



Spatial Clustering and Machine Learning for Crime Prediction: A Case Study on Women Safety in Bhopal

Yamini Sahu  and Vaibhav Kumar 

Indian Institute of Science Education and Research Bhopal (IISERB)
{yamini21,vaibhav}@iiserb.ac.in

Abstract. A crime is an unlawful action subject to punishment by a governing authority, causing harm not only to individuals but also to the well-being of a community, society, or the state. According to the latest annual report from the National Crime Records Bureau of *India*, there were 445,256 cases of crimes against women registered in 2022, representing a 4% increase from the 428,278 cases reported in 2021. India experiences an alarming rate of 51 cases of crimes against women per hour. These figures underscore the urgent need for proactive measures to ensure women safety and secure their continued contribution to the country's development.

Our research employs a predictive approach to address crimes against women, utilizing spatial analysis and crime prediction models to pinpoint high-risk areas in urban settings and accurately forecast crime trends. We focused on *Bhopal*, the capital of *Madhya Pradesh*, one of India's 28 states, and gathered crime data against women from 30 police stations in *Bhopal*. Our study showcases the application of spatial clustering techniques to identify hotspots for various crimes against women (murder, attempted murder, rape, gang rape, kidnapping, dowry-related offenses, and molestation). Additionally, we employed machine learning regression models and advanced forecasting techniques to predict crime rates. Our model, based on decision tree regression, exhibited a very low mean squared error of 0.0417 and a mean absolute error of 0.083. Furthermore, our analysis revealed that classical machine learning regression models outperformed advanced forecasting models, such as long short-term memory, given our limited dataset. Thus, detecting and predicting crime hotspots derived from historical crime data can enable law enforcement agencies to develop targeted intervention strategies customized for specific crimes occurring at particular locations.

Keywords: Spatiotemporal · Crime against women · Pattern analysis · Urban planning

1 Introduction

The 21st century has seen an unprecedented global migration into urban areas, leading to the term "Century of the City" [1, 2]. This ongoing urbanization is

causing significant social, economic, and environmental changes in urban areas. Organizations responsible for city management and providing essential services, such as transportation, air and water quality, public safety, and resource planning (e.g., water and power), may need help addressing these issues [3]. Moreover, in cities with higher crime rates, crime spikes have emerged as a critical social issue, impacting not only public safety but also adult socioeconomic status, health, education, and child development [4, 5].

A growing amount of urban-related data, including spatial and temporal attributes ranging from weather to economic activity, is available to public organizations, including police departments, for integration with internal data, such as crime happening in the cities. Urban-related data presents an opportunity to apply data analytics methodologies to derive predictive health, water, and energy management models, etc. Similarly, these urban data can be used to derive predictive models related to crime events. Such models enable police departments to optimize their limited resources and develop more effective crime prevention strategies.

Research on criminal justice indicates that criminal events are not evenly distributed within a city, and the crime rates vary based on geographic location (e.g., low-risk and high-risk areas) and temporal patterns (e.g., seasonal fluctuations, peaks, and dips). Therefore, an accurate predictive model must automatically identify areas more susceptible to crime events and understand how crime rates fluctuate over time in each specific area. This knowledge can help police departments allocate resources efficiently, focusing on high-risk areas and adjusting strategies based on changing crime trends.

A crime constitutes an unlawful act subject to punishment by a governing authority, posing harm not only to individuals but also to the well-being of a community, society, or the state. Research suggests that crimes display detectable patterns across geographical regions within specific timeframes, offering insights for proactive measures by law enforcement. The recurrent incidence of specific criminal events in a given area over time can be termed a spatiotemporal event, facilitating the prediction of future crime incidents using historical spatiotemporal data. Machine learning, deep learning, and data mining techniques serve as effective tools for crime prediction.

In India, the number of reported cases of crime against women (CAW) has shown a concerning trend over recent years. According to data released by the National Crime Records Bureau (NCRB) [6], in 2020, there were 371,503 reported cases, which increased to 428,278 in 2021 and further rose to 445,256 in 2022. These figures indicate a worrying escalation in such incidents year after year. Notably, 51 First Information Reports (FIRs) were filed every hour on average across the country, highlighting the frequency and severity of these crimes. The consistent rise in reported cases underscores the urgent need for effective measures to address and prevent CAW [7].

Mitigating CAW is essential for fostering safer urban environments, which directly impacts social and economic development. High crime rates discourage women's participation in public life, reducing workforce diversity and overall pro-

ductivity. A city designed with gender-sensitive urban planning—incorporating well-lit streets, safe public transport, and community surveillance—enhances women’s safety, contributing to social cohesion. Additionally, lowering crime rates leads to improved mental and physical well-being, reducing healthcare costs. Addressing these issues is vital for creating inclusive cities where everyone can thrive, which is critical for sustainable urban development and long-term societal progress.

