Crime Prediction Using Machine Learning

[1]Nidhi Singh M, [2]Alidini Ajay Kumar Reddy, [2]Betharasi Anand, [2]Konjeti Naga Venkata Sai Himanth,
[2]Simham Nagasai
[1]Assistant Professor, Department of Artificial Intelligence and Machine Learning, R L Jalappa institute of technology
Doddaballapur, Karnataka, India. [2]Student,
[1]nidhisinghm@rljit.in,[2]Ajaykumarreddyalidini@gmail.com ,[2]anandanand7597@gmail.com ,
[2]Konjetihimanth24@gmail.com ,
[2] simhamnagasai0@gmail.com

ABSTRACT

Introduction The current state of law enforcement is changing; it is transitioning from a mere response to crimes after they occur to crime prevention prior to occurrence. The issue with traditional analysis techniques is that they are ineffective in dealing with the complex, huge amount of data that a city generates. To address this concern, we developed a hybrid model that blends traditional machine learning techniques with deep learning techniques. In our model, we apply a multi-step strategy; we apply K-Nearest Neighbors (KNN), Random Forest, and XGBoost to recognize various types of crimes simultaneously. Meanwhile, we employ Long Short-Term Memory (LSTM) networks to recognize patterns within time-series data. The results of our experiments verify that it is indeed much more accurate to apply ensemble algorithms together with neural nets compared to applying a single method.

Keywords: Crime Prediction Models, Artificial Neural Network Models, Ensemble Models, LSTM Models, Temporal Analysis.

I. INTRODUCTION

When crime levels rise, this is a danger to all of us and to the economy. With a growing urban population and increased travel patterns, crime patterns also evolve. Law enforcement now has to adapt to these patterns. In previous decades, a typical method of crime analysis was looking backwards in order to predict outcomes based on past events.

The problem is that these traditional approaches simply take too long for the age of "big data." Traditional policing will typically only take action when it receives a call to act. This serves as a further reason why we find ourselves at a point in time when better data systems will be able to radically transform this process altogether. With machine learning, we now have the opportunity to search through complex data sets for unknown connections between the "where," "when," and "what" of a crime. Furthermore, since crime tends to be seasonal or show "long-term" cycles, we incorporated the use of LSTM deep learning models in order to better account for this.

II. RELATED WORK

Research into data-driven safety has expanded significantly as scientists explore how mining algorithms can predict hotspots.

Sathya Devan et al. [1] utilized data mining on historical records to find patterns based on geography, though their model struggled with long-term time dependencies.

Gupta & Jain [2] tested Naïve Bayes and Decision Trees for classification; while accurate at identifying crime types, they could not forecast future trends effectively.

Mishra et al. [3] used K-Means clustering to map "hotspots," but found the temporal accuracy to be only moderate.

Keyvan pour et al. [4] focused on the importance of feature selection, but their system required too much manual effort to be truly scalable.

Deepak Kumar et al. [5] proved that ensemble methods like Gradient Boosting increase classification accuracy, yet they ignored the time-series element of the data.

Singh et al. [8] demonstrated that LSTM networks are superior for temporal dependencies, though they require massive datasets to function well.

Hybrid models, that is, the integration of machine learning concepts with those of deep learning, have also been considered. **Hassan et al.** in [9] developed an approach that combined ensemble classifiers along with an LSTM model for better prediction.

Likewise, Aishwarya et al. in [10] developed an approach for real-time crime prediction that had optimal computation cost, but upon analyzing large data, it had limitations.

Mandalapu, V., Elluri, L., Vyas, P., & Roy, N.[11] (2023). Crime prediction using machine learning and deep learning: A systematic review and future directions. IEEE Access, 11, 60153–60179.

[12] **Zhang, X., Liu, L., Xiao, L., & Ji, J. (2020).** Comparison of machine learning algorithms for predicting crime hotspots. IEEE Access, 8, 170383–170394.

[13] **Safat, W., Asghar, S., & Gillani, S. A. (2021).** Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques. IEEE Access, 9, 70080–70094.

