Scientific Publications

Journal of Artificial Intelligence and Big Data, 2021, 1, 1334
www.scipublications.com/journal/index.php/jaibd
DOI: 10.31586/jaibd.2021.1334

Article

An Analysis of Crime Prediction and Classification Using Data Mining Techniques

Anuj Kumar Gupta [1,*], Dheeraj Varun Kumar Reddy Buddula [2], Hari Hara Sudheer Patchipulusu [3], Achuthananda Reddy Polu [4], Bhumeka Narra [5], Navya Vattikonda [6]

[1] Oracle ERP Senior Business Analyst, Genesis Alkali, USA
[2] Software Engineer, Anthem Inc, USA
[3] Software Engineer, Iheartmedia, USA
[4] SDE3, Goldman Sachs, USA
[5] Sr Java Developer, Statefarm, USA
[6] Business Intelligence Engineer, International Medical Group Inc, USA

*Correspondence: Anuj Kumar Gupta

Abstract: Crime is a serious and widespread problem in their society, thus preventing it is essential. Assignment. A significant number of crimes are committed every day. One tool for dealing with model crime is data mining. Crimes are costly to society in many ways, and they are also a major source of frustration for its members. A major area of machine learning research is crime detection. This paper analyzes crime prediction and classification using data mining techniques on a crime dataset spanning 2006 to 2016. This approach begins with cleaning and extracting features from raw data for data preparation. Then, machine learning and deep learning models, including RNN-LSTM, ARIMA, and Linear Regression, are applied. The performance of these models is evaluated using metrics like Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). The RNN-LSTM model achieved the lowest RMSE of 18.42, demonstrating superior predictive accuracy among the evaluated models. Data visualization techniques further unveiled crime patterns, offering actionable insights to prevent crime.

Keywords: Crime Prediction, Crime Data, Data Mining Visualization, Machine Learning, Deep Learning

How to cite this paper:
Gupta, A. K., Reddy Buddula, D. V. K., Patchipulusu, H. H. S., Polu, A. R., Narra, B., & Vattikonda, N. (2021). An Analysis of Crime Prediction and Classification Using Data Mining Techniques. *Journal of Artificial Intelligence and Big Data, 1*(1), 156–166.
DOI: 10.31586/jaibd.2021.1334

Received: August 22, 2021
Revised: November 26, 2021
Accepted: December 23, 2021
Published: December 27, 2021

1. Introduction

Crime poses a significant threat to society, and addressing it effectively has become increasingly complex due to its non-systematic and non-random nature. Modern technologies have not only advanced crime-solving methods but have also empowered criminals to carry out sophisticated offenses. According to the Crime Records Bureau, while some crimes, such as burglary and arson, have decreased, others, including murder, sexual abuse, and gang rape, have shown a significant rise [1]. Understanding the probability of crime in specific hotspot locations is crucial for devising effective preventive measures.

Crimes occur at various scales, from small villages to major urban centers, and they encompass a wide range of offenses such as murder, kidnapping, robbery, rape, assault, and more. Rising crime rates increase the urgency for law enforcement to address and resolve cases efficiently [2]. Predictive policing, which utilizes analytical and predictive techniques to identify potential crimes, has proven effective. However, as the crime rate increases and criminals become more technologically advanced, manual analysis of crime data stored in large warehouses becomes impractical. This necessitates the adoption of

advanced technologies, incorporating machine learning and data mining, to quickly and precisely examine, evaluate, and act upon this data.

Crime prediction security systems combine cutting-edge technology with AI, big data analytics, and predictive modeling to predict criminal activity and enhance law enforcement capabilities. The field of scientists focused on predictive policing system enhancement in 2019 by resolving outcome bias and privacy violation issues [3]. AI crime forecasting systems need open visibility and responsible handling to stop discriminatory behavior in operations. Crime prediction systems require effective development solutions to overcome two essential challenges data security protection and citizen rights protection.

