Predictive Crime Analysis and Visualization in Chicago

Sumaiya Farook, Celina Anwar, Sahana Hariharan, Sonika Tamilarasan

Abstract Crimes and acts of violence pose significant threats to the well-being of individuals and communities, necessitating innovative approaches for prediction and prevention. In our technologically advanced age, the potential for using machine learning models to gain insights into various fields has grown significantly. This paper explores the application of two predictive models—the Random Forest (RF) model and the Long Short-Term Memory (LSTM) model—focusing on their effectiveness in predicting crime in Chicago. By analyzing publicly available crime data from the City of Chicago, we aim to assess the accuracy of these models in forecasting the likelihood of various crimes based on geographic location. Additionally, we discuss the potential biases inherent in the data and consider the ethical implications of deploying such models, particularly in marginalized communities. Our findings not only provide insights into the predictive capabilities of Random Forest and LSTM models but also highlight the challenges and limitations that arise from biases in crime data. This study contributes to the ongoing discussions on the role of machine learning in public safety, emphasizing the need for careful consideration of the ethical and social consequences of crime prediction technologies.

Sumaiya Farook
University of Illinois Urbana-Champaign, e-mail: sumaiya.far123@gmail.com

Celina Anwar
University of Illinois Urbana-Champaign, e-mail: celina.anwar@gmail.com

Sahana Hariharan
University of Illinois Urbana-Champaign, e-mail: sahanahariharan@gmail.com

Sonika Tamilarasan
University of Illinois Urbana-Champaign, e-mail: sonikatam0@gmail.com

1 Introduction

Historically, Chicago has suffered from violence and high crime rates in many parts of the city. However, in recent years (as early as 2020), there has been an alarming increase in crime in Chicago, specifically violent crime. In fact, when compared with 2023 alone, violent crime increased by 11.5% from the previous year [2]. Crime can have many negative consequences, from impacting civilians and the general public to harming the local economy and businesses, ultimately making communities less livable. Crime is often random and can experience upticks at certain periods, making it difficult to model; however, if developed accurately, models such as those explored in this study have the potential to improve safety and wellness in communities.

In previous studies, crime prediction has been used in Chicago; however, those studies focused on overall crime densities and predicted changes over time [5]. There have also been efforts to predict crime using other methods, such as heat maps [4]. Recent research shows the potential for crime prediction to be a useful tool for people across cities and various other communities.

Crime prediction methods are a sought-after strategy, albeit they are quite contested. They provide those in charge of public safety, such as police, with insights into which areas should be monitored. These methods can also offer information on the uptick of certain crimes. If made accessible to the general public, they could help individuals make informed decisions about where to live, work, or visit. However, as with many machine learning models today, there are concerns about potential biases that could lead to ethical issues. Due to the over-reporting of crimes in impoverished areas or those with a majority-minority demographic, these models may make these areas appear more dangerous, while wealthier areas may seem safer by comparison.

Although many factors impact crime and the types of crime in an area, this study focused on analyzing crimes in Chicago based on their location. The City of Chicago has a public data portal that is updated daily, and this data was used for the models in this study. Given a location, we aimed to assess how accurately the RF and LSTM models could predict the likelihood of various crimes being committed, and then compare the results between the two. To de-bias our data and observe how the models may suffer from bias, we also created a geographical heat map of Chicago based on arrests made, allowing us to compare our results.

This study hopes to provide insights on reporting biases in data and how predictive crime models may be impacted from it by looking at a specific case.

2 Related Work and Background

Our project on predictive crime analysis in Chicago builds upon insights from several key studies, emphasizing the importance of localized and multi-dimensional crime prediction models.

