SafeTrace: Visualization and Analysis of Chicago Crime Data

Team 79: Lucia Etchecopar, Regina Kang, Jeremy Warren, Tarun Khosla, Chinmay Pathare

November 29, 2023

1 Introduction

The internet age has sparked increased global awareness, sometimes overshadowing the significance of local issues that directly affect individuals. Paradoxically, people prioritize matters closer to home, which drives demand for information regarding their local areas. To address this, platforms like Zillow provide local housing information. Similarly, Nextdoor, Facebook groups, and Reddit forums facilitate community discussions, though user engagement may vary. Crime concerns are universal, but crime statistics often lack local context, causing unwarranted worry. This is more critical for newcomers who lack neighborhood familiarity, highlighting the need for accessible and localized crime data.

2 Problem Definition

Crime is a major issue impacting cities all around the world, including Chicago. For people who are planning on moving to Chicago, starting a business, etc., crime is likely an important factor in making a decision. While there is crime data scattered around the internet, there is no platform from which a user can easily and effectively find reliable crime statistics and crime predictions in one place. They are usually forced to look at multiple different sources, many of which have conflicting information. Our project aims to tackle this issue along with addressing the lack of interactivity in how crime data is presented.

In this project, we construct an interactive platform where users can use crime type and date filters to find the specific crime information they are looking for. These filters control a map which can be zoomed into and dragged to focus on a specific area. The platform also includes functionality in which users can input two addresses and compare the crime counts for the crimes types and dates chosen. The platform also forecasts crime scores at those addresses. These crime scores are a novel approach at categorizing crime with more detail below. We test multiple forecasting models and implement the best model amongst the ones we tested. Using machine learning and visual analytics, as well as simple, functional, interactive features, we create a project that takes advantage of publicly available crime data. With this project, we bridge the gap between analytic output and usability, and provide a platform that gives users a holistic view of crime in Chicago.

Overall, given the need for reliable, accessible, and interactive crime data for decision-making in Chicago, we developed a platform that aggregates this data, allows for user interaction, and provides crime predictions. The inputs are publicly available crime data, user-selected filters, and user-inputted addresses. The outputs are a map displaying crime data according to the filters, a comparison of crime counts between two addresses, a historical crime score at these addresses, and a forecast of crime scores at these addresses. We also focused on ensuring that the platform is user-friendly, that the data is reliable, and that the crime predictions are accurate.

3 Literature Survey

Numerous research papers, which focus on the visualization and analysis of crime data, have significantly enriched our project. Zhang et al. [15] delve into a study that investigates the intricate connection between crime patterns and various factors, underscoring the limitations of conventional regression models in crime prediction. These models often make the assumption that crime events are spatially and temporally clustered. Similarly, Wang et al. [14] employ a variety of models, including regression-kriging and geographically weighted negative binomial models, to forecast crimes in Chicago. They explore the impact of variables like point-of-interest and taxi-flow data, revealing results that vary depending on the region and type of crime. In contrast to these methodologies, our project adopts a distinct path by excluding such data points to eliminate noise, focusing on the development of crime score forecasting models by postcode.

Our research led us to various articles that have shed light on the work and models implemented in this domain. Tayebi et al.'s Crime Tracer model [12] employs crime locations and personalized random walk models

to predict crime based on offenders' activity spaces. While this model aligns with our objectives, it presents scalability and privacy challenges as it requires unique offender identifiers. Our project steers clear of such identifiers to avoid these concerns. Greenberg's paper [7] provides methodological insights into time series analyses of crime rates, despite potential limitations in focus. Its publication date, 2001, is also a cause for some concern. Kim et al. [8] analyze crime data in Vancouver over a 15-year period, employing machine learning predictive models such as K-nearest-neighbor and boosted decision trees. While their methods are highly relevant, disparities in geography and demographics between Vancouver and Chicago made us conclude that their models did not have direct applicability to our project.