Data mining is a diverse and growing field of study that helps reveal hidden patterns and insightful information in data [4]. The technique is vital because it helps discover new knowledge and clarify existing occurrences. Data mining procedures help organizations detect crime patterns within specific locations alongside time-related relationship patterns. Through mathematical approaches along with machine learning applied to time-based criminal data communities achieve better results in crafting prevention and control strategies.

1.1. Significance and Contribution of the Paper

Evaluating crime prediction through data mining methods is fundamental in enhancing public security while backing police activities. The research produces vital knowledge about crime activities by using machine learning methods with background crime information to make decisions based on data-driven principles. The findings highlight the efficacy of predictive models such as Decision Trees in providing accurate forecasts and facilitating timely and proactive interventions. Moreover, integrating data visualization techniques supports a deeper understanding of spatial and temporal crime dynamics, improving resource allocation and preventive measures. The contributions of study are:

- Develop effective ML and DL models for crime prediction by crime dataset.
- Demonstration of advanced data preprocessing techniques, feature extraction, and visualization to reveal spatial and temporal crime patterns and their correlations.
- Comparative evaluation of traditional machine learning models (e.g. ARIMA, Linear Regression) and a deep learning model (RNN-LSTM) for crime prediction accuracy.
- Recommendations for integrating real-time datasets and exploring ensemble and hybrid models for enhanced accuracy in future studies.

1.2. Structure of the paper

The study is organized as follows: Relevant research on data mining-based crime prediction is presented in Section II. The methods and supplies employed are described in full in Section III. The experimental results of the suggested system are shown in this section. Section V wraps up the inquiry and presents a summary of its results.

2. Literature Review

This section reviews and analyses surveys on data mining for crime prediction. Some of them are mentioned in this section.

Almaw and Kadam (2018), The trial found that Random Tree's accuracy was greater, at 82.0227%. The second category combines ensemble learning models with base classification models. The 3-ensemble model, which employed three distinct base classifiers, and the 1-ensemble model, which employed the same kind of classifier with a range of training sets, were the two forms into which the ensemble learning was divided. Based on the experiment's findings, the 1-ensemble model's accuracy is higher at 81.6073%,

while the 3-ensemble model's accuracy is higher at 79.2353%. The third category, statistical data visualization, examines how crime rates vary by season, time of day, and month[5].

Feng et.al. (2019) According to prediction findings, Neural network models are outperformed by the Prophet model and Keras stateful LSTM, and the optimal training data size is three years. These positive findings will aid police departments and law enforcement organizations better understand crime issues and provide information that will enable them to monitor activity, predict the likelihood of incidents, allocate resources effectively, and make more informed judgments [4].

Putri and Kurniadi (2019) Predictions of criminal activity are done through multiple data sources which include both spatiotemporal crime data and zoning district information. GBMs were trained as classifiers to conduct the tests on a subset of characteristics. The best result was obtained when all features were used, including KDE with smoothing and zoning district characteristics; on the validation set, the multiclass logarithmic loss was 2.356104, and on the test set, it was 2.35443 [6].

Kim et.al. (2018) looks at crime prediction based on machine learning. In this study, crime statistics for Vancouver for the last 15 years are examined using two different data-processing approaches. The accuracy of crime prediction using machine learning predictive models, such as K-nearest-neighbor and boosted decision trees, varies between 39% and 44% [7].

Almaw and Kadam et al. (2018) A model group consisting of Random Tree, J48 and Naive Bayes serves as the platforms for analysis. The accuracy of Random Tree was 82.0227%, according to the experiment data. Ensemble learning systems and base learning models are combined in the second categorization category. Two distinct approaches were employed for ensemble learning: the 3-ensemble model utilized three basic classifiers, while the 1-ensemble model used the same classifiers but trained them using different data sets. According to experimental data, the accuracy of an analysis utilizing the 3-ensemble model is 81.6073%, whereas an analysis using the 1-ensemble model is 79.2353%. The third field of study examines the statistical relationship between crime rates and daily time intervals, as well as the corresponding monthly and seasonal trends [5].

