Hybrid Machine Learning Approach for Crime Prediction and Forecasting: Integrating K-Means Clustering and LSTM Networks

KARTHIKEYAN M[1], NAVEEN KUMAR N[2], MARUTHUPANDI M[3]

[1, 2, 3]*Sri Krishna College of Engineering and Technology*

Abstract- ***For law enforcement organizations to maintain public safety and effectively deploy resources, crime prediction and forecasting are essential duties. In order to produce precise crime forecasts, this study offers a novel strategy that blends machine learning and deep learning methodologies. In order to take use of their complementing strengths in recognizing crime patterns and modeling sequential data, we suggest combining k-means clustering with Long Short-Term Memory (LSTM) networks. The first step in the procedure is gathering and preparing various crime data, such as demographic data and crime reports. The chronological, spatial, and social components of the extracted variables offer important insights into criminal incidents. Similar crime data points are organized into clusters using k-means clustering, which indicates unique crime patterns or hotspots. After these clusters are converted into sequential data, LSTM models can use historical crime sequences to forecast future crime patterns. In terms of crime prediction and forecasting, the hybrid approach shows encouraging results, giving law enforcement useful information to stop and efficiently address criminal activity. However, while using sensitive crime-related data, ethical issues and data protection laws should be strictly followed.***

Indexed Terms- ***Crime Prediction, Machine Learning, Analysis and Forecast, Metropolitan Safety***

I. INTRODUCTION

An essential component of maintaining security and allocating resources as efficiently as possible is crime analysis. Better strategic planning and decision-making are made possible by spotting patterns and trends in crime data. Numerous analytical methods can be used to glean valuable insights from a variety of datasets, such as statistics on crime rates, demographic characteristics, and environmental circumstances. It is feasible to find hidden links and anticipate unexpected incidents by employing sophisticated computer techniques. Assessing risk variables and creating efficient preventive actions are made easier by putting organized approaches to data processing and pattern recognition into practice. Maintaining ethical standards and making sure data is used responsibly are still crucial factors in this analytical process.

1.1 CRIME PREDICTION

The practice of predicting where and when specific sorts of crimes are likely to occur using data and statistical analysis is known as crime prediction. This idea is part of the larger field of predictive policing, which seeks to maximize law enforcement resources and tactics by utilizing data and technology.

This is the general procedure for crime prediction:
1. Data Collection: Law enforcement organizations collect a variety of data, such as socioeconomic characteristics, historical crime data, weather patterns, location, time of day, and other pertinent information.
2. Data Analysis: To find trends and connections between various factors and crime episodes, statistical and machine learning methods are used to evaluate the gathered data. These algorithms are able to identify patterns and connections that human analysis could miss.
3. Predictive Models: Predictive models are constructed using the data analysis. These models forecast future crime events based on historical data. Neural networks, decision trees, and random forests

are popular machine learning methods for this. 4. Deployment and Monitoring: By integrating the predictive models with law enforcement systems, police are able to focus patrols and resources on regions and times where there is a greater chance of crime. These models' efficacy is regularly assessed, and modifications are made as needed to increase their accuracy.

1.2 MACHINE LEARNING

Computational models known as machine learning algorithms are made to allow machines to learn from data and make judgments or predictions without needing to be explicitly programmed for every task. These algorithms have led to major advances in artificial intelligence and are at the heart of many machine learning applications.

Machine learning algorithms come in a variety of forms, each with a distinct function depending on the type of data and the issue they are intended to address. The following are some typical types of machine learning algorithms:

1. Supervised Learning: In this method, an algorithm is trained using a labeled dataset in which every data point is linked to an outcome variable or target. In order for the algorithm to be able to anticipate new, unknown data, it must learn the mapping between the input features and the appropriate target variable. The following are some instances of supervised learning algorithms: For regression issues with a continuous target variable, linear regression is utilized. b. When the target variable contains two classes, logistic regression is utilized to solve binary classification problems. c. Support Vector Machines (SVM): Especially useful in high-dimensional domains, SVM can be used for both regression and classification tasks. d. Decision Trees: A hierarchical framework for classification and regression applications. e. Random Forest: An ensemble technique that lessens overfitting and increases accuracy by combining several decision trees.