Varshitha et al. [13] explore prediction using deep learning, crime casting, and sentiment analysis, highlighting the advantages and disadvantages of each approach. This emphasizes the importance of testing various models for crime prediction, a principle we adopted as well. Almaw and Kadam [1] use ensemble methods and compare their performance against conventional classification models, concluding that ensemble methods offer enhanced predictability. This suggested the potential benefit of testing ensemble methods when creating predictive models.

Dakalbab et al.'s literature review [3] outlines the distinctions between supervised and unsupervised learning approaches in evaluating crime data, noting that supervised learning is the more applied method in crime analysis. Safat et al. [11] provide insights into relevant techniques for our dataset through the exploration of common machine learning techniques for forecasting crime trends. We found this paper to be relevant to our project despite their focus on the broader Chicago metropolitan area. Nitta et al. [10] employ simpler predictive techniques while incorporating latitudinal and longitudinal data for precise geographic modeling, a strategy we also explore. Feng et al. [5] present a study aimed at identifying crime patterns through visualization and the exploration of time series models. While many of their visualizations were ineffective, in our project, we created clear, useful, and interactive visualizations. Visual Informatics Vol. 5 [2] emphasizes the significance of understanding relationships among data components and how visual analytics aids in constructing a coherent mental model.

In our pursuit of a visual approach, we investigated several articles for practical application. Deng et al. [4] provide a comprehensive review of urban visual analytics, emphasizing the integration of visualization techniques and computational models. While it lays a foundational understanding, it lacks a specific focus on crime data analysis, which presented an opportunity for our project to delve deeper into crime-specific visual analytics. George Mohler's paper [9] utilizes hotspot maps for crime patterns, aligning with our project's goals of simplicity and effectiveness. Garcia-Zanabria et al. [6] present Mirante, a crime mapping visualization system that uses spatiotemporal analysis to highlight crime patterns on a street level. While this tool could have been useful for our analysis, we decided not to use it, as it is not publicly accessible.

In conclusion, these diverse literature sources provided a wide array of perspectives and methodologies to our project, guiding our approach in creating an effective and user-friendly crime data visualization platform. By addressing the limitations observed in existing studies, our project provides interactive visualizations that offer a holistic view of crime in Chicago, coupled with a crime score index forecast that provides insights into future crime in Chicago.

4 Proposed Method

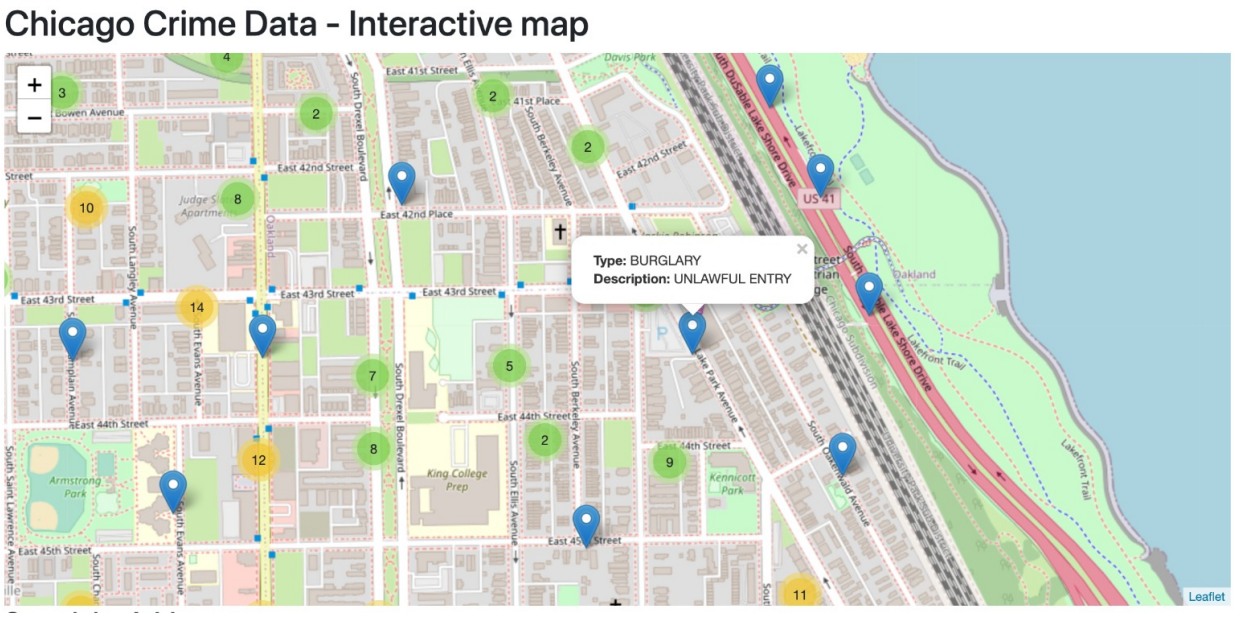
Our product, a web-based application, offers a variety of features. At its core, it allows users to compare crime between different locations within Chicago, using our extensive data set. Users can also employ filters by crime type and date to refine their search.