One such study is ”Area-Specific Crime Prediction Models” [1] on IEEE Xplore. This research critiques the conventional use of global models for crime prediction, highlighting their failure to account for the heterogeneous relationship between crime and criminogenic factors across different areas. The study introduces area-specific crime prediction models that leverage hierarchical and multi-task statistical learning. These models balance detailed local predictions by sharing information across ZIP codes to mitigate data sparsity while addressing non-homogeneous crime patterns. The effectiveness of these models is demonstrated through out-of-sample testing on real crime data, showing significant predictive improvements over several state-of-the-art global models. This study underscores the importance of localized modeling in crime prediction, aligning well with our approach of focusing on specific regions within Chicago to improve predictive accuracy.

Another pivotal paper, ”Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques” [5], further supports our methodology. This research applied various machine learning algorithms, including logistic regression, support vector machine (SVM), Naïve Bayes, k-nearest neighbors (KNN), decision tree, multilayer perceptron (MLP), random forest, and eXtreme Gradient Boosting (XGBoost), along with time series analysis using long short-term memory (LSTM) and autoregressive integrated moving average (ARIMA) models. Notably, the LSTM model showed promise in predicting crime trends, with the study’s exploratory data analysis revealing important temporal patterns in crime rates. The ARIMA model’s forecast of a moderate increase in Chicago’s crime rate and the superior performance of XGBoost, with an accuracy of 94 percent for Chicago, highlight the value of diverse machine learning approaches for crime prediction.

A third study, ”Marked Point Process Hotspot Maps for Homicide and Gun Crime Prediction in Chicago” [4], explores the use of crime hotspot maps for visualizing spatial crime patterns and allocating police resources. Traditional hotspot maps often suffer from high variance in risk estimates for low-frequency crimes and fail to capture emerging trends. This study extends point process models to include leading indicator crime types through a marked point process approach, combining several years of data and various crime types to create accurate hotspot maps. By applying this methodology to a large, open-source dataset from the Chicago Police Department, the study demonstrates improved predictive accuracy for gun-related crimes and homicides. The use of kernel-based hotspot maps, which we employed in our project, provides a robust framework for visualizing crime patterns and supports the development of predictive policing strategies.

Another important study, ”Crime Prediction and Forecasting Using Machine Learning Algorithms” [6], investigates the application of several machine learning models to forecast crime in major metropolitan cities. The research focuses on predicting the severity of reported crimes and trends in crime data by year. Utilizing models such as Random Forest, K-Nearest Neighbors, AdaBoost, and Neural Networks, the study finds that the Neural Network model performs the best, achieving an accuracy of 90.77 percent. This research leverages the Chicago Police Department’s CLEAR system, analyzing over six million records to provide insights into crime trends and the effectiveness of different machine learning models. The use of data visualization tools like Folium further enhances the study's ability to illustrate crime patterns and trends, providing a comprehensive approach to crime prediction.

The final study, ”Using Machine Learning Algorithms to Analyze Crime Data” [3], employs WEKA, an open-source data mining software, to compare violent crime patterns from the Communities and Crime Unnormalized Dataset with actual crime statistical data for Mississippi. By implementing Linear Regression, Additive Regression, and Decision Stump algorithms, the study finds that the linear regression algorithm performs the best in predicting crime data. The study highlights the effectiveness and accuracy of machine learning algorithms in predicting violent crime patterns and underscores the potential of data mining in law enforcement for determining criminal ”hot spots,” creating criminal profiles, and identifying crime trends. Despite the challenges law enforcement officials face in sifting through large volumes of data, the precision in crime prediction and prevention makes the effort worthwhile for enhancing public safety.

Our project integrates these insights, employing LSTM and RandomForest models to predict crime probabilities in different regions of Chicago. By leveraging detailed crime data and advanced machine learning techniques, we aim to contribute to the development of more effective crime prediction models, ultimately supporting proactive policing strategies and improving public safety measures in Chicago.

