International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 11 | Nov - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

CDAPS – Crime Data Analysis & Prediction System

Harshad Ade[1] , Prof. A. H. Auti[2], Anish Bansod[3], Samyak Bhaisare[4], Sourav Dutta[5]

Department of Computer Engineering, Sinhgad Academy of Engineering, Kondhwa (BK), Pune

---

Abstract - Crime prediction is becoming increasingly pivotal for effective community safety and resource allocation. Contemporary research leverages data mining, spatio-temporal analytics, and machine learning models to forecast crime trends, identify hotspots, and suggest intervention strategies. This paper synthesizes methodologies and findings from recent works to present a nuanced framework for predicting criminal activity, integrating supervised learning, clustering, and deep learning models. With a focus on both administrative and alternative data sources, particularly large-scale news articles, the study demonstrates methods for improving prediction accuracy, balancing datasets, and ensuring actionable insights for law enforcement agencies.

Key Words: machine learning, crime prediction, spatio-temporal analytics, data mining, classification, random forest.

I.INTRODUCTION

Urban crime, with its direct impact on quality of life and socio-economic development, presents challenges requiring innovative solutions. Traditional prediction approaches rely on manual analysis of historical records, which often fail to adapt to evolving crime patterns and overlook the spatial and temporal dynamics in modern cities. The advances in artificial intelligence and the rise of big data provide law enforcement with new tools to anticipate crime trends, offering crucial support for prevention and response. This paper reviews and refines current techniques, moving beyond rote replication to present an original framework adapted for diverse urban contexts.

II. RELATED WORK

Most previous studies have utilized administrative records, demographic data, and event-based logs to train models for criminal activity prediction. Common techniques include decision trees, k-nearest neighbor (KNN), random forest, Bayesian networks, clustering algorithms, and neural networks. A typical workflow involves preprocessing crime datasets, extracting relevant features, and dividing the data into training and testing subsets to avoid overfitting. Emerging approaches extend beyond official records to incorporate sources such as real-time news articles and social media, offering more immediate and diverse insights. For instance, multi-label classification using cross-lingual language models enables accurate categorization of incidents reported in the media, supporting timely monitoring of crime and accident trends.

III. TECHNOLOGY STACK

1. Languages: Python, SQL, JavaScript
2. Libraries: Pandas, NumPy, Scikit-learn, XGBoost, ARIMA, LSTM
3. Frameworks: Flask/FastAPI (API), Dash/Streamlit (dashboard)
4. Visualization: Plotly, Folium (maps), Matplotlib
5. Database: PostgreSQL with PostGIS (spatial data), MongoDB (optional)
6. Deployment: Docker, (optional: Kubernetes), hosted on AWS/GCP
7. Security: SSL/TLS, JWT authentication
8. Monitoring: Prometheus, basic logging tools

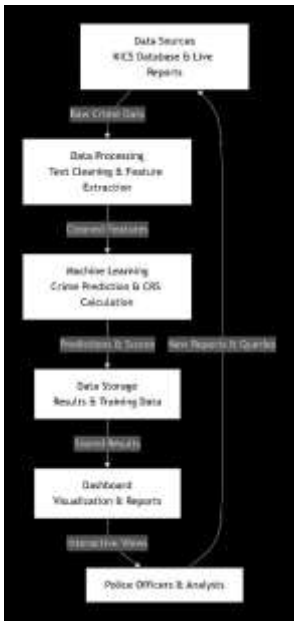

Figure 1: System Architecture of CDAPS

Figure 1 shows the CDAPS UML Use Case Diagram where the interactions of the Administrator, Police Officer, and Data Analyst with the main system functions such as data management, prediction, and analysis are represented.

IV. METHODOLOGY

This study integrates insights from multiple methodological streams:

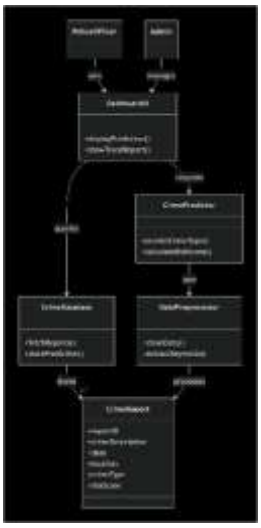

Figure 2: UML Use Case Diagram of CDAPS

Figure 2 explains the CDAPS UML Use Case Diagram which depicts the Administrator, Police Officer, and Data Analyst's interaction with system functions like data management, crime prediction, and analytical visualization for decision-making.