To deliver this platform, we initiated the data collection process by obtaining data from the City of Chicago. This data set includes crime records dating back to 2001 (with the exception of murder data), contains seven millions records, and is updated weekly. Those records have latitude and longitude data to indicate the reported location of the crime. We improved the data set with postcode and population for each record from 2022 to the present. To get the postcode, we used Nominatim's geocoder for OpenStreetMap data, and we were only able to populate 20 months of data because of the limitations of the public API. To add the population, we used data from the latest census and joined that in our data set.

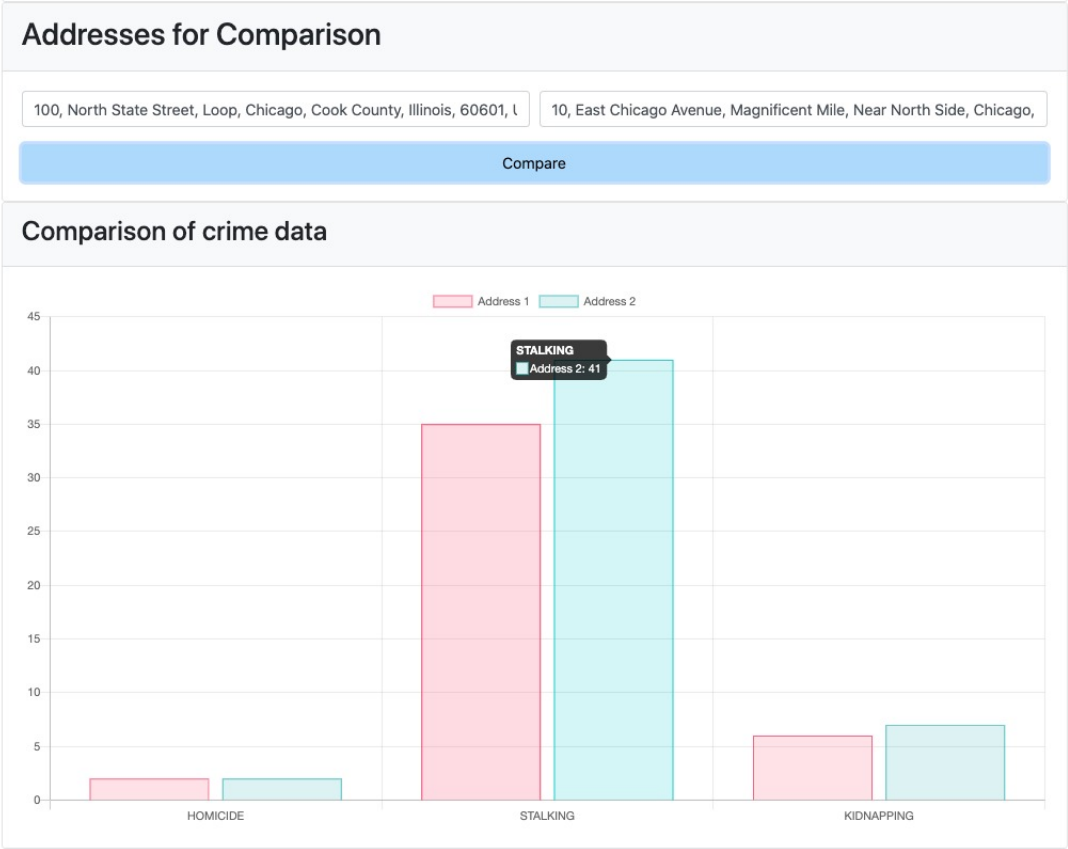
The initial data cleaning is performed using the Pandas library in Python. Then, the data is loaded into our SQL Alchemy database. This database is connected to our Flask front end, enabling users to access the platform through their web browsers.