[14] **Meenakshi Sundaram, S., Sivakumar, P., Parthiban, M., & U. Revathy (2022).**Crime hotspot analysis using six layered deep recurrent neural networks. International Journal of Early Childhood Special Education (INT-JECSE), 14(14), 508–520.

III. PROPOSED MODEL

Our model creates a "synergy" between standard Machine Learning and Deep Learning. The goal is to turn raw historical files into "actionable intelligence," allowing police to move from a defensive posture to a proactive one.

A. Data Processing and Feature Engineering

We ingest data including crime categories, GPS coordinates, and precise timestamps. To ensure the model is reliable, we perform "deep cleaning"—this involves filling in missing values via statistical imputation and removing duplicate records that might bias the results. We also normalize numeric features so that different scales (like zip codes vs. hours) don't confuse the algorithm.

B. The Dual-Stream Analysis Module

Classification (The "What"): We use KNN as a baseline to find similarities in local data distributions. However, the "heavy lifting" is done by

Random Forest and XGBoost, which combine multiple decision paths to reach a more accurate conclusion about the likely type of crime.

Forecasting (The "When"): We process the data into a sequential format for an LSTM network. Unlike standard regression, the LSTM has a "memory cell" that remembers seasonal fluctuations—such as spikes in theft during holiday seasons.

IV. System Overview

This framework is essentially a deep analytics tool. It translates huge amounts of history into smart information. By using machine learning for classification and Deep Learning for forecasting, we help police organizations shift from reacting to problems to managing safety proactively.

- **Data Acquisition & Preprocessing:** We gather data from police databases. Since raw data often has errors, this module cleans it up, handles missing values, and removes repetitions.

- **Crime Classification Module:** This determines the most probable crime based on location and attributes. We use KNN for baseline analysis and Ensemble Learning (Random Forest, XGBoost) to find complex, non-linear patterns.

- **Crime Forecasts Module:** Running at the same time, this module predicts *when* crimes will happen. It uses LSTM networks because they are excellent at learning from sequences and spotting long-term trends like seasonal spikes.

- **Evaluation and Output Module:** Finally, we validate the results. We check the models using accuracy, precision, recall, and F1 scores. The output gives law enforcement a list of probable crimes and their expected frequency.

V. METHODOLOGY

Imagine our framework to be a pipeline that processes input data to create intelligence. It is modular in nature so that we can improve a module without tearing down the entire infrastructure

A. Data Acquisition and Deep Cleaning

We begin with the data set from the law enforcement agency. Each entry contains the type of crime, location in the form of GPS co-ordinates, and the date. We begin our data set with dirty data. We delete missing values for crime type and get rid of the duplications so that our model won't be skewed towards incidents that happened more than once. "Noise," the useless data, is removed

B. Feature Engineering and Transformation

Since machine learning algorithms don't work on raw text, it's important we transform these inputs into a numerical format. We use encoding techniques for this purpose: to convert categorical data like neighborhood names or specific crime descriptions into numbers understandable by the system. We also decompose the timestamps into Day, Month, and Year. This is useful in finding the cyclic patterns; for example, the incident rate may increase on weekends or during certain months.

C. The Dual-Stream Analysis Engine The processing environment has two streams:

_Classification: _ A multi-level classification is used in order to classify a type of crime. The local patterns from the data are identified using a technique known as K-Nearest Neighbors (KNN). Furthermore, to avoid overfitting, a situation where the system learns the data, a Random Forest, an XGBoost, or a combination of these two can be applied since these two methods are able to learn complex decisions.

Forecasting: To know when these events will occur, we employ the use of an LSTM network. This is mainly because the network has a unique architecture that retains data for a long period of time and is highly effective in detecting patterns in trend data over a period of a few years. Evaluation and Visualization. Before any insight is presented to the user, the models undergo rigorous validation. We split the dataset into an 80:20 ratio (Training vs. Testing) to ensure the system is predicting unseen events rather than memorizing history. Performance is measured using metrics such as Accuracy, Precision, Recall, and F1-Score. The final output is visualized on a dashboard, translating complex probabilities into clear trend charts and heatmaps that allow police commanders to make strategic decisions quick.