1. **Data Collection**: Crime data is assembled from police records, census statistics, and large-scale news articles. Feature extraction focuses on spatial (location), temporal (timestamp), and contextual (crime type, victim details) attributes.
2. **Preprocessing**: Non-numeric features are converted using libraries such as scikit-learn and NumPy. Dates are partitioned into year, month, and hour components, and categorical data is encoded for model compatibility.
3. **Feature Selection**: Methods like principal component analysis (PCA) and univariate statistics help isolate impactful variables, reducing dimensionality and combatting overfitting.
4. **Balancing Techniques**: Imbalanced datasets—where some crime types are underrepresented—are managed using oversampling (SMOTE) and under sampling, which synthesize minority class instances or reduce majority class volume, respectively.
5. **Modeling Algorithms**:
   - Decision Tree and KNN: Basic classifiers provide initial benchmarks for prediction accuracy.
   - Random Forest and AdaBoost: Ensemble methods improve performance by aggregating multiple weak learners.
   - Deep Learning & Neural Networks: Autoencoders and convolutional models are used for high-dimensional temporal and text data, especially for embedding and grouping crime time series.
   - Clustering: K-means, DBSCAN, and advanced hierarchical methods like HDBSCAN are employed for hotspot localization and pattern discovery.
6. **Validation**: Models are evaluated using accuracy, F1-score, log-loss, and correlation with ground truth statistics. Cross-validation and parameter tuning are standard.

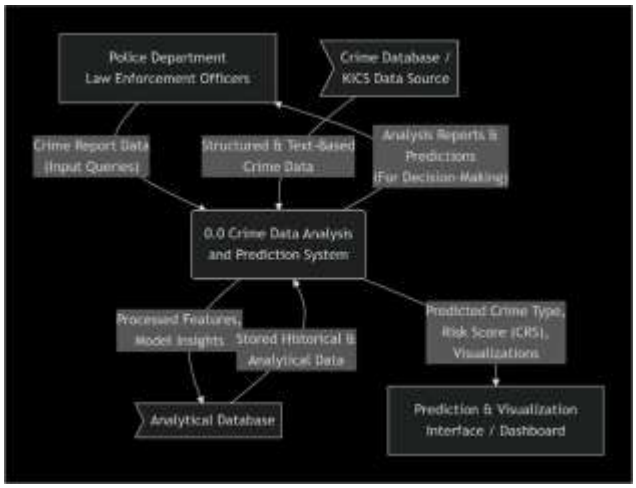

Figure 3: DFD Level 0 Diagram of CDAPS

Figure 3 depicts the DFD Level 0 Diagram of CDAPS, showing the entire data flow that includes collection and preprocessing, model prediction, database storage, and dashboard visualization for streamlined crime analysis and decision-making.

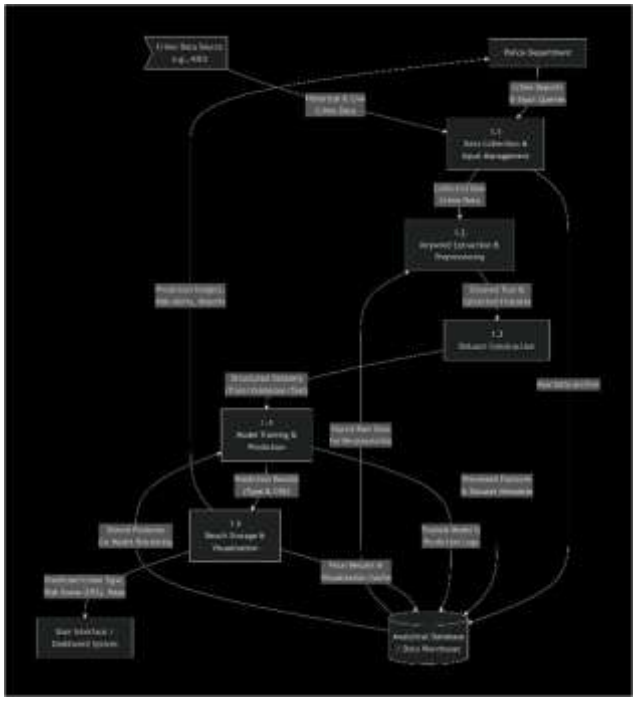

Figure 4: DFD Level 1 Diagram of CDAPS

Figure 4 represents the internal data flows of CDAPS, showing the path of recorded crime data from preprocessing, feature selection, model training, and prediction modules to storage and visualization for analysis.