When accessing this application there are different things users can do. The initial functionality is that users can select different crime types and a period of time to get a filtered view of crime. The information is presented in a map, and if two addresses are inputted, it is also presented in a grouped bar chart that provides a historical overview of the filtered crime around the inputted addresses. Users are also able to obtain the historical crime score of each address and forecast the crime score of each address for the next six months. In the following paragraphs, we will discuss each of these innovations in more detail.

Our first innovation introduces an interactive map that combines zooming, dragging, and filtering capabilities, enabling users to easily access the specific data they seek. While existing online crime maps of Chicago offer zooming features, our platform stands out by providing advanced filtering options and a single point of contact for all crime concerns. This allows users to refine their data search effectively and save time with a singular web page rather than scouring multiple news and police sites.



Our next innovation gives users the ability to input two addresses and receive information on the selected crimes around those locations, facilitating side-by-side comparisons of crime rates. This feature is particularly valuable for individuals interested in assessing crime differences between two addresses. Currently, this task is cumbersome as people must look up each address separately and rely on memory or written notes for comparisons, given that crime statistics are typically presented by neighborhood, town, or city, rather than by individual addresses. Our platform streamlines and enhances this process, offering convenient and precise address-based comparisons.



The crime score and its forecasts are our final innovations that we believe helps users gain a better understanding of crime in a given area. Currently users are bombarded with crime reports that present confusing and sometimes even conflicting views of crime in specific areas. For example, a person interested in the 60611 postcode in Chicago may be presented with statistics stating car thefts are up 37% while assaults are down 63%. Does this mean the area is safer or more dangerous than it used to be? Is crime going up or down? The crime score is intended to give users a single number to associate with crime. Philosophically, we decided to rank crime by how bad the crime is. Determining badness can be inferred from the punishment. We used Cook County sentencing data from the Cook County State's Attorney Office to get sentences for crimes. We also chose to only use the most serious charge a person was convicted with in terms of sentencing. Cleaning this data involved normalizing the sentence in terms of days, ensuring that only guilty verdicts were included, and matching the crimes from our multiple data sets. Our method of scoring crimes based on sentencing data is an innovative technique that is unique to our platform and has not been used in the past.

The Attorney's Office and the Chicago Police use different categorizations for crime. The police use IUCR codes such that each crime has a unique code. Initially we tried and failed to do a fuzzy match with the 'Disposition Charged Offense Title' column in the sentencing data set with the 'Primary Type' column in the crime data set. Fuzzy matching produced too many errors and 1443 missing values but other NLP techniques may be used here in the future. We then used another data set to manually map these two columns. Once these categories were mapped, we averaged the sentence length for each crime to determine how bad each crime was. Then, aggregating this score within each postcode and dividing by the population in each postcode gives us a crime score that allows us to much more easily predict crime and determine the safety of each postcode. In our GUI, when users input the addresses that they want to compare, our application determines the postcodes from those addresses, and returns the historical crime scores and crime score forecasts by postcode.

	December 2023	January 2024	February 2024	March 2024	April 2024	May 2024
Address 1	8.39	12.03	11.99	12.17	12.02	11.88
Address 2	15.55	23.68	22.98	24.59	24.38	21.83

In summary, our platform is better than state of the art because it offers an innovative and user-friendly way to access relevant crime data for informed decision-making, while incorporating a novel way of categorizing and interpreting crime in an area. Crime data is available from various sources so we consolidated the most pertinent information to assist individuals in making decisions based on crime-related factors. Through this application, we created a simple and usable link between the public, the police, and the attorney's office.