In this paper, we used spatial analysis and predictive modeling to automatically identify high-risk crime regions in urban areas and reliably forecast crime trends in each region. Our proposed algorithm consists of several steps: first, identifying high crime density areas (referred to as crime-dense regions or hotspots) through spatial analysis; then, discovering specific crime prediction models for each detected region based on the analyzed data. The resulting micro-level spatio-temporal crime prediction model comprises a set of crime-dense regions and associated crime predictors, with each predictor representing a model to forecast the expected number of crimes in its respective region.

1.1 Contributions

We have demonstrated the use case of our algorithm using a case study. We present an analysis of crimes within a particular region of *Bhopal*, encompassing approximately 4,500 crime events over four years. The crime data used in this work was obtained from the *Bhopal* Law Enforcement Agency¹. Experimental evaluation results demonstrate the effectiveness of our proposed approach, achieving good accuracy in spatial and temporal crime forecasting over time.

Our primary contribution lies in creating a micro-level spatiotemporal crime forecasting model. Our work is significant because it enables implementation on a smaller scale rather than attempting to apply the solution to a larger region, such as the entire state (in our case, *Madhya Pradesh*). Furthermore, our work demonstrates the utilization of limited data (from the years 2020 to 2023) to predict crime rates among women.

1.2 Paper Organization

Our paper is organized as follows. Section 2 provides a comprehensive review of approaches utilized in crime hotspot detection and prediction. In Sections 3, we present the proposed methodology used in this study. Section 4 discusses the proposed algorithm. The implementation and evaluation of the proposed method are presented in Section 5. Subsequently, Section 6 presents the results, while in Section 7, we discuss the model performance and implications of our work. Section 8 outlines the study’s limitations. Finally, Section 9 presents the conclusion, and section 10 outlines the future directions of our work.

¹ <https://ptsbhopal.mppolice.gov.in/>

2 Related Work

Existing studies have utilized various data mining techniques for crime data analysis and crime detection. Some researchers have concentrated on predicting crime locations, while others have aimed to detect crime patterns (hotspot detection). We categorize existing research on crime data into four categories: (i) crime prediction using machine learning, (ii) analysis of crime patterns, (iii) spatiotemporal crime analysis, (iv) utilizing diverse data sources for crime analysis.

2.1 Crime Prediction Using Machine Learning

Biswas et al. [8] utilized polynomial regression, linear regression, and random forest algorithms to forecast crime scenarios in Bangladesh, achieving success with the polynomial model. Similarly, Hajela et al. [9] employed machine learning (ML) and hotspot analysis for crime prediction, with REPTree showing promising results. Hossain et al. [10] achieved high accuracy using supervised ML algorithms on San Francisco crime records for crime prediction. Similarly, Kumar et al. [11] developed a crime prediction model using the KNN algorithm.

2.2 Analysis of Crime Patterns

Das et al. [12] conducted a behavioral analysis of violence against women in India, identifying perpetrator clusters using the Infomap clustering algorithm. Similarly, Tamilarasi et al. [13] focused on identifying regions where major crimes frequently occur and the types of crimes that persist over time, employing various ML algorithms. Lavanyaa et al. [14] delved into crimes against women in Tamil Nadu, India, aiming to discern patterns and predict occurrences of crime, thereby enhancing the operational efficiency of Tamil Nadu Police.

2.3 Spatiotemporal Crime Analysis

Ibrahim et al. [15] conducted spatiotemporal crime hotspot analysis using Chicago's crime dataset, comparing SARIMA (Seasonal AutoRegressive Integrated Moving Average) with LSTM (Long Short-Term Memory) for crime prediction. Similarly, Li et al. [16] analyzed urban crime in China using quantitative methods and ARIMA (AutoRegressive Integrated Moving Average) for spatiotemporal crime predictions. Butt et al. [17] focused on crime prediction in New York City, employing HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) and SARIMA models. However, Yi et al. [18] proposed a clustered CRF model for predicting crime based on spatiotemporal factors.

2.4 Utilizing Diverse Data Sources for Crime Analysis

Belesiotis et al. [19] explored diverse online data sources (six) to analyze and predict crime distribution in large urban areas, demonstrating improved prediction accuracy through integrated data. Similarly, Sivanagaleela et al. [20] focused on crime analysis and prediction employing the Fuzzy C-Means algorithm, showcasing its efficacy in crime analysis.

3 Proposed Methodology

Figure 1 shows the proposed methodology utilized in this study. It comprises data cleaning, crime hotspot detection, and crime prediction. Now, we will discuss the crime hotspot detection and crime prediction steps in detail. The discussion on data cleaning is provided separately in section 5.1.

In this study, we used CAW data, including incidents of rape, murder, and kidnapping, as input variables. We applied the K-nearest neighbor (K-NN) spatial clustering method to analyze the spatial distribution of these crimes. The output of this algorithm identifies spatial clusters of crime hotspots, allowing for a detailed examination of crime patterns and their potential impact on urban planning.