Sivaranjani, Sivakumar and Aasha et.al. (2017) The K-Means clustering process is displayed through Google Maps for better human interaction and comprehension. The KNN classification works as the chosen method for predicting criminal activities. The performance of different clustering algorithms is evaluated by accuracy, recall and F-measure measurement and these results exhibit comparison [8].

The literature review of crime prediction and classification through data mining covers the information summarized in Table 1.

Table 1. Summary of Crime Prediction and Classification Using Data Mining

Authors	Methods	Dataset	Key Findings	Limitations and Future Work
Almaw & Kadam (2018)	Random Tree, 1-ensemble, 3-ensemble, Statistical Analysis	Crime Dataset	Random Tree: 82.02% accuracy	More ensemble techniques needed and extend crime trend analysis.
Feng et.al.	Stateful LSTM with Keras and Prophet Model	Crime data (3 years training)	Compared to conventional neural network models, the Prophet model and Keras LSTM produced superior prediction results, which helped law enforcement allocate resources.	Further optimization of training dataset sizes and exploration of hybrid deep learning methods.

Crimes prediction using spatiotemporal data and kernel density estimation et.al.	Gradient Boosting Machine (GBM)	Spatiotemporal and zoning datasets	KDE with zoning district characteristics and smoothing improves model performance; achieved a multiclass logarithmic loss of 2.356104 on validation and 2.35443 on test sets.	Expand to real-time prediction applications and evaluate generalizability across various cities and regions.
Kim et.al.	Enhanced Decision Tree with K-Nearest Neighbour	Vancouver crime data (15 years)	The prediction accuracy of KNN and Boosted Decision Tree models varied between 39% and 44%.	Improve accuracy through advanced preprocessing, feature engineering, and incorporating contextual external data.
Almaw and Kadam et.al.	Naive Bayes, J48, and Random Tree	Experimented dataset	Random Tree outperformed others with 82.0227% accuracy. Ensemble models showed 81.6073% (1-ensemble) and 79.2353% (3-ensemble).	Limited focus on computational efficiency and need to explore ensemble models with novel classifiers.
Sivaranjani, Sivakumari and Aasha et.al.	K-Means and K-Nearest Neighbor (KNN)	Crime data visualized on Google Maps	K-Means clustering visualized with Google Maps enhances usability; KNN used for prediction and evaluated using precision, recall, and F-measure.	Need to refine spatial accuracy and investigate more advanced algorithms for geospatial clustering and prediction.

3. Methodology

The main objective of this research employs data mining methods to both forecast criminal activities and examine crime pattern changes. The research uses classification models together with statistical analysis to predict crime occurrence while identifying the main elements that shape criminal activity rates. The methodology studies. The crime classification and prediction process use an organized method with data mining methods. The figure below represents the machine learning model-based flowchart used for crime prediction. The collected crime dataset originates from public sources before undergoing pre-processing operations where missing values get eliminated, and features receive enhancements for improved model execution. After processing, the data is divided into subsets for testing (30%) and training (70%). For crime prediction, various machine learning models are used, including ARIMA, Linear Regression, and Recurrent Neural Networks (RNN-LSTM). Model performance is evaluated using metrics like to assess accuracy, use MAPE and RMSE. The most effective model is then selected to predict crime patterns based on historical data, aiding law enforcement in proactive crime prevention.