D. Model Evaluation and Visualization

These models must be validated for any results before they are given to the final user. We divide our data into 80:20 to see that our model performs better in new events that it has not been trained to do. Then the results are given in the form of an interactive dashboard that changes the graph values into possible results that must be taken by the command of the law enforcement.

VI. Dataset Description

A structured data set has been used in this paper that has been collected over the years from all over India. This data set consists of tens of thousands of distinct data sets. The most minute columns in this data set are nature of incidents like Theft and Assault, geographic data that consists of District/State information, and then there are features related to time. Various categories of distinct data can be used for training in this regard that have been restructured in terms of time series patterns; this means that it has “seen” an incident happen in terms of an occurrence sequence in training for the LSTM model.

Model Evaluation:

In All these are related to a set of metrics: accuracy, precision, recall, as well as F1 metrics. In relation to the forecasting process, there is relevance to investigate the metrics that can be used in determining the accuracy of were compared on the basis of Mean Absolute Error and Root Mean Square Error. This is an essential function it performs in making some predictions possible.

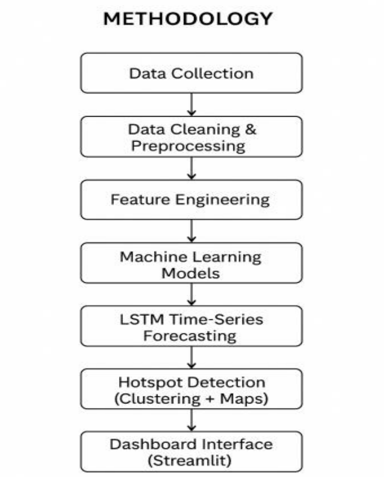

METHODOLOGY
Data Collection
Data Cleaning & Preprocessing
Feature Engineering
Machine Learning Models
LSTM Time-Series Forecasting
Hotspot Detection (Clustering + Maps)
Dashboard Interface (Streamlit)

Fig 1: Methodology

Work Flow Summary:

1. Data Foundation - Acquisition and Preparation: In fact, it all starts with crime statistics from past years because our whole system is grounded on these facts as our ‘ground truth.’ The bottom line is that data from the true world can or cannot contain ‘noisy patterns’; thus, it is necessary to include a process for data preparation prior to all these. Accordingly, we will eliminate the potential errors existing in the data by removing the repeating data to ensure that we acquire the best data for our model to produce good decision-making.

2. Feature Engineering - After completing data cleaning, it moves to Feature Extraction & Encoding. This is where we undergo the process of turning data into a language understandable to the computer language. We make use of the categoric data available pertaining to the type of crime or/and zone where it was committed to transform this data into numbers. We also transform this data into a format where the system can extract a pattern with concern to the factor of time pertaining to days, months, or season.

3. Dual-Stream Intelligent Analysis –

The central processing task is split into two streams:

Crime Classification: For categorizing what sort of crime might happen, it leverages a highly effective blend of algorithms. K-Nearest Neighbours (KNN) is used as a benchmark for calculating similarity, whereas sophisticated ensemble algorithms such as Random Forest and XGBoost are used for dealing with intricate, nonlinear relationships.

Time Series Forecasting: On the other hand, in order to determine the periods when the crime rates may exceed, the design uses LSTM (Long Short-Term Memory) networks. This is a part of the AI solution that uses the data to provide forecasts of future trends.

4. Validation and Output

The final stage entails Model Evaluation based on performance metrics (Accuracy, Precision, Recall, and F1-Score) to ensure that the predictions made (Classification of crime types and Trend prediction) are statistically valid before being used for decision-making purposes by the law enforcement agencies.

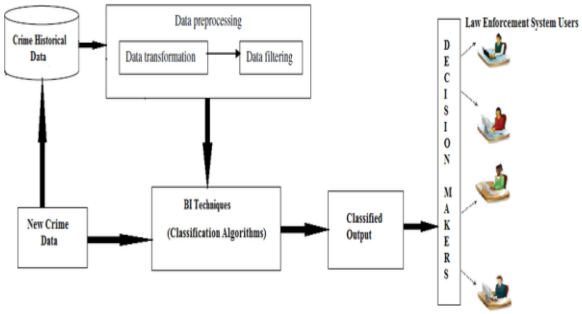


Fig 2: Workflow Diagram

Data Collection Process

The data gathering phase is actually the backbone on which this crime prediction model is designed. Given that the quality of any AI model is highly dependent on the data fed to this model, we made sure to obtain high-quality data from law enforcement agencies and other research platforms.