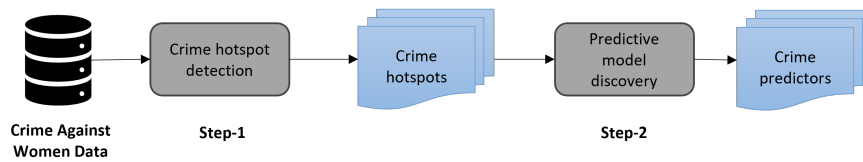


Fig. 1: Spatiotemporal crime against women prediction steps.

3.1 Crime Hotspot Detection and Crime Prediction

The crime hotspot detection method uses spatial clustering on the processed dataset, with each cluster representing a concentrated region of criminal activity. This spatial clustering approach utilizes clustering techniques like density-based, which aims to group objects together based on their proximity and density in the geographical space. This can be achieved by using spatial pattern analysis techniques to identify clusters by analyzing the estimated density distribution of the dataset.

3.2 Crime Prediction

Once the crime-prone areas are identified, predictive models can be employed to forecast future occurrences of crimes within these regions. These predictive models can be developed using classical machine learning algorithms and advanced forecasting techniques. Classical machine learning regression models, including Decision Trees (DT) [21], Random Forests (RF) [22], and K-Nearest Neighbors (KNN) [23], can be utilized for this purpose. Additionally, advanced forecasting models such as ARIMA (AutoRegressive Integrated Moving Average) [24], SARIMA (Seasonal ARIMA) [25], and LSTM (Long Short-Term Memory) [26] can be explored for predictions.

4 Proposed Algorithm

We have merged sections 3.1 and 3.2 to propose a crime prediction algorithm 1, which consists of two sequential steps: section 3.1 as step 1 and section 3.2 as step 2. In the first step, the algorithm focuses on crime hotspot detection, which identifies regions with a higher density of criminal activity by treating it as a geospatial clustering problem. The clustering method analyzes both spatial and temporal crime data to produce K clusters, each representing a high-crime-density region. The goal of this step is strictly to detect these hotspots. In the second step, the algorithm transitions to crime prediction. It utilizes the hotspots identified in step 1 as input to a predictive model, which forecasts future crime occurrences based on CAW data. Thus, the crime prediction model not only considers the historical geospatial clustering but also integrates additional predictive features such as temporal patterns and crime types.

Algorithm 1 Spatio-Temporal Crime Prediction

Require: D : Crime against women dataset

Ensure: $CAWH$: Crime against women hotspot, CP : Crime Predictors

- 1: **L1**: Crime hotspot detection // *Step 1 starts*
 - 2: Identify crime hotspots using spatial analysis techniques.
 - 3: $CAWH \leftarrow$ Locations identified as crime hotspots.
 - 4: $CP \leftarrow$ Predictors for crime prediction // *Step 1 ends*
 - 5: **L2**: Apply various algorithms // *Step 2 starts*
 - 6: Train various machine learning algorithms, including decision tree, random forest, ARIMA, SARIMA, LSTM, and KNN on the dataset.
 - 7: Evaluate the performance of each algorithm using metrics such as MSE and MAE.

 - 8: $MSE, MAE \leftarrow$ Performance metrics for each algorithm.
 - 9: **return** $CAWH, CP, MSE, MAE$ // *Step 2 ends*
-

5 Case Study

Crime data are sensitive in nature and typically not shared by law enforcement agencies. However, this data proves effective in understanding crime occurrences in various geographical locations and can be used for analysis to inform actions by law enforcement agencies accordingly.

To illustrate our proposed methodology and the algorithm, we have chosen Bhopal as our study area. Figure 2 shows the geographic coordinates of the study area, with latitude 23.28023161715492 and longitude 77.36346614827542, covering an area of 547.53 square kilometers. This case study focuses on the 30 police stations within the urban landscape of *Bhopal*. These stations span across diverse localities, including *TT Nagar*, *Ratibad*, *Habibganjh*, *Sahpura*, *Ashoka garden*, *Jhagirabad*, *Aishbag*, *Bajariya*, *Govindpura*, *Piplani*, *Awadhपुरi*, *M.P.*

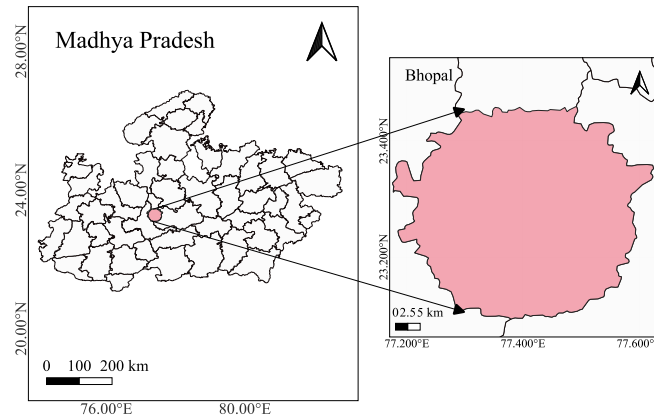


Fig. 2: Madhya Pradesh and the study area (in pink color) (left); *Bhopal* study area (right)

Nagar, Arera Hills, Bagsewaniya, Katara Hills, Misrod, Ayodhyanagar, Kotwali, Talaiya (Budhwara), Shahjhabad, Kohefiza, Tila Jamalpura, Mangalwara, Bairagadh, Khajuri Sadak, Kolar Road, Nishatpura, Gandhi Nagar, and Chhola Road.