3 Data and Methodologies

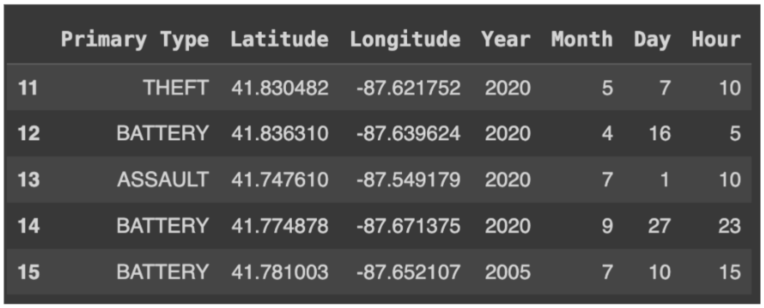
For this study, the data was obtained from the City of Chicago's data portal, specifically from the portals related to crime history and arrests. This portal is updated daily and includes data from 2001 to the present year. The dataset provides detailed information on the location and timing of crimes, along with labels for crime types. Overall, the large dataset was easily accessible and well-organized, making it the most suitable choice for our analysis.

3.0.1 Model Data Collection

For our model training, we utilized the crime dataset available from the Chicago Data Portal. This dataset includes comprehensive records of reported crimes in Chicago, providing detailed information necessary for predictive analysis. As mentioned, the dataset contains many useful data points, such as where the crime was committed (i.e., X and Y coordinates), along with timestamps and dates for each crime. The crimes are also categorized by ”Primary Type,” a label that the City of Chicago uses to identify crimes. These labels were used in this study when predicting crime type. In total, there were 22 columns.

Because of the large dataset, many rows contained null values that could not be easily filled in or estimated, so they were dropped. This was not an issue since the total number of rows exceeded 8,000,000. Prior to this, it was decided to keep only the following columns: Date, Primary Type, Latitude, Longitude, and Year.

Initially, the ’Location Description’ column was included and used to predict crime based on that information. However, when comparing the accuracy scores, it was found that longitude and latitude values were more accurate, and the ’Location Description’ column was subsequently dropped.



	Primary Type	Latitude	Longitude	Year	Month	Day	Hour
11	THEFT	41.830482	-87.621752	2020	5	7	10
12	BATTERY	41.836310	-87.639624	2020	4	16	5
13	ASSAULT	41.747610	-87.549179	2020	7	1	10
14	BATTERY	41.774878	-87.671375	2020	9	27	23
15	BATTERY	41.781003	-87.652107	2005	7	10	15

Fig. 1 Chicago data set after cleaning

3.0.2 Heat Map Data Collection

It was decided that a heat map should be added to help de-bias the data and make other observations about the model. To do so, each arrest was represented as a point with a color based on the race of the individual arrested.

To create this heat map, another dataset provided by the City of Chicago was used. This dataset had 23 columns, and most were dropped, leaving only ’Case Number’ and ’Race’. The case number was used to merge this dataset with the previously mentioned crime dataset used to train the models. The resulting dataset had 4 columns: Case Number, Race, Latitude, and Longitude. This included everything needed to create the heat map.

3.1 Methods

Python 3 was used as the programming language for developing the back end of both models in this study. The Random Forest (RF) model utilized the scikit-learn library, specifically the random forest classifier. To evaluate the model’s performance, we employed metrics from the same library, including the classification report and accuracy score. For the Long Short-Term Memory (LSTM) model, TensorFlow was used, along with additional performance metrics from the scikit-learn library.

3.1.1 Random Forest Model

The Random Forest model is an ensemble learning method primarily used for classification tasks. It operates by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes (classification) of the individual trees. This model helps reduce overfitting and improve accuracy. In this study, the RF model was trained on 67% of the dataset, with the remaining 33% reserved for testing. Two hundred decision trees were used to train the model, and it was found that setting the maximum number of nodes to 25 resulted in higher accuracy. Given the large size of the dataset, significant training time issues were encountered. To address this, the training process was divided into smaller parts, which alleviated most of the runtime problems. This approach ensured efficient model training without compromising the integrity of the data or the accuracy of the model.