Discussion of the Data

What is the Data ?

The dataset is constructed based on organized records of distinct events of crimes. To allow the models to form an understanding of what exactly is happening, where, and when, each record contains:

•Type of Crime: (e.g. Theft, Assault

•Location: This is where the event occurred.

•Timestamp: The actual date and time.
•Frequency: Similar events occur with what frequency in this region.

Pattern Recognition in Time-Series Data Sets
The system will need information that will take a number of years in order for it to ensure that it is not interpreting ‘snapshots’ but rather trends. The information is needed in order for it to be identifiable for predictive purposes. The pattern has to be identifiable for it to be predictable.

Privacy and Ethics

Data Integrity: This involves not just accuracy, but responsibility as well. We naturally ensured anonymization of all data. There were no references made to names, specific addresses, or anything of the sort. This was an academic endeavor, based on trends regarding public safety.

Preparation for Analysis
However, This data was formatted in a standard format like CSV. The first screening was performed in an attempt to point out the missing values as well as repeated values in an attempt to have our data ready for modelling.

5.THE HUMAN PREPARATION

1. **"Deep Clean"**
First, we cleaned the dataset before carrying out the analysis. The cleaning operations involved:

• Handling Missing Information: Imputation techniques are used to fill in the missing information or removal of incomplete information when that would be important and fair to prevent the network from learning any bias unfair information.

•Removing Duplicates. We eliminated all duplicates from being considered, so that certain events were not inadvertently given 'extra weight' by the model.

•Filtering Noise: The irrelevant information that wasn't useful in making predictions about crimes was removed, leaving the most influential information.

2. Translation & Scaling

Algorithms do not read words but numbers. In the attempt to fill this gap, two approaches have been employed:

•Encoding: We encoded categories such as "Crime Type" or "Location Name" to numerical values.

• Normalization: All numerical variables have been normalized to have the same scale. It helps to counter the effect that could result in an insignificant variable becoming too important due to its large number compared to another significant variable such as the time of the day in the case of the zip codes.

3. Structuring for Time-Series Analysis

Because our system is built using LSTM (Long Short-Term Memory) networks, time treatment is essential in our system.

•We extracted dates by breaking it down into individual properties: Day, Month, and Year. *After that, it was necessary to implement this data in a sequential fashion so that it is visible to the model itself and ‘sees’ it in that order. It is this that makes it possible to ‘understand’ the concept of temporality, such that, perhaps, there is an increase in certain crimes that take place during holidays or seasons.

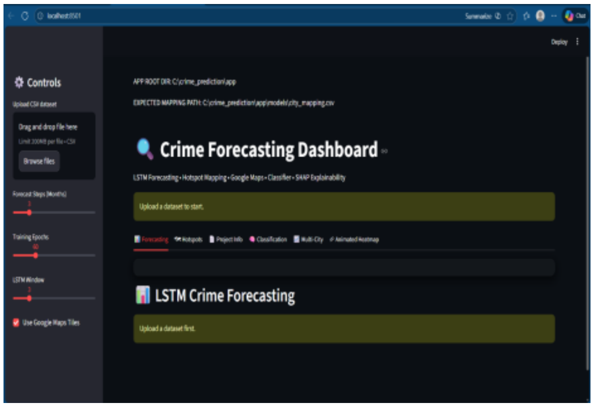
4. The Final Split: Training vs. Testing

In order to ensure that the system truly works well in the real world and is more than just "memorizing" history, we split the final data set into two:

•The Training Set: It is used to train the models on patterns and trends.

•Testing Set: To be kept distinct and utilized as a 'final exam', testing how effectively the developed model can predict the crime on unseen data.

