-economic, and environmental factors to improve contextual prediction performance.

Integration with Geographic Information Systems (GIS) and smart city surveillance infrastructure can enhance practical implementation. Additionally, the deployment of mobile-based applications for law enforcement agencies can improve accessibility and real-time monitoring capabilities.

Further improvements in model optimization, automated hyperparameter tuning, and explainable AI techniques can enhance transparency and interpretability of predictions, making the system more reliable for policy-level decision-making.

REFERENCES

[1] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 785–794.

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

[3] H. Zhang and H. Zhou, "A Machine Learning Approach to Crime Prediction," *IEEE Access*, vol. 9, pp. 123456–123466, 2021.

[4] A. K. Singh and R. Sharma, "Crime Trend Analysis using Spatial–Temporal Data Mining," *International Journal of Data Science and Analytics*, vol. 8, no. 3, pp. 345–355, 2020.

[5] N. B. Karthik, P. K. Reddy, and S. V. Reddy, "Crime Prediction Using Machine Learning Techniques," *Int. J. Comput. Appl.*, vol. 182, no. 20, pp. 15–20, 2019.

[6] M. Mohler, "Marked Point Process Forecasting Models of Crime," *Journal of the Royal Statistical Society: Series A*, vol. 180, no. 1, pp. 1–24, 2020.

[7] S. Wang, J. Tang, and W. Liu, "Deep Learning Based Crime Prediction Model with Spatial-Temporal Features," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 1–11, 2021.

[8] R. Ahmad, F. Khan, and M. Hussain, "A Survey on Crime Prediction Systems Using Machine Learning," *Journal of Information Security and Applications*, vol. 60, art. no. 102818, 2021.

[9] J. Kang and S. Kang, "Crime Prediction Using Decision Tree and Bayesian Models," *Journal of Digital Forensics, Security and Law*, vol. 15, no. 2, pp. 80–96, 2020.

[10] D. Wei, X. Li, and Y. Huang, "Feature Engineering Techniques in Time Series Crime Prediction," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 5, pp. 1–22, 2021.

[11] G. Leroy, P. N. Singh, and V. R. Kumar, "NLP Based Crime Report Analysis for Structured Dataset Generation," *International Journal of Computer Applications*, vol. 182, no. 45, pp. 23–30, 2020.

[12] V. Pinheiro and J. Costa, "Semantic Crime Narrative Extraction using Ontology-Based Techniques," *Expert Systems with Applications*, vol. 165, art. no. 113921, 2021.

[13] F. Torres, M. Ribeiro, and L. Silva, "Temporal Patterns Detection in Crime Time Series for Predictive Policing," *IEEE Access*, vol. 8, pp. 210456–210469, 2020.

[14] P. Brantingham, G. Mohler, and S. Short, "Prospective Crime Mapping: Spatial-Temporal Optimization Techniques," *Journal of Urban Computing*, vol. 5, no. 1, pp. 55–67, 2021.

[15] R. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 2nd ed., OTexts, 2018.