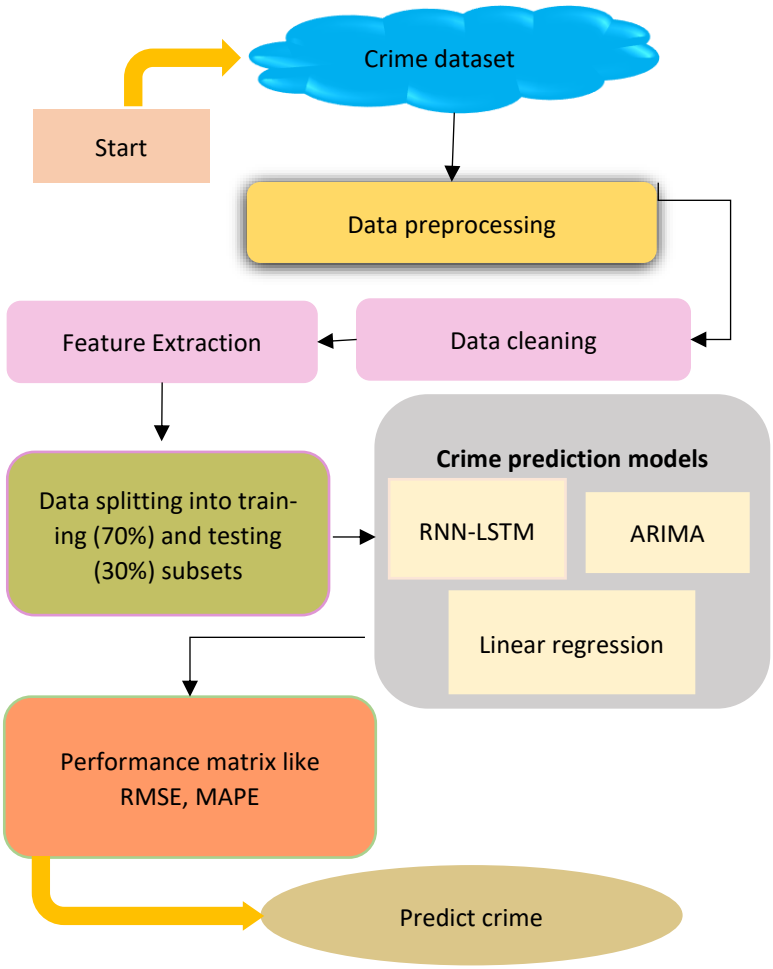

Figure 1. System flowchart for crime prediction using ML & DL models.

The successive steps of a flowchart are briefly detailed in the following paragraphs:

3.1. Data collection

The API retrieves the dataset from Philadelphia Crime's official website. The dataset includes statistics on several types of crimes from 2006 to 2016. The crime scene location and timing data are used to predict short- and medium-term crimes.

3.2. Data preprocessing

This process includes techniques to remove null or infinite values that might jeopardize the accuracy of the system. The key steps are, cleaning data and feature extraction. The cleaning procedure aims to eliminate or correct incomplete or missing data. Common practice dictates that datasets undergo pre-processing before classification algorithms are used.

- **Data cleaning:** To eliminate inaccurate numbers, data cleaning is used. The most crucial and difficult aspect of achieving great precision. Features with above 60% missing data are eliminated since they are useless for more research.
- **Feature Extraction:** The technique of identifying the pertinent and necessary data by classifying all of the data into particular categories is known as feature extraction [9]. Getting all the required information or minimizing the loss of pertinent data while dealing with an enormous dataset is critical.

3.3. Data splitting

The training and testing subsets of the crime dataset were separated, with 70% going towards training and 30% towards testing.

3.4. Classification with RNN-LSTM model

One potent kind of recurrent neural network (RNN) that can recognize long-term dependencies is the LSTM model. Because LSTMs can preserve the state while identifying patterns throughout the time series, they are especially helpful in prediction when time series feature auto-correlation, or the presence of correlation between the time series and lagged copies of itself [4]. The states can be maintained or exchanged between updated weights as each epoch goes on thanks to the recurrent design. The LSTM cell layout can also improve the RNN by permitting short-term and long-term persistence. The mathematical representation are (1 to 4) as:

$$f(t) = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \quad (1)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (4)$$

where ft is a sigmoid function that indicates if the prior condition should be maintained, The old cell state is denoted by C_{t-1}, the updated cell state by C_t, the prior values in each layer by W_f, W_i, and W_C, the input values by h_{t-1} and x_t, and the constant values by b_f, b_i, and b_C, which determine which value will be selected to update the state. C_t represents the new candidate's values.