The Result: A well-structured and normalized high-quality dataset that is a perfect engine for crime classification and prediction.



APP ROOT DIR: C:\crime_prediction\app

EXPECTED MAPPING PATH: C:\crime_prediction\app\models\city_mapping.csv

Crime Forecasting Dashboard

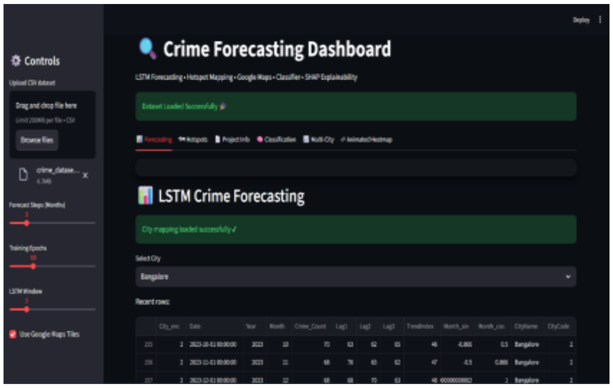
LSTM Forecasting • Hotspot Mapping • Google Maps • Classifier • SHAP Explainability

Upload a dataset to start.

LSTM Crime Forecasting

Upload a dataset first.

Fig 3: Home Page



Crime Forecasting Dashboard

LSTM Forecasting • Hotspot Mapping • Google Maps • Classifier • SHAP Explainability

Dataset Loaded Successfully 🎉

LSTM Crime Forecasting

City mapping loaded successfully ✓

Select City: Bangalore

Recent rows:

	City_enc	Date	Year	Month	Crime_Count	Lag1	Lag2	Lag3	TrendIndex	Month_sin	Month_cos	CityName	CityCode
255	1	2023-10-01 00:00:00	2023	10	70	63	62	65	46	-0.866	0.5	Bangalore	1
256	1	2023-11-01 00:00:00	2023	11	68	70	63	62	47	-0.5	0.866	Bangalore	1
257	1	2023-12-01 00:00:00	2023	12	68	68	70	63	48	[illegible]	1	Bangalore	1

Fig 4: Dataset Upload Page

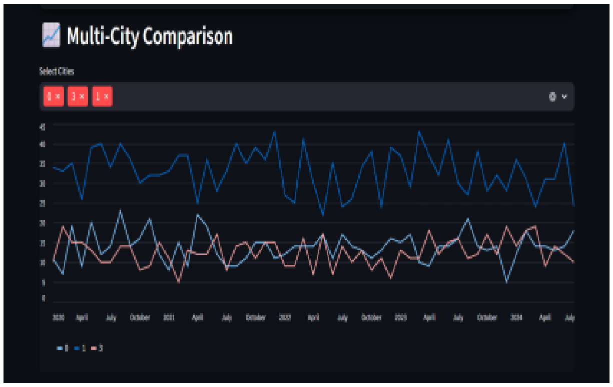


Fig 5: Multi-City Comparison

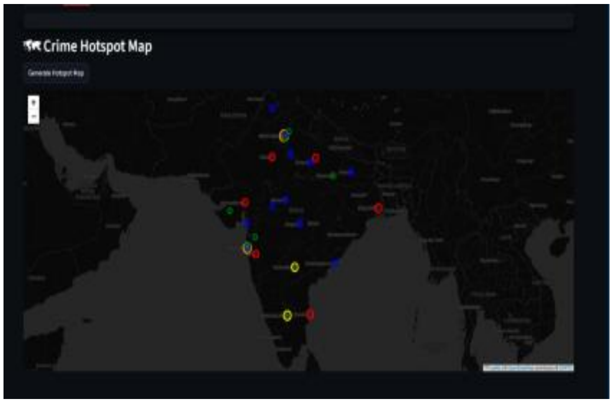


Fig 6: Hotspot Map

Dataset Description:

In the proposed research, the crime data from the Indian subcontinent is utilized. It spans several years. This data is the basis of the training of the proposed ML/Deep Learning algorithm. It contains a vast number of records, where every record is a crime. It is not topics of classification that are detailed, but the features are spatiotemporal.