3.1.2 Long Short-Term Model

The Long Short-Term Memory (LSTM) model is a type of recurrent neural network (RNN) capable of learning long-term dependencies. LSTMs are particularly effective for time series prediction and natural language processing tasks due to their ability to retain information over long periods. In this study, the LSTM model was implemented using TensorFlow. The dataset was split into 80% for training and 20% for testing. Initially, the model experienced extended training times, which were mitigated by increasing the batch size to 128 and reducing the number of epochs to 10. These adjustments significantly improved training efficiency while maintaining the model's performance and accuracy.

4 Results

The results of our predictive crime analysis in Chicago were derived from two primary models: the Random Forest (RF) model and the Long Short-Term Memory (LSTM) model. These models were evaluated based on their accuracy in predicting the likelihood of various crimes in different regions of Chicago. Below, we present a summary of our current findings for each model, which we plan to improve on.

4.0.1 Random Forest Model Results

The Random Forest model was trained on 67 percent of the dataset and tested on the remaining 33 percent. Despite the large dataset size and training time issues, the RF model provided useful insights into crime prediction. The following are the key results:

- **Accuracy:** The RF model achieved an overall accuracy of approximately 17 percent in predicting the type of crime based on location data (latitude and longitude). This result indicates that while the model captured some patterns, there is significant room for improvement in predictive performance.
- **Feature Importance:** The most important features for the RF model were latitude, longitude, and primary crime type. These features played a crucial role in determining the model's predictions.

4.0.2 LSTM Model Results

The LSTM model was trained on 80 percent of the dataset and tested on the remaining 20 percent. Adjustments to batch size and the number of epochs were made to optimize training efficiency. The key findings from the LSTM model are as follows:

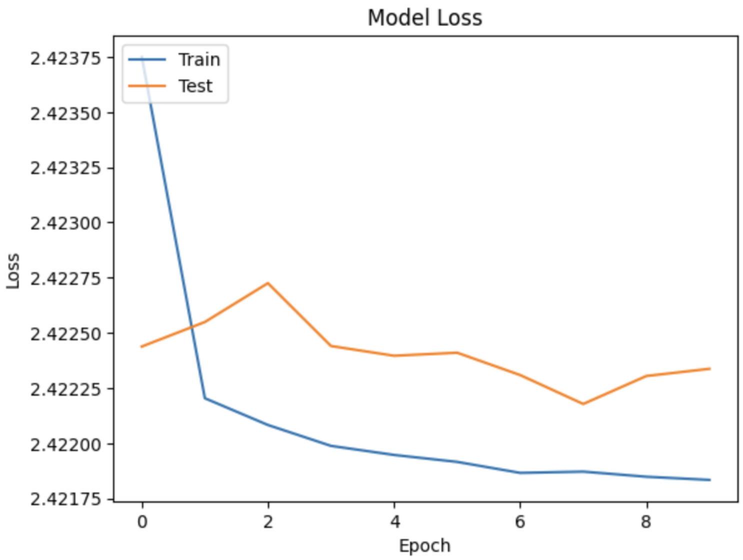


Fig. 2 LSTM Loss Curve

- **Accuracy:** The LSTM model achieved an overall accuracy of approximately 21 percent in predicting the likelihood of various crimes based on location data. This performance, while higher than that of the RF model, still highlights the challenges in our project.
- **Training Efficiency:** By increasing the batch size to 128 and reducing the number of epochs to 10, training efficiency improved significantly without compromising the model's performance.

4.0.3 Comparative Analysis

Comparing the RF and LSTM models, the LSTM model demonstrated a higher overall accuracy in predicting crime types. This suggests that the LSTM's ability to handle sequential data provides an advantage in this context. However, both models exhibited relatively low accuracy, indicating the inherent complexity and variability in crime data, as well as possible inaccuracies in our project.

4.0.4 Summary

The results of our study underscore the challenges in predictive crime analysis, particularly the need for more sophisticated models and better data quality. While the LSTM model showed promise with higher accuracy, both models require further refinement. Additionally, the visualization of arrest data highlighted critical issues of racial disparity, which must be addressed to ensure the ethical and effective use of predictive policing technologies.