4. Result Analysis and Discussion

There are several subsections within this section. First, provide the analysis of the crime data using EDA. then evaluated using a performance matrix. Finally, presents a comparison of the ML and DL concepts. The following experiments are conducted on a Dell desktop with 16 GB of RAM, Intel i3, Python 3.7 is used for programming.

4.1. Data analysis and visualization

Data visualization is both a science and an art. It's a visual communication method. It entails the development and analysis of data visualization. Using statistical visuals and charts to efficiently and clearly explain data is the main objective of data visualization. Effective visualization aids in data and evidence analysis and reasoning. In order to assist crime analysts in examining crime trends, the work generates maps of crime density. In order to investigate and prevent crimes, law enforcement and intelligence organizations must have a thorough understanding of patterns of criminal activity. The Crime data visualization graphics are provided below:

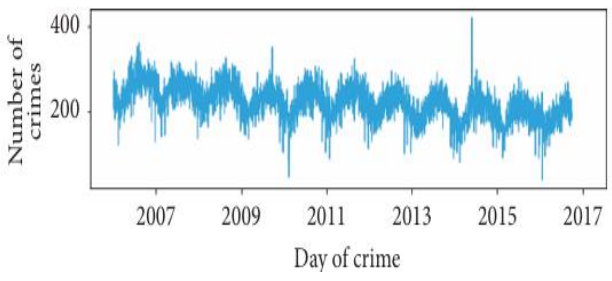

Figure 2. Daily crimes on crime data

This Figure 2 illustrates a time series of daily crime counts. The x-axis displays the date, while the y-axis displays the total number of infractions reported each day. The dataset spans an extended period, possibly over several years, and demonstrates significant fluctuations, characterized by peaks and troughs. These patterns suggest the presence of seasonality or recurring trends in crime occurrences.

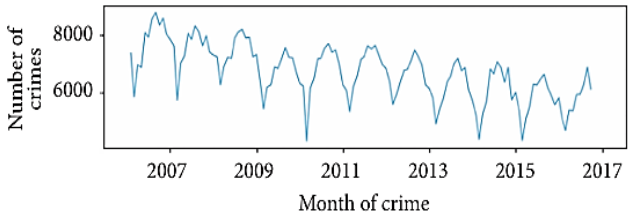

Figure 3. Monthly crimes

Figure 3 presents a line graph depicting the monthly crime count from 2007 to 2017. The x-axis represents the 'Month of Crime,' with specific months omitted and only years indicated. The y-axis denotes the 'Number of Crimes,' ranging approximately from 6,000 to 8,000. The graph reveals a cyclical pattern in crime rates, characterized by distinct peaks and troughs, suggesting potential seasonal or periodic influences. Additionally, the data exhibits a fluctuating trend with a noticeable overall decline in crimes towards the latter part of the observed period.

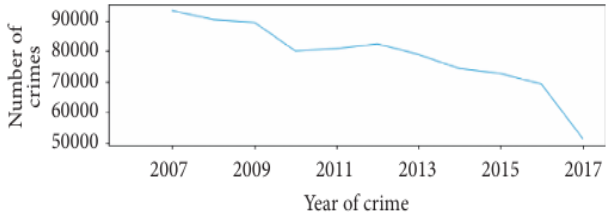

Figure 4. Number of crimes by Year

Figure 4 illustrates the line graph illustrates a downward trend in the number of reported crimes from 2007 to 2017. A notable decline is observed between 2007 and 2009, followed by relative stabilization lasting until 2013. After 2013, the crime rate declined gradually, with a steeper drop recorded between 2015 and 2017. This decrease may be attributed to several factors, including improved crime prevention measures, shifts in law enforcement strategies, or enhancements in socioeconomic conditions.