The most informative data, as far as relevance is concerned, would involve the type of crime, the state or district where it has occurred, and the exact time when it took place. We have further developed these initial variables to make the data more informative when it comes to analysis by breaking down these variables into their distinct parts, which would involve the year, month, day, and time.

Data is to be divided, preferably in an 80:20 ratio, into training and test datasets for model evaluation. This training data was then used to train models on crime pattern detection, while the test data tested the accuracy levels of the predictions. The crime data provided does have enough efficacy to present a good base for establishing crime prediction models with the help of machine learning algorithms, such as Random Forest, Support Vector Machine, or even deep learning networks such as LSTM models. The data does have adequate structure to analyze the trends in criminal activities and, based upon these, easily forecast further incidences.

VII. SYSTEM DESIGN

1. Layers of Data Input & Preprocessing

The Data Input Layer is the very foundation of this system, and works to ingest historical crime data in the form of CSVs or databases, which include the crime types, their locations, and their timestamps.

The Preprocessing Layer works like a filter because raw data is never perfect. It removes duplicates, fills in missing values, and normalizes numbers. Importantly, it also handles "data balancing," which means it ensures the model won't get biased toward common crimes and ignore the rarer ones that are more serious.

2. Feature Engineering Layer

This layer provides the "translation." It takes raw timestamps and breaks them down into specific Temporal Features such as the day of the week, the month, or even specific time intervals. This is crucial to be comprehended by the forecasting module in order to understand "when" crimes are most likely to take place, such as spikes during holidays or weekends.

3. "Brain": Classification & Forecasting Modules

The system analyzes crime on a two-pronged approach:

Classification of Crime - "What"

In this process of incident categorization, the machine learning algorithm applies models such as Random Forests and XGBoosting for this purpose of incident classification. Though KNN acts as a baseline in this process of incident classification, it is actually these ensemble methods of machine learning techniques that perform the major task of arriving at an even more precise solution through multiple routes of decision-making in the process of coming.

• Crime Forecasting (The “When”): In relation to the module which uses the LSTM network, LSTMs are deep learning architectures that boast the capability of memorizing past trends.

4. Application Layers of Evaluation &

These models determine an intended outcome before the result and are compelled before the result to pass a test of evaluation or assessment entitled "Evaluation Module". In view of "What's at stake", the one for "Accuracy" is the guarantee of "all is well as regards the system"; therefore, there is a concern about "validation" regarding "The "Results," in view of "Precision", "Recall", "Measures regarding F1."

The name of the last process outcome of the analysis of data performed is “Output & Visualization.” Data interpretation performed through the production of “Valuable Insights” within the process could be very helpful once the data analysis process is over. The process has been designed in a manner that its final development outcome helps in creating data emerging out of the process which can easily be interpreted by the law and enforcement department using the trend charts.

Why this design works:

Modularity: It follows a modular architecture that greatly supports maintenance. Changing or tuning any AI model involves merely local changes, and neither the entire system has to be dismantled nor rebuilt.

Scalability: The architecture is designed to be scalable and handle high volumes of data. This facilitates parallel processing of inputs from multiple cities or even regions without the loss of performance.

Actionable means going beyond 'data points' to proactive, data-driven strategy on public safety.

VIII. CONCLUSION

This research successfully establishes an intelligent crime prediction model that synthesizes machine learning classification with deep learning forecasting. By integrating Random Forest and XGBoost with LSTM networks, we captured both the spatial and temporal nuances of criminal activity . Our evaluations confirm that while baseline methods like KNN are insufficient for complex datasets, our hybrid ensemble-LSTM approach significantly reduces errors and better handles non-linear patterns.

The system provides a robust framework for proactive policing, offering data-driven insights for resource planning. However, the model's efficacy remains tied to the quality of historical data; predicting extremely rare "black swan" events remains a challenge . Future iterations of this project will focus on integrating real-time data streams and exploring advanced deep learning architectures to further refine predictive capabilities.

IX. ACKNOWLEDGMENT

We extend our sincere gratitude to **Dr. P. Vijayakarthik**, Principal, RLJIT, and **Dr. Manjunatha B N**, Head, Dept. of AI&ML for their support. We profoundly thank our guide, **Ms. Nidhi Singh M,** for his invaluable technical guidance.