Future work will focus on improving model accuracy, incorporating additional features, and addressing ethical concerns to enhance the overall utility and fairness of predictive crime analysis tools.

5 Conclusion

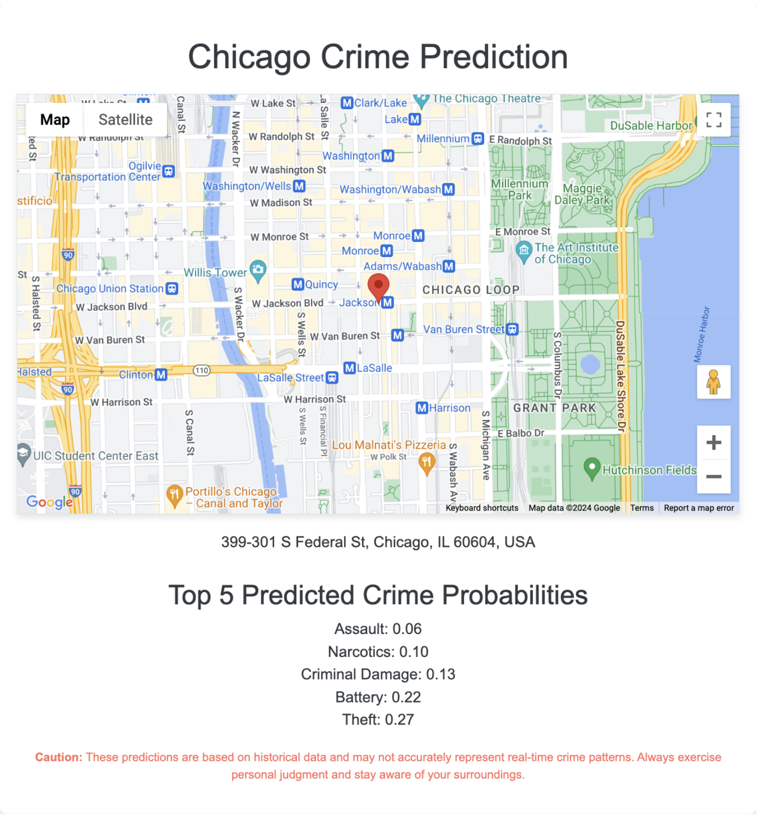
5.1 Applications

This section details the practical use of our models, including an interactive website that provides real-time crime probabilities and a heat map visualization of arrest data by race. We also discuss the accuracy and ethical considerations of these predictive tools.

5.1.1 Building the Interactive Website

To enhance the accessibility and practical application of our predictive crime analysis, we have developed an interactive website. This platform provides users with a user-friendly interface to automatically receive real-time probabilities of various crimes occurring nearby. The website's structure is designed to integrate seamlessly with the predictive capabilities of our Long Short-Term Memory (LSTM) model, which has been trained on extensive crime data from Chicago.

Upon accessing the website, the user's location is automatically detected using the Google Maps API. The backend system, powered by our trained LSTM model, processes this location data and returns the probabilities of different types of crimes that might occur in that vicinity. This setup allows residents and visitors to make informed decisions about their safety based on up-to-date, data-driven insights.



Chicago Crime Prediction

399-301 S Federal St, Chicago, IL 60604, USA

Top 5 Predicted Crime Probabilities

Assault: 0.06

Narcotics: 0.10

Criminal Damage: 0.13

Battery: 0.22

Theft: 0.27

Caution: These predictions are based on historical data and may not accurately represent real-time crime patterns. Always exercise personal judgment and stay aware of your surroundings.

Fig. 3 A plot map depicting arrest data by race

5.1.2 Heat Map Visualization of Arrest Data by Race

In addition to crime prediction, our website features a heat map that visualizes arrest data across different racial groups in Chicago. This heat map is constructed using data from the City of Chicago's public data portal, specifically focusing on the racial demographics of arrests. Each arrest is plotted on the map, with color coding representing different races, providing a clear visual representation of racial patterns in arrests across the city.