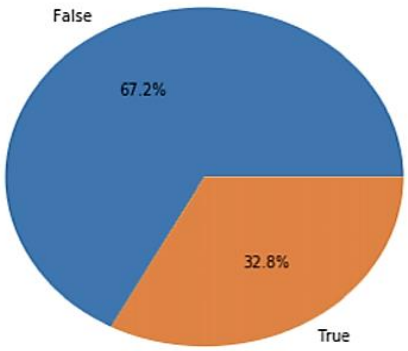

Figure 5. Percentage of crimes vs No crimes occurred

Figure 5 shows the pie chart provides a categorical breakdown of incidents, distinguishing between those that resulted in a crime (32.8%) and those that did not (67.2%). The data demonstrates that a significantly larger proportion of incidents did not lead to criminal activity, highlighting the disparity between reported incidents and those classified as crimes.

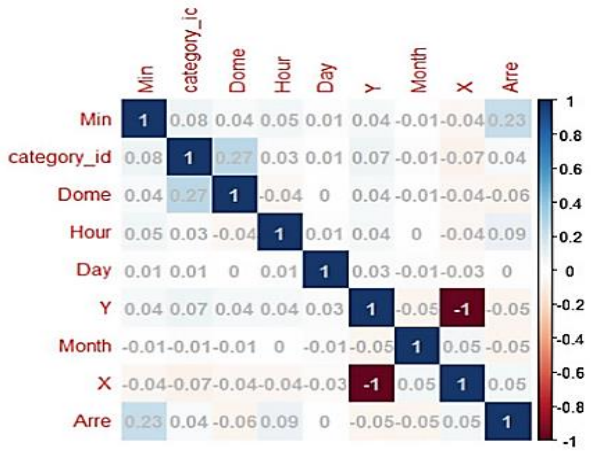

Figure 6. The correlation coefficient diagram

The correlation coefficient matrix in Figure 6 illustrates the dataset's relationships among various features (Min, category_id, Dome, Hour, Day, Y, Month, X, Arre). Strong negative correlations are observed between X and Month (-1) and a strong positive correlation between X and Arre (1), indicating inverse and direct relationships, respectively. Moderate correlations include a positive association between category_id and Dome (0.27) and a negative association between Hour and Arre (-0.09). Most other feature pairs show weak correlations, indicating limited linear relationships. This analysis provides insights into the dataset's structure, facilitating feature selection and guiding model development decisions.

4.2. Performance Measures

After algorithm installation, two metrics are produced that assess the algorithms' efficacy and efficiency: the correlation coefficient, MAPE, and RMSE. The determined measures will serve as inputs to analyze the comparative patterns between crime data. This study investigates the capability of algorithms to spot patterns in criminal activity through its main focus. The performance metrics require the following equations for their computation:

- **Root Mean Squared Error:** The square root of the average square of the total error is known as the RMSE. The root mean squared error belongs to the assessment tools for numerical forecasting accuracy evaluation. The root mean squared error obtains its definition through the formula which includes xpred as predicted values and xobs as observed values. It is given equation (5):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{pred} - x_{obs})^2} \quad (5)$$

- **MAPE:** The average absolute % difference between expected and actual data is the MAPE. It is given equation (6):

$$MAPE = \sum \frac{|A-F|}{A} \times 100 \quad (6)$$

4.3. Experimental results

This section presents experiment results from using RNN-LSTM with the crime dataset. The performance of ML classifiers, including RMSE and MAPE shown in Table 2.

Table 2. Model performance on crime dataset

Model	RMSE	MAPE
LSTM	18.42	6.03

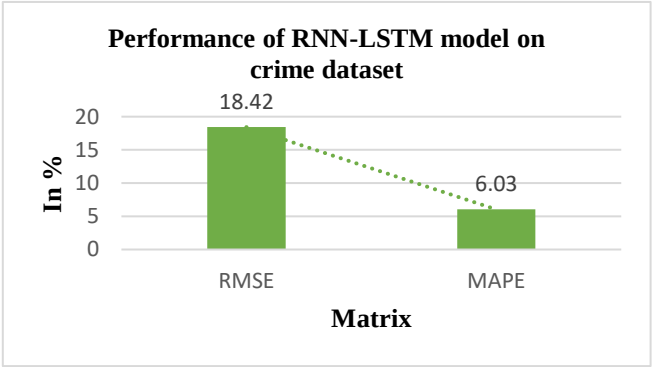

Figure 7. Performance of RNN-LSTM Model

The evaluation metrics MAPE and RMSE are used to show the performance of the LSTM model on the crime dataset in Table II and Figure 7. The mean percentage difference between the projected and actual values, or MAPE, was 6.03%, and the mean magnitude of error in the predicted values, or RMSE, for the LSTM model was 18.42.