X. REFERENCES

[1] K. Sathya Devan et al. proposed a crime analysis and prediction system that utilizes data mining techniques on historical crime data. The work focused on identifying crime patterns based on location and time using classification algorithms. The proposed system demonstrated an improved prediction accuracy for structured datasets but limited in handling temporal dependencies.

[2] A. Gupta and R. Jain proposed a model of crime prediction based on the machine learning algorithms Decision Trees and Naïve Bayes. For this model, it is observed that with the help of some analysis of records of crimes, the system could classify the type of crimes and predict them in the future. The performance results indicated its best performance in the classification task, but it is not strong enough to present correct forecasts of long-range crime trends.

[3] S. Mishra et al. proposed a framework for crime hotspot prediction using clustering and classification techniques. The system utilized K-Means clustering to identify crime-prone regions and performed classification using Random Forest. The results exhibited a very effective spatial analysis of the crime, but the accuracy regarding time prediction was found to remain moderate.

[4] M. R. Keyvan pour et al. proposed an integrated crime prediction method using both data mining and machine learning. In this work, the author highlighted feature selection and preprocessing to enhance model performance. The developed system has reached reasonable results, but its development needs exhaustive manual feature engineering.

The paper "Crime Prediction using Ensemble Learning" by Deepak Kumar et al. proposed a crime prediction model using ensemble learning techniques including random forests and gradient boosting. Their model significantly improved the classification performance by assembling multiple decision trees. However, the time-series crime patterns have not been captured in an explicit manner.

[6] Harshita Verma et al. proposed a crime prediction system that utilized Support Vector Machines and Random Forest algorithms. The system employed historical crime data to forecast categories and locations of crimes. Results showed a high accuracy on the known patterns of crimes but a reduced performance in the case of rare types of crimes.

[7] R. K. Sharma et al. proposed a crime forecasting model by analyzing the criminal activities using time-series analysis. Their method identified that month and year can be considered as important temporal features in predicting the criminal trends, while in practice, traditional methods of time-series failed in capturing the complex long-term dependencies.

It presented a deep learning-based crime prediction system, P. Singh et al. introduced a deep learning-based crime prediction system using Long Short-Term Memory (LSTM) networks. This model was successful in extracting the temporal dependencies of the data in crime incidents and thus provided better performance with regards to accuracy in forecasting than traditional machine learning techniques. The limitation of this approach was its requirement for large training datasets.

Mona Hassan et al. proposed a hybrid crime prediction system that used the combination of machine learning classification and deep learning-based forecasting. In their work, they have introduced an integrated ensemble model with LSTM for time series-based prediction, which resulted in higher accuracy. Their results have proved that hybrid models work more effectively in crime prediction.

[10] Aishwarya R. et al proposed a real-time crime

prediction architecture using machine learning models to minimize computational cost. The proposed technique works well with the urban crime datasets, whereas when applied to complex ones with large-scale information, the model produces lesser accuracy.

[11] Mandalapu, V., Elluri, L., Vyas, P., & Roy, N. (2023). Crime prediction using machine learning and deep learning: A systematic review and future directions. IEEE Access, 11, 60153–60179. https://doi.org/10.1109/ACCESS.2023.3286384

[12] Zhang, X., Liu, L., Xiao, L., & Ji, J. (2020). Comparison of machine learning algorithms for predicting crime hotspots. IEEE Access, 8, 170383–170394. https://doi.org/10.1109/ACCESS.2020.3026420

[13] Safat, W., Asghar, S., & Gillani, S. A. (2021). Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques. IEEE Access, 9, 70080 70094. https://doi.org/10.1109/ACCESS.2021.3078817

[14] Meenakshi Sundaram, S., Sivakumar, P., Parthiban, M., & U. Revathy (2022). Crime hotspot analysis using six layered deep recurrent neural networks. International Journal of Early Childhood Special Education, 14(14), 508–520. https://doi.org/10.48047/INTJECSE/V14I8.309.