1.3 ANALYSIS AND FORECAST

Two crucial components of planning and decision-making in a variety of domains, such as business, economics, weather forecasting, and more, are analysis and forecasting. Let's examine each of these ideas in detail:1. Analysis: Analyzing data is looking at and interpreting it to find trends, patterns, connections, and other pertinent information. Gaining a greater comprehension of the data and extracting useful information from it are the objectives of analysis. Depending on the goals and the type of data, there are various kinds of analysis:
• Data analysis: This entails examining and purifying unprocessed data to find trends, abnormalities, and significant characteristics. Statistical approaches, data visualization strategies, and exploratory data analysis (EDA) are frequently used in data analysis. Analyzing financial data to determine the health and performance of an investment or business is known as financial analysis. Trend analysis, cash flow analysis, and ratio analysis are common techniques in financial analysis.
• Sentiment analysis: In natural language processing, this kind of analysis is used to identify the sentiment or feelings conveyed in textual data, including customer reviews or posts on social media.

II. LITERATURE REVIEW

Mohle, George et al. Crime hotspot maps, as suggested in this study, are a popular and effective way to allocate police resources and show spatial crime patterns. However, only one crime type is frequently used to construct hotspot maps across a single timeline. Particularly for low frequency crimes like homicide, risk estimations suffer from a huge variance when using short-term hotspot maps that use several weeks of crime data. Near-repeat effects and new hotspot trends are not considered in long-term hotspot maps that use data spanning several years. In this study, we demonstrate how a marked point process technique can be used to extend point process models of crime to incorporate leading indicator crime types, while capturing both short-term and long-term patterns of risk. Accurate hotspot maps that can be utilized for predictive policing of gun-related crime are produced by methodically combining years' worth of data and a wide variety of crime types. We use the methodology on a sizable open source data collection that the Chicago Police Department has made publicly accessible online. The methodology for predicting homicide and precursory gun crimes is developed in this research and applied to predictive policing [1]. This study makes a proposal by Goudy A. Leroy et al. It must be possible to report crimes around-the-clock. There are a number of alternative reporting options, from in-person reporting to internet submissions, even though 911 and tip-lines are the most well-known. Crime victims and witnesses can report occurrences to police at any time, from any location, using internet-based crime reporting systems. However, witnesses' memory recall is not well supported by the current email and text-based systems, which results in reports that are less accurate and contain less information. Additionally, these solutions do not make it easier to integrate and reuse the supplied data with other information systems. In order to gather pertinent crime information from witness accounts and ask follow-up questions based on that information, we are creating an anonymous online crime reporting system. We support memory recall by using investigative interviews and natural language processing techniques, and we facilitate knowledge reuse by mapping the data straight to a database. We present the assessment of the system's Suspect Description Module (SDM). 70% (recall) of the information from witness accounts is accurately captured by our interface [2].

Pinheiro Vládia et al. This article outlines the architecture for web-based information extraction systems that are based on natural language processing (NLP) and specifically designed for the investigation of crime-related data. The NLP module, which is based on the Semantic Inferential Model, is the architecture's primary feature. By using the design to contribute to WikiCrimes, a collaborative web-based crime registration system, we show that the architecture is feasible. Textual narratives are one of the primary information sources that managers and analysts of public safety employ. These reports include details about crimes that have been committed as well as profiles of individuals that intelligence services keep an eye on. In order to generate knowledge and take practical steps to enhance public safety, these reports highlight the traits, quirks, and connections between the events and individuals. They also enable one to identify patterns. However, it takes a lot of effort and time to read the large amount of information in natural language. As a result, a Natural Language Processing (NLP) system that aids in the comprehension of these documents is extremely beneficial. Understanding meaning is limited to explicit textual input in the majority of NLP techniques now in use [3].