The inclusion of this heat map serves multiple purposes. Firstly, it offers a deeper understanding of the geographic distribution of arrests and potential racial biases. Secondly, it raises awareness about disparities in the criminal justice system, encouraging users to consider the broader social and ethical implications of predictive policing technologies.

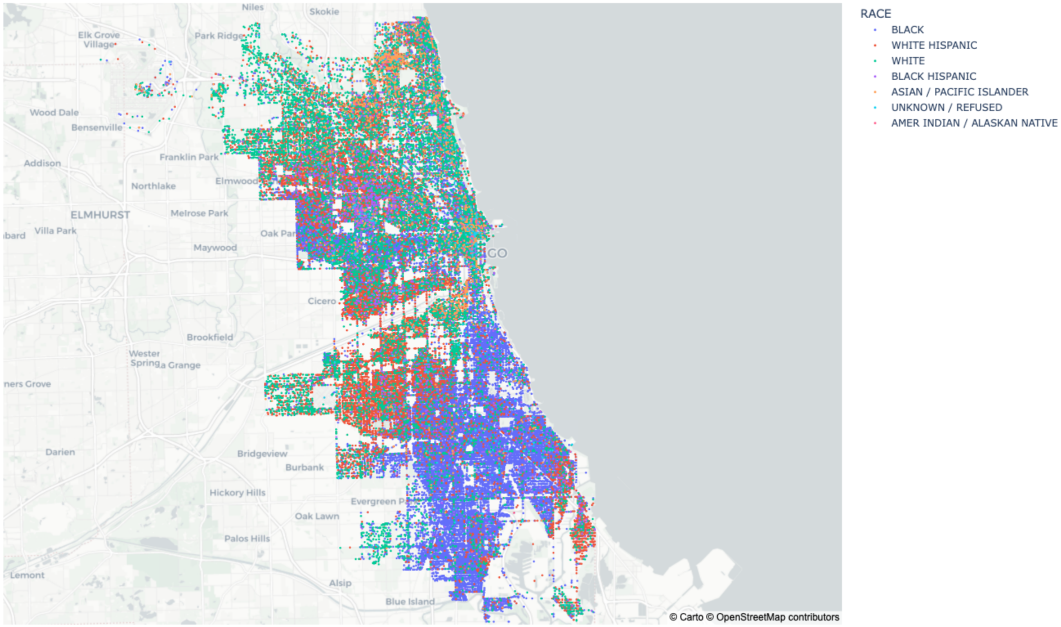


Fig. 4 A plot map depicting arrest data by race

- **Geographic Distribution:** TThe heat map revealed significant variations in arrest rates across different regions of Chicago, with certain areas showing higher concentrations of arrests.
- **Racial Disparities:** The visualization highlighted racial disparities in arrest data, with minority-majority areas experiencing higher arrest rates. This finding underscores the importance of considering socio-economic and demographic factors in crime prediction and policing strategies.

5.1.3 Awareness and Accuracy Considerations

While our website strives to provide accurate and useful information, it is important to acknowledge that predictive models and visualizations may not always be perfectly accurate. Factors such as data quality, reporting biases, and model limitations can affect the accuracy of predictions and the representation of arrest data. Users are advised to use the information provided as a supplementary tool rather than a definitive guide for personal safety decisions.

By combining predictive analytics with insightful visualizations, our interactive website aims to empower users with knowledge and foster a deeper understanding of crime patterns and systemic issues in Chicago. However, it is crucial to approach the data with a critical eye and remain aware of the inherent limitations and potential biases in the models and data sources used.