5.1 Data Collection and Pre-processing

The data was collected from the law enforcement agency in Bhopal after a long process, which involved submitting numerous requests, making appeals, and obtaining permission from multiple higher authorities. The collected data for the police stations listed in section 5 was organized by year (temporal information) and only contained women-related crimes. This dataset covers the period from January 2020 to October 2023 and encompasses details on 25 categories of crimes, including murder, attempted murder, rape, gang rape, kidnapping, dowry-related offenses, molestation, and various others.

During the data preprocessing step, we identified instances of missing data and observed that local languages were used for crime definitions in some cases. To address this, we dropped the columns containing missing values and utilized Google Translate² to convert the local language terms into English. As a result, our final dataset has dimensions of 120 rows by 10 columns, where the first three columns denote latitude, longitude, and the year of the offense, respectively. The remaining seven columns correspond to seven categories of crimes (murder, attempted murder, rape, gang rape, kidnapping, dowry-related offenses, and molestation). Further, we divided the dataset into training and testing sets in a ratio of 80:20 while experimenting with the forecasting model. The split was

² <https://translate.google.com/>

performed without considering the temporal sequence of the data. While this method does not involve a time-based separation, it provides a representative distribution of the dataset across both sets. We acknowledge that this approach might allow cases from later in the dataset to appear in both the training and testing sets. However, for the purposes of this study, we aimed to maintain an equal distribution of data points across the two sets to maximize the performance and generalization ability of the model.

5.2 Detection of Crime-Dense Regions from the Data set

The objective of this study is to predict spatiotemporal crime patterns related to women. Figure 3 shows the cumulative number of crimes against women per year from 2020 to 2023. Similarly, Figure 4 shows the cumulative number of specific crimes against women, such as murder, kidnapping, rape, etc., per year from 2020 to 2023. From visual analysis, we observed a consistent increase in the number of crimes against women each year.

We utilized the proposed algorithm 1 to identify crime-dense regions within our study area. Figures 5, 6, 7, and 8 display the crime-dense regions in the Bhopal area, highlighting instances of rape, murder, kidnapping, and molestation against women, respectively.

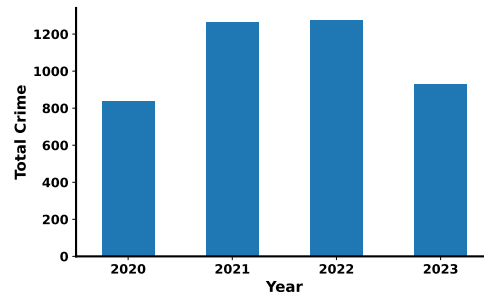


Fig. 3: According to the data shown in the figure, the number of reported cases of CAW has shown a consistent increase over the years. Specifically, in 2020, the number of cases was lower compared to subsequent years, namely 2021, 2022, and 2023. This trend indicates a continuous rise in reported cases from 2020 to 2023. However, it is worth mentioning that the data for 2023 only includes information up to October, which could explain the lower number of reported cases compared to the full-year data for 2021 and 2022.

5.3 Training and Evaluating the Predictive Model

We employed various regression techniques to forecast crime occurrences. The regression technique used in this paper includes classical machine learning mod-

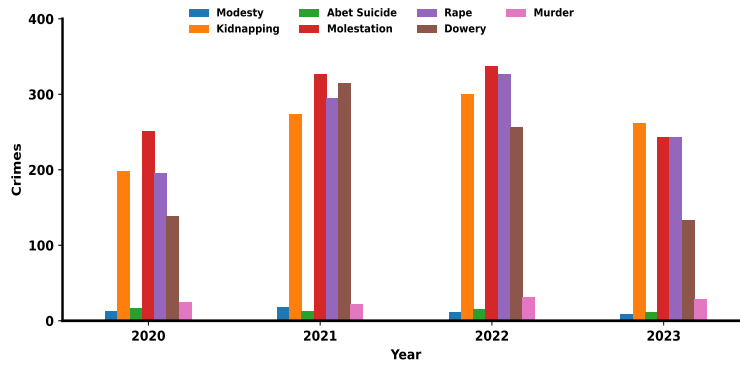


Fig. 4: From 2020 to 2023, a range of different crime cases have been observed. Especially the categories investigated in this study include modesty, kidnapping, abet suicide, molestation, rape, dowry-related crimes, and murder. Across all years, molestation cases notably stand out as being exceptionally high compared to other types of crimes, followed by rape cases and kidnapping incidents. There is a consistent upward trend in the number of cases reported annually. However, it is essential to note that the data for the year 2023 only includes information up to October, leading to a lower count compared to previous years. Furthermore, modesty and abet suicide cases consistently appear to be relatively low in number each year.