According to Bao Wang et al., real-time crime predictions is crucial in this approach. It is challenging to make an exact prediction about the time and location of the next crime, though. Such a complicated system has no known physical model that can reasonably approximate it. There is a weak signal of interest and a lack of historical crime data in both space and time. We begin this task by providing an accurate depiction of crime data. In order to forecast the distribution of crime in Los Angeles at the hourly scale in neighbourhood-sized parcels, we next modify the spatial temporal residual network on the well-represented data. These tests, together with comparisons to a number of other prediction methods already in use, show how accurate the suggested model is. In order to solve the resource consumption problem for its practical implementation, we lastly introduce a ternarization technique. It is a significant scientific and practical challenge to forecast crime in micro-geographic regions at hourly or even finer temporal scales. Finding out where and when crimes are most likely to happen opens up new avenues for crime prevention. However, it is very difficult to forecast crime accurately at fine spatial temporal scales. Crime is influenced by a variety of intricate elements, many of which are impossible to quantify. Crime is statistically sparse and highly chaotic in both space and time. 21 Mathematical and statistical modeling of crime has been the focus of recent work [4].

According to Sharmila Chackravarthy et al., this approach makes it clear that protecting any home depends on the prompt and precise identification of criminal activities. Crime detection system integration aims to increase security in light of smart cities' explosive growth. To accomplish this, traditional video surveillance has always been heavily relied upon. This frequently results in a backlog of video data that needs to be watched over by a supervisor. Error rates rise in large urban areas as a result of the supervisory officers' ever heavier workload. Workload reduction strategies have been put into place. Although they have a number of drawbacks, auto regressive models are currently utilized to more accurately predict criminal activity. We suggest a method for analyzing video stream data that combines neural networks with a Hybrid Deep Learning algorithm. The workload for the supervisory officials will be lessened as a result of our system's ability to swiftly recognize and evaluate illicit activities. An effective and flexible criminal detection system will be possible if it is integrated into the infrastructure of smart cities. It is now challenging to police and keep an eye on places with a high crime likelihood due to the recent population boom in urban areas. Insecurity and criminal activity have increased in certain regions as a result of this lack of control. There is a chance to develop innovative solutions to these issues as smart city infrastructure develops [5].

III. EXISTING SYSTEM

Crime and violations are intended to be managed because they pose a threat to justice. Computationally, accurate crime prediction and trends for the future can help improve urban safety. Early and precise crime prediction and forecasting are hampered by humans' limited capacity to process complicated information from huge data. Numerous computational opportunities and challenges arise from the precise calculation of the crime rate, types, and hot locations based on historical patterns. Even with extensive research, improved predictive algorithms are still required to guide police patrols toward criminal activity. Prior research on learning models for crime forecasting and prediction accuracy is weak. In order to better fit the crime data, this study used a variety of machine learning algorithms, including logistic regression, support vector machine (SVM), Naïve Bayes, k-nearest neighbors (KNN), decision trees, multilayer perceptrons (MLP), random forests, and extreme gradient boosting (XGBoost). Additionally, time series analysis was done using the autoregressive integrated moving average (ARIMA) model and long-short term memory (LSTM). Additionally, the key regions for both cities were used to further identify the outcomes of the crime predictions. All things considered, these findings are helpful in guiding police tactics and practice since they offer early detection of crime, hotspots with greater crime rates, and future trends with better predictive accuracy than with other approaches.

IV. PROPOSED SYSTEM

By utilizing machine learning and deep learning techniques, the proposed system seeks to create a sophisticated framework for crime prediction and forecasting. To improve crime analysis and prediction skills, the system will combine k-means clustering with Long Short-Term Memory (LSTM) neural networks. Through k-means clustering, the system will use past crime data to find hotspots and underlying patterns of criminal activity, giving law enforcement authorities important information about how crimes are distributed. LSTM, which is excellent at modeling sequential dependencies in time series data, will then be used to record the temporal features of crime incidents. By combining clustering with LSTM, the system will be able to predict probable criminal situations, giving law enforcement proactive capabilities to deploy resources effectively and carry out focused preventive actions. Additionally, the system will provide interpretable information and visuals to help decision-makers comprehend crime trends and create successful crime reduction plans. Communities will become safer and more secure as a result of society adopting a more proactive and data-driven approach to combating crime under the proposed system.

4.1 LOAD DATA

The module starts by loading historical crime data, including crime kinds, timestamps, locations, and any other pertinent characteristics, from a variety of sources. To supplement the crime data, additional information is gathered, such as socioeconomic and environmental aspects.

4.2 DATA PREPROCESSING

The loaded data is preprocessed at this point to guarantee its quality and analysis-suitability. The module resolves any discrepancies in the data, eliminates outliers, and handles missing numbers. Numerical features are normalized to a common scale, and categorical variables are encoded into numerical representations. For the purpose of evaluating the model, the dataset is then divided into training and testing sets.