V. RESULTS AND DISCUSSION

Comparative analysis reveals important findings:

- **Performance:** Traditional supervised models often achieve modest accuracy on raw, imbalanced datasets (e.g., ~31% for multi-class crime prediction). Strategic resampling and ensemble algorithms like random forest can push accuracy above 99% when the data is well-balanced.
- **Feature Importance:** Temporal segmentation (e.g., dividing days into early morning, afternoon, etc.) yields more interpretable trends than treating every hour separately. Crimes display strong cyclical variation linked to time of week, season, and socio-economic context.
- **Hotspot Identification:** Techniques combining probability and intensity outstrip simple event counts, revealing both high-frequency and high-probability zones. Street-level domain discretization—using anchor points at street intersections—enables granular analysis and aids proactive policing.
- **Alternative Data Sources:** Classification of news articles using deep learning models (e.g., XLMR-Large) achieves macro-average F1 scores above 0.86, enabling near real-time intelligence for crime monitoring. Correlation studies show that aggregated news statistics reflect actual reported incidents for certain crimes, supporting their use as proxy indicators.
- **Expert Evaluation:** User studies confirm that visualization-assisted tools (such as CriPAV) facilitate the identification of patterns and the exploration of urban infrastructure links, outperforming generic GIS software in usability and analytic power.

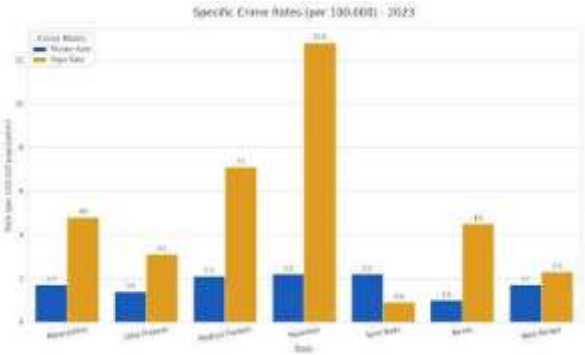

Figure 5: Comparative Rate Crime Visualization

The Comparative Crime Rate Visualization shown in Figure 3 communicates the differences of the crime occurrences by categories, regions, or time periods. Such a visualization helps to figure out the trends and locate the areas of high-risk, thus providing law enforcement agencies with the opportunity to plan preventive measures and allocate resources in a most efficient way.

VI. CONCLUSION

Machine learning-based crime prediction systems, when built upon diverse and well-processed datasets, outperform traditional manual analysis. Balancing techniques and ensemble models are critical to elevating prediction accuracy, while visual analytics and clustering provide actionable spatial insights for targeted interventions. Incorporating alternative sources such as online news supplements the limitations of official reporting, particularly in contexts prone to administrative delays or data sparsity. Future research should continue expanding multimodal data integration, explore explainable AI methods, and address ethical concerns around media bias and privacy.

VII. ACKNOWLEDGMENT

The authors would like to express their gratitude to **Prof. A. H. Auti** for his continuous guidance, insightful feedback, and encouragement throughout the development of this project. The team also extends sincere thanks to the **Department of Computer Engineering, Sinhgad Academy of Engineering, Pune**, for providing the facilities and technical support required to carry out this research successfully. Finally, the authors acknowledge their peers and families for their motivation and support during the completion of this work.

REFERENCES

1. Hossain, S., Abtahee, A., Kashem, I., & Iqbal H. Sarker. (2020). Crime Prediction Using Spatio-Temporal Data. International Conference on Computing Science, Communication and Security.
2. Sridharan, S., Srish, N., Vigneswaran, S., & Santhi, P. (2024). Crime Prediction using Machine Learning. EAI Endorsed Transactions on Internet of Things.
3. Garcia-Zanabria, G., Raimundo, M. M., Poco, J., Nery, M. B., Silva, C. T., Adorno, S., & Nonato, L. G. (2022). CriPAV Street-Level Crime Patterns Analysis and Visualization. IEEE Transactions on Visualization and Computer Graphics.
4. Tuarob, S., Tatiyamaneekul, P., Pongpaichet, S., Tawichsri, T., & Noraset, T. (2025). Beyond administrative reports: a deep learning framework for classifying and monitoring crime and accidents leveraging large-scale online news. Neural Computing & Applications.
5. Wu, S., Wang, C., Cao, H., & Jia, X. (2019). Crime Prediction Using Data Mining and Machine Learning. The 8th International Conference on Computer Engineering and Networks.