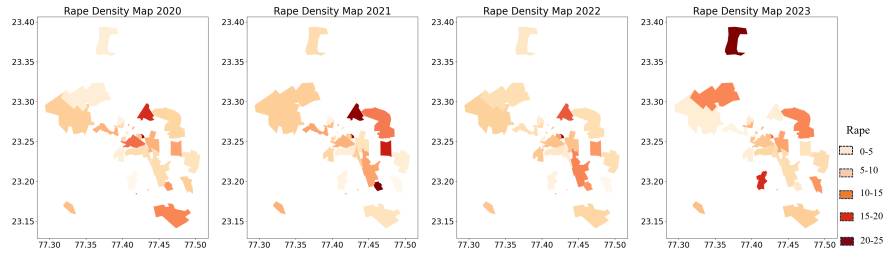


Fig. 5: Rape cases hotspot regions in *Bhopal* spanning from 2020 to October 25, 2023.

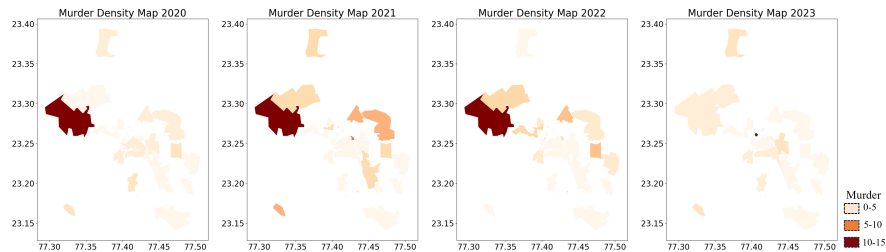


Fig. 6: Murder cases hotspot regions in *Bhopal* spanning from 2020 to October 25, 2023.

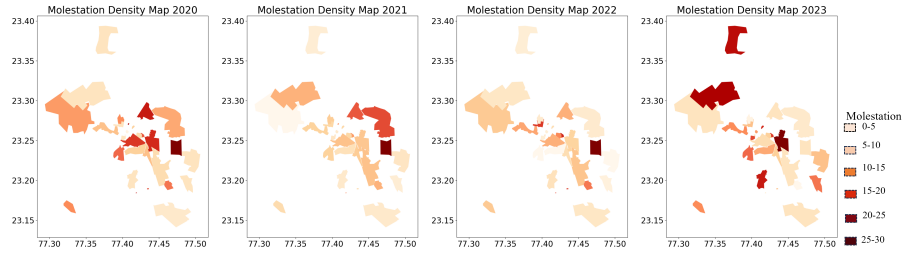


Fig. 7: Molestation cases hotspot regions in *Bhopal* spanning from 2020 to October 25, 2023.

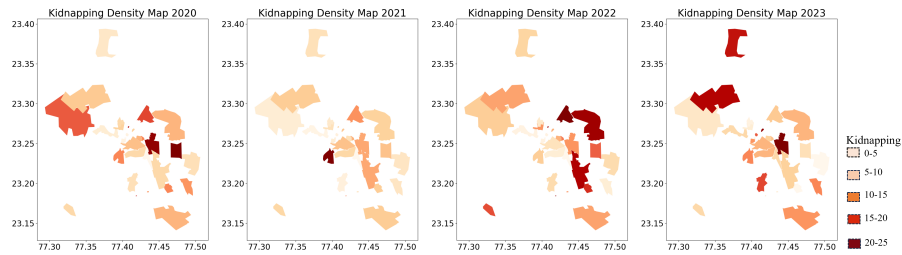


Fig. 8: Kidnapping cases hotspot regions in *Bhopal* spanning from 2020 to October 25, 2023.

els such as Decision Trees (DT), Random Forests (RF), and K-Nearest Neighbors (KNN), as well as advanced forecasting models including Long Short-Term Memory (LSTM) networks, Seasonal Autoregressive Integrated Moving Average (SARIMA) models, and Autoregressive Integrated Moving Average (ARIMA) models. The selection of these models aimed at leveraging their distinct capabilities and characteristics to predict crime patterns accurately.

DT offers a non-linear approach, breaking down data into hierarchical structures for classification, while RF is an ensemble learning method that excels at handling large datasets and minimizing overfitting by aggregating multiple DTs. However, KNN makes predictions based on the similarity of data points. LSTM is a recurrent neural network that captures long-term dependencies in sequential data and is potentially valuable for recognizing patterns in data over time. Additionally, SARIMA and ARIMA models, both time series forecasting techniques, are particularly suited for analyzing temporal patterns and trends in data.