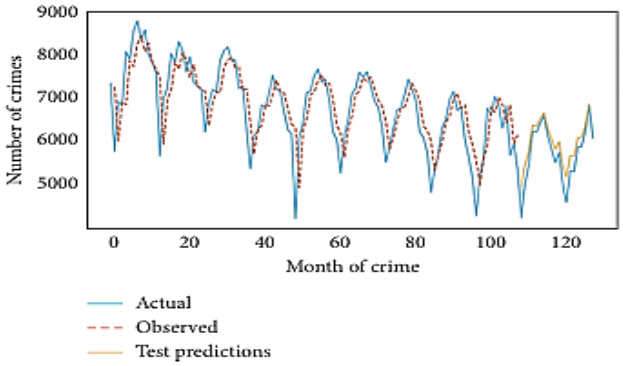

Figure 8. Actual and observed values for monthly crimes for RNN-LSTM

Figure 8 illustrates a time series analysis of crime data, plotting the "Number of Crimes" against the "Month of Crime." It features three distinct lines: the actual crime figures (blue line), the observed crime figures (13red line), and the test predictions (yellow line). The graph spans from approximately month 0 to month 120, highlighting a fluctuating trend in crime incidents over time. The observed and predicted values align closely with the actual trend, particularly in the later months, demonstrating the accuracy of the predictive model based on the RNN-LSTM approach.

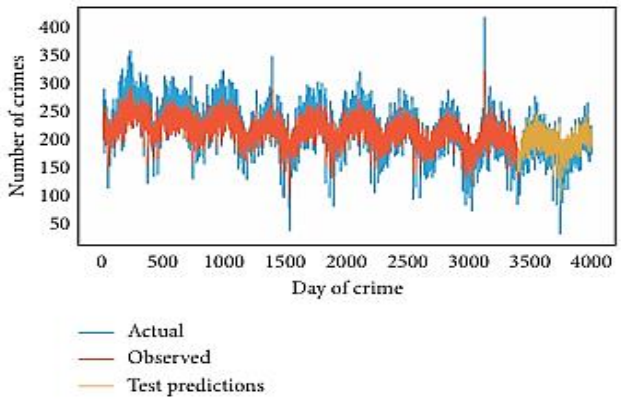

Figure 9. Actual and observed values form daily crimes of LSTM-RNN model

Figure 9: The graph, which plots the "Number of Crimes" against the "Month of Crime," shows a time series study of crime statistics." It features three distinct lines: the actual crime figures (blue line), the observed crime figures (red line), and the test predictions (yellow line). The graph spans from approximately month 0 to month 120, highlighting a fluctuating trend in crime incidents over time. The observed and predicted values align closely with the actual trend, particularly in the later months, demonstrating the accuracy of the predictive model based on the RNN-LSTM approach.

4.4. Comparative analysis

The comparative analysis for crime prediction on the crime dataset are provided in this section. The comparison of different ML and DL models RNN-LSTM, Linear regression [10], ARIMA [11], based on performance matrices like RMSE are given below:

Table 3. ML and DL models comparison on Crime dataset

Model	RMSE
RNN-LSTM	18.42
Linear regression [10]	23.3
ARIMA [11]	56.94

In Table 3, the comparison of models based on RMSE values reveals notable differences in their predictive accuracy. The RNN-LSTM model performed the best with the lowest RMSE of 18.42, demonstrating relatively accurate predictions. Linear Regression followed with a higher RMSE of 23.3, indicating a moderate increase in prediction error compared to RNN-LSTM. ARIMA showed the highest RMSE of 56.94, signifying the least precise forecasts and a significant margin of error compared to the other models. Overall, the RNN-LSTM model outperformed the alternatives, making it the most effective for this dataset.