4.3 FEATURE SELECTION

In order to concentrate on the most instructive and pertinent characteristics for crime prediction, feature selection is essential. The module creates clusters of related crime incidents based on geographic proximity by using k-means clustering to find spatial patterns in the crime data. As extra features, the k-means cluster labels are included. To capture time-dependent patterns, temporal variables are retrieved, including season, time of day, and day of the week. To enhance the dataset, contextual information such as local locations, social activities, and weather are also taken into account.

4.4 TRAINING AND TESTING

The LSTM model is trained using the preprocessed and feature-selected data. LSTM may learn temporal connections in crime occurrences and is well-suited for sequential data. To maximize performance, the model is trained on the training set and verified on the testing set.

4.5 EVALUATION AND PERFORMANCE

The module uses suitable evaluation metrics, including accuracy, precision, recall, F1-score, and AUC-ROC, to assess the performance of the trained LSTM model. By contrasting the model's performance with and without the additional cluster labels, the contribution of k-means clustering to the LSTM model is evaluated. Performance analysis directs possible enhancements and aids in comprehending the model's advantages and disadvantages.

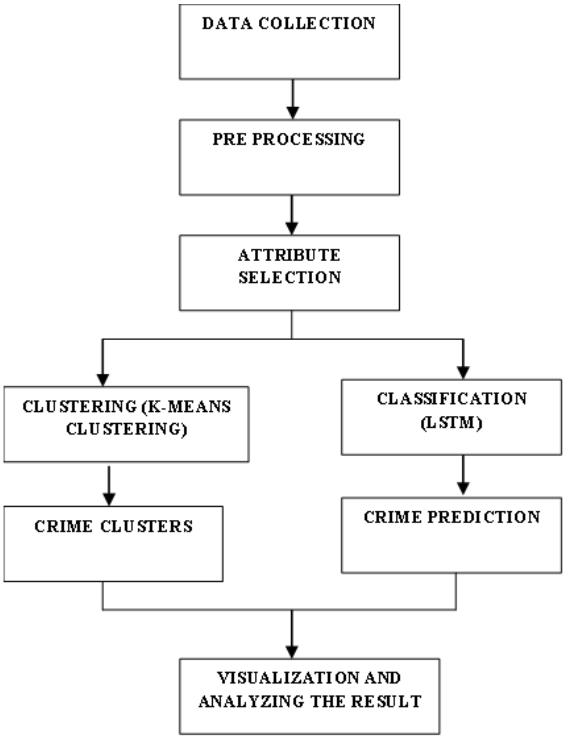

FIGURE 1. SYSTEM FLOW DIAGRAM

V. RESULT ANALYSIS

The examination of the findings is centered on assessing how well the suggested method works to spot trends and forecast future events using historical data. To evaluate the model's predictive power, a number of metrics are used, including accuracy, precision, recall, F1-score, and AUC-ROC. The incorporation of clustering algorithms improves feature representation by strengthening the capacity to group related patterns. Models with and without clustering-based features are compared to show how important spatial and temporal characteristics are for improving predictions. The results show that using structured data processing techniques produces more dependable results, highlighting how crucial it is to choose pertinent characteristics and adjust model parameters for optimal performance.