Model Evaluation:

Mean square error (MSE) and mean absolute error (MAE) are standard metrics used to evaluate the performance of predictive models. MSE measures the average squared difference between the actual and predicted values, providing insight into the overall accuracy of the model’s predictions. It is calculated by taking the average of the squared differences between each predicted value and its corresponding actual value. On the other hand, MAE computes the average absolute difference between the actual and predicted values, offering a measure

of the model's precision. It is determined by averaging the absolute differences between each predicted value and its corresponding actual value. MSE and MAE are essential tools for assessing the effectiveness and reliability of predictive models, with lower values indicating better performance.

6 Results

We will discuss our findings on two main topics: the detection of crime-dense regions and crime prediction.

Crime-Dense Regions

Figure 5 showcase the concentration of rape cases in the Bhopal area from 2020 to 2023, offering insights into evolving patterns over time and across different locations. In 2020, *Ashoka Nagar* stood out as a hotspot for these incidents, signifying a cluster of cases in *Ashoka Nagar*. The subsequent year, 2021, witnessed a notable surge in rape cases in *Khajuri*. As time progressed to 2022, there was a significant increase in incidents in *Govind Pura* and *Pipilani*. By 2023, *TT Nagar* emerged as the area with the highest number of reported rape cases.

Similarly, Figure 6 shows the regions with the highest density of murder cases in the Bhopal area from 2020 to 2023, shedding light on evolving patterns over time and across various locations. *Bairagadh* recorded the highest number of reported crimes during this period. Notably, *Bairagadh* emerged as the central hub of criminal activity, with the highest rate of crime recorded from 2020 to 2022. However, in 2023, there is a noticeable shift, with the *Mangal Wara Momim Pura Gate* region experiencing a surge in murder incidents. This analysis delves into the evolving patterns of crime over time, highlighting the necessity for targeted interventions to address specific geographic areas in response to emerging trends. Understanding these trends is crucial for effectively implementing law enforcement measures and initiatives to ensure community safety.

Similarly, Figure 7 shows the regions with dense occurrences of Molestation cases in the Bhopal area. In 2020, *Piplani* had the highest number of reported molestation cases, while *Nishantpura*, *Govindpura*, *Jhangirabad*, and *TT Nagar* recorded moderate case numbers. The following year, 2021, witnessed a continuation of cases in *Piplani*, alongside moderate instances in *Ayodhyanagar* and *Nishantpura*. In 2022, *Piplani* reported the highest number of cases again, with *Sahajanbad* and *Ashokanagar* documenting moderate occurrences. This suggests that *Piplani* is a hotspot for molestation cases, and the necessary intervention is needed.

Similarly, Figure 8 illustrates the regions with dense occurrences of kidnapping cases in the Bhopal area. In 2020, *Govindpura* and *Piplani* emerged as hotspots with the highest number of reported kidnapping of women cases, indicating concentrated criminal activity in these regions. *Bairagadh* and *Nishantpura* recorded moderate case numbers during the same period. However, the dynamics shifted in 2021, with *TT Nagar* taking the lead in reported cases. By 2022, *Nishantpura* regained its position with the highest reported cases,

alongside moderate numbers in *Habib Ganjh* and *Ayodhya Nagar*. In the following year, 2023, *Govindpura* witnessed the highest number of cases again, while *Gandhi Nagar* and *Khajuri* reported moderate case numbers.

Crime Prediction

The performance metrics for MSE and MAE of the used models is shown in Table 1. Classical ML models outperformed the advanced forecasting models in terms of both MSE and MAE. Random Forest exhibited the best performance with the lowest MSE of 0.019 and MAE of 0.074. Among the advanced forecasting models, ARIMA outperformed SARIMA and LSTM, achieving an MSE of 1.009 and an MAE of 0.66, respectively.

RF is better due to its ensemble nature, which combines multiple DTs to make predictions. The ensemble nature of RF helps reduce overfitting and capture complex relationships in the data. Additionally, RF is robust to noisy data, which might be beneficial given our limited data. On the other hand, advanced forecasting models, i.e., LSTM, SARIMA, and ARIMA, did not perform well due to their sensitivity to data characteristics and hyperparameters. For instance, LSTM requires large amounts of data and longer sequences to learn meaningful patterns, which our dataset might not adequately meet. Similarly, SARIMA and ARIMA models struggled to capture complex patterns in the data, particularly if the underlying crime trends were non-linear or seasonal.

Table 1: CAW Model Performance for Crime Prediction

Model	Mean Squared Error	Mean Absolute Error
DT	0.0417	0.083
RF	0.019	0.074
KNN	0.021	0.076
LSTM	7.72	2.77
SARIMA	1.21	0.69
ARIMA	1.009	0.66

7 Discussion

The confidentiality of crime-related data is paramount, leading law enforcement agencies to be cautious about sharing it with external parties. However, gaining access to this information is crucial for developing solutions to address the root causes of crime and formulating targeted actions for law enforcement. Furthermore, creating a safe and secure city for women hinges on comprehending and mitigating the escalating crime rates by understanding the crime pattern against women. By implementing crime forecasting solutions, we can foster a safer environment, a necessity that is particularly critical for the safety of women and children.