5. Conclusion and Future Work

The problem of crime affects society and people daily all around the globe. The current trend in their culture is data mining and crime prediction systems. It aims to lower crime. Occurrence by forecasting future criminal activity based on the crime information that is accessible. This study analyzed crime prediction and classification using various data mining techniques and machine learning models. A structured crime dataset from 2006 to 2016. Linear Regression and RNN-LSTM models were evaluated using performance metrics like RMSE and MAPE. RNN-LSTM outperformed others, achieving the lowest RMSE of 18.42, highlighting its superior predictive accuracy. While the ARIMA

model demonstrated potential for handling complex temporal patterns, its higher error suggests the need for refinement in temporal applications.

Future work will focus on enhancing prediction accuracy by exploring ensemble models and hybrid approaches. Extending datasets to include recent and diverse records, integrating additional features like economic or demographic data, and leveraging advanced deep learning models like Transformer architectures could further improve outcomes.

References

[1] S. Sathyadevan, M. S. Devan, and S. Surya Gangadharan, "Crime analysis and prediction using data mining," *1st Int. Conf. Networks Soft Comput. ICNSC 2014 - Proc.*, pp. 406–412, 2014, doi: 10.1109/CNSC.2014.6906719.

[2] C. Chauhan and S. Sehgal, "A review: Crime analysis using data mining techniques and algorithms," in *Proceeding - IEEE International Conference on Computing, Communication and Automation, ICCCA 2017*, 2017. doi: 10.1109/CCAA.2017.8229823.

[3] M. Liu and T. Lu, "A Hybrid Model of Crime Prediction," in *Journal of Physics: Conference Series*, 2019. doi: 10.1088/1742-

[4] M. Feng *et al.*, "Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data," *IEEE Access*, 2019, doi: 10.1109/ACCESS.2019.2930410.

[5] A. Almaw and K. Kadam, "Crime Data Analysis and Prediction Using Ensemble Learning," in *Proceedings of the 2nd International Conference on Intelligent Computing and Control Systems, ICICCS 2018*, 2018. doi: 10.1109/ICCONS.2018.8663186.

[6] V. K. Putri and F. I. Kurniadi, "Crimes prediction using spatio-temporal data and kernel density estimation," in *2019 Asia Pacific Conference on Research in Industrial and Systems Engineering, APCoRISE 2019*, 2019. doi: 10.1109/APCoRISE46197.2019.9318972.

[7] S. Kim, P. Joshi, P. S. Kalsi, and P. Taheri, "Crime Analysis Through Machine Learning," in *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON 2018*, 2018. doi: 10.1109/IEMCON.2018.8614828.

[8] S. Sivaranjani, S. Sivakumari, and M. Aasha, "Crime prediction and forecasting in Tamilnadu using clustering approaches," in *Proceedings of IEEE International Conference on Emerging Technological Trends in Computing, Communications and Electrical Engineering, ICETT 2016*, 2017. doi: 10.1109/ICETT.2016.7873764.

[9] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J. M. Benítez, and F. Herrera, "A review of microarray datasets and applied feature selection methods," *Inf. Sci. (Ny).*, 2014, doi: 10.1016/j.ins.2014.05.042.

[10] L. McClendon and N. Meghanathan, "Using Machine Learning Algorithms to Analyze Crime Data," *Mach. Learn. Appl. An Int. J.*, 2015, doi: 10.5121/mlaij.2015.2101.

[11] P. Chen, H. Yuan, and X. Shu, "Forecasting crime using the ARIMA model," in *Proceedings - 5th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2008*, 2008. doi: 10.1109/FSKD.2008.222.

[12] Routhu, K., Bodepudi, V., Jha, K. M., & Chinta, P. C. R. (2020). A Deep Learning Architectures for Enhancing Cyber Security Protocols in Big Data Integrated ERP Systems. *Available at SSRN 5102662*.