5.2 Concerns and Next Steps

While our website aims to enhance safety and provide actionable insights, several concerns and considerations must be addressed:

5.2.1 Concerns

- **Data quality and completeness:** The accuracy of crime predictions and visualizations is contingent on the quality and completeness of the data. Missing or inaccurate data points can lead to false predictions and misrepresentations. The reliance on historical crime data and arrest records may not fully account for underreported or unreported crimes, potentially skewing the results.
- **Model limitations:** Both the Random Forest (RF) and Long-Short Term Memory (LSTM) models have inherent limitations. RF models, while effective for classification tasks, may struggle with highly imbalanced datasets or complex interactions between features. LSTM models, though powerful in capturing temporal patterns, can be sensitive to hyperparameter choices and training duration, which may affect performance.
- **Ethical and privacy concerns:** Predictive crime models and arrest visualizations raise ethical concerns regarding privacy and the potential for misuse. The dissemination of detailed crime predictions and arrest data may unintentionally stigmatize neighborhoods or individuals. Ensuring that the technology is used responsibly and ethically is paramount to avoid exacerbating social inequalities.
- **Model updates and maintenance:** Predictive models need regular updates to incorporate new data and adapt to changing crime patterns. Without continual refinement, the models may become outdated and less effective over time. Maintenance of the models and the underlying datasets is essential for sustained accuracy and relevance.

5.2.2 Next Steps

- **Enhancing data quality:** Future research should focus on improving data quality by addressing missing values, correcting inaccuracies, and incorporating a wider

range of data sources. Collaborating with local law enforcement and community organizations could provide more comprehensive datasets and insights into underreported crimes.
- **Refining models:** Continuous refinement of the RF and LSTM models is necessary to improve predictive accuracy. Experimenting with additional machine learning algorithms, such as gradient boosting methods or neural networks with attention mechanisms, could offer better performance. Incorporating feature engineering techniques and exploring advanced time-series methods might enhance model robustness.
- **Addressing biases:** Implementing fairness-aware algorithms and techniques can help mitigate biases in crime predictions and arrest visualizations. Conducting fairness audits and incorporating feedback from affected communities can guide improvements and ensure that the models do not perpetuate existing inequalities.
- **Expanding visualizations:** To provide a more nuanced understanding of crime patterns, future work should expand visualizations to include various dimensions, such as socio-economic factors and community resources. Integrating additional data layers, such as neighborhood development and local initiatives, could offer a more holistic view of crime dynamics.
- **Promoting transparency and education:** Promoting transparency about the limitations and potential biases of the models is essential for responsible use. Providing educational resources and clear explanations of the predictive outputs can help users make informed decisions and understand the broader context of crime data.

By addressing these concerns and pursuing these next steps, the project aims to improve the accuracy, fairness, and utility of predictive crime analysis and visualization tools, ultimately contributing to enhanced public safety and informed decision-making.

Acknowledgements Thank you to Christof Teuscher and the rest of the altREU team for providing feedback and the resources to pursue this project.

References

1. M. A. Boni and M. S. Gerber. Area-specific crime prediction models. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 671–676, Anaheim, CA, USA, 2016.
2. Illinois Policy Institute. Violent crime up 18%, arrests down 43% in chicago over 10 years. https://www.illinoispolicy.org/press-releases/violent-crime-up-18-arrests-down-43-in-chicago-over-10-years/#: :text=years%20%7C%20Illinois%20Policy., 2024. Accessed: 2024-07-24.
3. Lawrence McClendon and Natarajan Meghanathan. Using machine learning algorithms to analyze crime data. *MLAIJ*, 2, 2015.
4. George Mohler. Marked point process hotspot maps for homicide and gun crime prediction in chicago. *International Journal of Forecasting*, 30:491–497, 2014.

5. Wajiha Safat, Sohail Asghar, and Saira Andleeb Gillani. Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques. *IEEE Access*, 9:70080–70094, 2021.
6. Azwad Tamir, Eric Watson, Brandon Willett, Qutaiba Hasan, and Jiann-Shiun Yuan. Crime prediction and forecasting using machine learning algorithms. *IJCSIT*, 12:26–33, 2021.