In this study, we have demonstrated the application of technology, such as clustering, to identify crime hotspot regions and forecasting models to predict

the number of CAW expected in the next or upcoming years. Our proposed algorithms offer a valuable tool for developing practical solutions to mitigate these crimes, including enhancing security measures, installing closed-circuit television (CCTV) cameras, and increasing police patrols.

7.1 Model Comparison

Our experiments conducted on the collected data indicate that classical ML regression models such as DT, RF, and KNN outperform the advanced forecasting models. This observation suggests that classical machine learning regression models may be more suitable than advanced models when dealing with a small dataset. Furthermore, it is worth noting that advanced models typically require a large amount of training data and often require high-end computing systems for training purposes.

7.2 Implications

Developing and implementing a spatiotemporal crime prediction and hotspot detection system for CAW can significantly contribute to achieving Sustainable Development Goals (SDGs³) 5⁴ and 11⁵, which prioritize gender equality and sustainable cities and communities, respectively. By accurately predicting and identifying areas susceptible to CAW, law enforcement agencies, and local authorities can take proactive measures to enhance safety and security for women, thereby advancing gender equality objectives.

By leveraging advanced technologies and data analytics, such as ML and geographic information systems (GIS) [27], the proposed system can analyze historical crime data to discern patterns, trends, and high-risk areas for CAW. This predictive capability empowers law enforcement agencies to allocate resources more efficiently, deploy patrols to high-risk areas, and implement targeted interventions to prevent crimes before they happen. Consequently, women feel safer and more empowered to engage in social and economic activities, thus advancing gender equality and upholding women's rights.

Furthermore, implementing such a system aligns with SDG 11, which aims to make cities and human settlements inclusive, safe, resilient, and sustainable. By enhancing safety and security through crime prediction and hotspot detection, cities can create environments that are conducive to the well-being and prosperity of all residents, including women and vulnerable populations. Safer cities foster community cohesion, social inclusion, and economic development, leading to more sustainable and resilient urban environments. Developing and deploying a spatiotemporal crime prediction and hotspot detection system for crimes against women contributes to achieving SDGs 5 and 11 and promotes overall societal well-being by creating safer, more inclusive, and sustainable cities for everyone.

³ <https://sdgs.un.org/goals>

⁴ <https://sdgs.un.org/goals/goal5>

⁵ <https://sdgs.un.org/goals/goal11>

8 Limitation

Our work showcased the effectiveness of forecasting models in predicting CAW in the Bhopal area with minimal error. However, our study is subject to several limitations outlined below:

1. Our analysis was based on only four years of crime data against women. Obtaining these data was challenging due to various hurdles, including the lack of public availability and the need to navigate through multiple law enforcement authorities. Moreover, the sensitive nature of crime-related data, particularly those concerning CAW, often leads to hesitancy among law enforcement agencies to share such information.
2. Although our study used Bhopal as a case study consisting of 52 police stations, we could only obtain data from 30 police stations in Bhopal. The absence of data from the remaining 22 police stations was due to the unavailability of these data from law enforcement agencies.

These limitations underscore the challenges inherent in researching CAW and highlight the need for greater accessibility to comprehensive and reliable data for more robust analyses and insights.

9 Conclusion

This study has successfully demonstrated the potential of data-driven approaches to identify crime hotspots and understand the underlying causes of crimes against women in Bhopal. By analyzing historical data and using predictive models, we can pinpoint areas where crimes are more likely to occur, helping law enforcement agencies to focus their interventions effectively. The research not only supports targeted law enforcement actions but also highlights the importance of proactive measures like increased patrols, better lighting, and CCTV surveillance. Overall, these insights contribute to safer urban environments and enhanced livability for women.

10 Future Work

Future work will focus on expanding the dataset to include more detailed data points and refining the predictive models for even greater accuracy. Additionally, collaboration with local authorities is crucial for implementing the insights gained from this study into actionable, data-driven interventions. Further exploration will include evaluating the long-term impact of these interventions on reducing crime rates and improving public safety. The research aims to create a framework that can be replicated in other cities, contributing to broader strategies for urban planning and women's safety across India.