algorithm	accuracy
EXISITNG SYSTEM	77
PROPOSED SYSTEM	83

Table 1. Comparison table

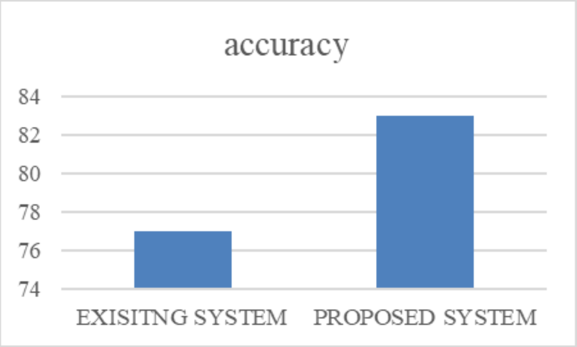

Figure 2. Comparison graph

CONCLUSION

To sum up, the suggested crime prediction and forecasting system is a major step forward in using deep learning and machine learning methods to improve law enforcement and public safety tactics. The system efficiently detects unique crime patterns and models temporal connections by fusing k-means clustering with LSTM networks, producing precise and timely crime predictions. Law enforcement organizations are able to proactively deploy resources and react quickly to new security risks because to the real-time capabilities. The thorough data analysis and intuitive interface offer insightful information that may be used to prevent and respond to crimes. All things considered, the crime prediction and forecasting system has a lot of potential to make communities safer by facilitating proactive and data-driven law enforcement tactics. The system can change and adapt as it keeps learning from fresh data, which could eventually increase its capacity for prediction. Its worth in assisting with efforts to prevent and respond to crime will surely be recognized when it enters operational deployment, creating safer and more secure communities.

FUTURE WORK

To improve the suggested approach even more, future studies and advancements in the field of crime prediction and forecasting can concentrate on a few crucial areas. First off, adding more sophisticated deep learning architectures, like Transformer-based models or Graph Neural Networks, may enhance the system's capacity to identify intricate linkages and patterns in crime data. Second, adding data from outside sources, such social media or urban sensors, may enhance the predictive models by offering more contextual details. Furthermore, investigating the application of ensemble techniques that integrate forecasts from several models may result in more reliable and accurate crime forecasting. Furthermore, examining the effects of adding current socioeconomic variables and law enforcement activities to the system may provide important new information about the dynamics of crime. Last but not least, extending the system's reach or combining it with other public safety initiatives may result in a more complete and interwoven ecosystem for crime prevention. These upcoming initiatives could revolutionize crime forecasting and prediction, leading to safer and more secure societies in the long run.

REFERENCES

[1] G. Mohler, "Homicide and gun crime prediction in Chicago using marked point process hotspot maps," Int. J. Forecasting, July 2022, vol. 30, no. 3, pp. 491–497.

[2] 'Natural language processing and e-government: Extracting reusable crime report information,' in Proc. IEEE Int. Conf. Inf. Reuse Integr., Las Vegas, IL, USA, Aug. 2020, pp. 221–226. A. Iriberri and G. Leroy.

[3] In Proc. IEEE Int. Conf. Intell. Secur. Informat., Vancouver, BC, Canada, May 2021, pp. 19–24, V. Pinheiro, V. Furtado, T. Pequeno, and D. Nogueira, "Natural language processing based on semantic inferentialism for extracting crime information from text."

[4] P. J. Brantingham, S. J. Osher, J. Xin, B. Wang, P. Yin, A. L. Bertozzi, and "Deep learning for real-time crime forecasting and its ternarization," Chin. Ann. Math., B, vol. 40, no. 6, pp. 949–966, Nov. 2020.

[5] S. Chackravarthy, S. Schmitt, and L. Yang, "Deep learning-based intelligent crime anomaly detection in smart cities," in Proceedings of IEEE 4th International Conference on Collaboration Internet Comput. (CIC), Philadelphia, PA, USA, October 2021, pp. 399–404.

[6] "Prediction of crime occurrence from multimodal data using deep learning," by H.-W. and H.-B. Kang Art. no. e0176244, PLoS ONE, vol. 12, no. 4, April 2019.

[7] As stated in Social Media Strategy in Policing (Security Informatics and Law Enforcement), B. Akhgar, P. S. Bayeri, and G. Leventakis, Eds. Cham, Switzerland: Springer, 2019, pp. 177–195, A. Fidow, M. Hassan, M. Imran, X. Cheng, C. Petridis, and C. Sule, "Proposing a hybrid approach mobile apps with big data analysis to report and prevent crimes."

[8] Brantingham, M. Valasik, and G. O. Mohler, P. J. "Does arrest bias result from predictive policing? A randomized controlled trial's findings, Statist. Public Policy, vol. 5, no. 1, pp. 1–6, January 2020.

[9] Adv. Sci. Technol. Lett., vol. 90, pp. 90–92, Dec. 2021; A. Nasridinov and Y.-H. Park, "A study on performance evaluation of machine learning algorithms for crime dataset,"

[10] "Forecasting crime with deep learning," by A. Stec and D. Klabjan arXiv:1806.01486, 2022. [Online]. http://arxiv.org/abs/1806.01486 is accessible.