References

1. United Nations Human Settlements Programme. *The State of the World's Cities 2004/2005: Globalization and Urban Culture*, volume 2. UN-HABITAT, 2004. 1
2. Witold Rybczynski. *Makeshift metropolis: ideas about cities*. Simon and Schuster, 2010. 1
3. Franco Cicirelli, Antonio Guerrieri, Giandomenico Spezzano, and Andrea Vinci. An edge-based platform for dynamic smart city applications. *Future Generation Computer Systems*, 76:106–118, 2017. 2
4. Hongjian Wang, Daniel Kifer, Corina Graif, and Zhenhui Li. Crime rate inference with big data. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 635–644, 2016. 2
5. Mohammad A Tayebi, Martin Ester, Uwe Glässer, and Patricia L Brantingham. Crimetracer: Activity space based crime location prediction. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, pages 472–480. IEEE, 2014. 2
6. Ncrb. <https://ncrb.gov.in/>, accessed date: 2024-03-10. 2
7. Toi. <https://timesofindia.indiatimes.com/india/india-records-51-cases-of-crime-against-women-every-hour-over-4-4-lakh-cases-in-2022-ncrb-report/articleshow/105731269.cms>, accessed date: 2024-03-10. 2
8. Al Amin Biswas and Sarnali Basak. Forecasting the trends and patterns of crime in bangladesh using machine learning model. In *2019 2nd international conference on intelligent communication and computational techniques (ICCT)*, pages 114–118. IEEE, 2019. 4
9. Gaurav Hajela, Meenu Chawla, and Akhtar Rasool. A clustering based hotspot identification approach for crime prediction. *Procedia Computer Science*, 167:1462–1470, 2020. 4
10. Sohrab Hossain, Ahmed Abtahee, Imran Kashem, Mohammed Moshiul Hoque, and Iqbal H Sarker. Crime prediction using spatio-temporal data. In *Computing Science, Communication and Security: First International Conference, COMS2 2020, Gujarat, India, March 26–27, 2020, Revised Selected Papers 1*, pages 277–289. Springer, 2020. 4
11. Akash Kumar, Aniket Verma, Gandhali Shinde, Yash Sukhdeve, and Nidhi Lal. Crime prediction using k-nearest neighboring algorithm. In *2020 International conference on emerging trends in information technology and engineering (IC-ETITE)*, pages 1–4. IEEE, 2020. 4
12. Priyanka Das and Asit Kumar Das. Behavioural analysis of crime against women using a graph based clustering approach. In *2017 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–6. IEEE, 2017. 4
13. P Tamilarasi and R Uma Rani. Diagnosis of crime rate against women using k-fold cross validation through machine learning. In *2020 fourth international conference on computing methodologies and communication (ICCMC)*, pages 1034–1038. IEEE, 2020. 4
14. S Lavanyaa and D Akila. Crime against women (caw) analysis and prediction in tamilnadu police using data mining techniques. *International Journal of Recent Technology and Engineering (IJRTE)*, 7(5C):261, 2019. 4
15. Niyonzima Ibrahim, Shuliang Wang, and Boxiang Zhao. Spatiotemporal crime hotspots analysis and crime occurrence prediction. In *Advanced Data Mining and Applications: 15th International Conference, ADMA 2019, Dalian, China, November 21–23, 2019, Proceedings 15*, pages 579–588. Springer, 2019. 4

16. Zhe Li, Tianfan Zhang, Zhi Yuan, Zhiang Wu, and Zhen Du. Spatio-temporal pattern analysis and prediction for urban crime. In *2018 Sixth International Conference on Advanced Cloud and Big Data (CBD)*, pages 177–182. IEEE, 2018. 4
17. Umair Muneer Butt, Sukumar Letchmunan, Fadratul Hafinaz Hassan, Mubashir Ali, Anees Baqir, Tieng Wei Koh, and Hafiz Husnain Raza Sherazi. Spatio-temporal crime predictions by leveraging artificial intelligence for citizens security in smart cities. *IEEE Access*, 9:47516–47529, 2021. 4
18. Fei Yi, Zhiwen Yu, Fuzhen Zhuang, Xiao Zhang, and Hui Xiong. An integrated model for crime prediction using temporal and spatial factors. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 1386–1391. IEEE, 2018. 4
19. Alexandros Belesiotis, George Papadakis, and Dimitrios Skoutas. Analyzing and predicting spatial crime distribution using crowdsourced and open data. *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, 3(4):1–31, 2018. 4
20. B Sivanagaleela and S Rajesh. Crime analysis and prediction using fuzzy c-means algorithm. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 595–599. IEEE, 2019. 4
21. Anthony J Myles, Robert N Feudale, Yang Liu, Nathaniel A Woody, and Steven D Brown. An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 18(6):275–285, 2004. 5
22. Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001. 5
23. Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. Knn model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*, pages 986–996. Springer, 2003. 5
24. Yong Zhuang, Matthew Almeida, Melissa Morabito, and Wei Ding. Crime hot spot forecasting: A recurrent model with spatial and temporal information. In *2017 IEEE International Conference on Big Knowledge (ICBK)*, pages 143–150. IEEE, 2017. 5
25. Mohammad Valipour. Long-term runoff study using sarima and arima models in the united states. *Meteorological Applications*, 22(3):592–598, 2015. 5
26. Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270, 2019. 5
27. Gis. <https://education.nationalgeographic.org/resource/geographic-information-system-gis/>, accessed date: 2024-03-10. 13