5 Experiments and Evaluation

Our experiments are structured to address several key questions. First, we seek to determine the areas where crimes are most concentrated. Second, we aim to establish the timing of these criminal occurrences. Third, we investigate the specific types of crimes that are prevalent in different locations. Furthermore, we examine the crime statistics surrounding particular addresses to provide users with relevant insights. Lastly, we experiment with predicting crime types and forecasting the expected level of crime by postcode, adding a predictive dimension to our analysis.

To address the first four questions, our approach involved experimenting with various visualizations designed to be user-friendly and intuitive. We conducted rudimentary A/B testing within our team to determine the optimal placement and integration of these visualizations with filters. As a result, we designed a single-page platform featuring an interactive map and a linked bar chart, controlled by a unified set of filters (but the linked bar chart has an additional input for users to specify the addresses they want to compare). This integration enhances the platform's cohesion, power, and user-friendliness.

The interactive map allows users to zoom in and out, drag, and apply filters based on crime type and date. This empowers users to explore specific data tailored to their needs, providing a comprehensive solution to the first four questions. The linked bar chart facilitates a side-by-side comparison of the number of crimes between two inputted addresses. Crimes within a 0.02-degree square area, equivalent to approximately a 1.2-mile square area under the assumption of a flat Earth, centered on each address, are counted and visualized. As the filters are unified, the crime type and date filters applied to the map also apply to the bar chart. These combined features deliver effective answers to the initial questions regarding crime concentration, timing, crime types, and localized crime statistics.

To address the fifth question involving forecasts, we integrated crime scores and crime score forecasts under the bar chart. While we agreed that the ideal implementation would involve forecasts for the 1.2-mile square areas around the addresses inputted, the computational demands of this approach were too significant. Given the constraints of our available processing power, we decided to generate forecasts by postcode. To facilitate this, we are using an API to extract postcodes based on the inputted addresses, allowing us to utilize the same address filter for forecasts.

While these predictions were not implemented, our initial testing involved decision tree and random forest classification models for predicting crime types, yielding accuracy rates of approximately 25.4% and 25.8%, respectively. Decision tree models, while interpretable, often suffer from accuracy issues and overfitting. While random forest classification models typically offer better accuracy, our results showed only marginal improvement.

As the previously obtained accuracy scores were extremely low, we decided that we could not use those models. To predict something as serious as crime with a model that has low accuracy can end up causing tremendous problems for those relying on our platform for making decisions. This is why we came up with the idea to forecast a crime score (the calculation for which we discussed in the 'Proposed Method' section) for each postcode. As we mentioned before, the final crime score is calculated by taking the sum of the crime scores for each data point within a postcode, divided by the population, which is sourced from US census data.

In order to find a better model, given the nature of the data, we first decided to experiment with time series models like ARIMA, SARIMA, and triple exponential smoothing. However, we found that in order to forecast with these models, we would have to create a different model for each postcode since each postcode would have different trends and behaviors. This would have created the need to make 1400 different models, which was not suitable for our project. In addition, if we decided to scale this project to different cities, the number of models needed would continue to multiply, which further solidified our determination that the use of any of these models was unsuitable. Therefore, we decided to explore alternative options.

We shifted our approach to explore machine learning models that can handle time series data in the context of regression. We tested a random forest regressor model, a gradient boosting machines model (XGBoost), and a neural networks model. However, we found that we did not have the computational resources to create a good neural networks model, so we eliminated this model early on. To compare performance between the random forest regressor model and the gradient boosting machines model, we conducted rigorous testing. As mentioned earlier, the main Chicago crime data set has around seven millions records, each representing a crime incident, with key features including date, crime type, and location. We joined this data with the population data set, which has around 1,400 records, each representing a postcode, with population being the key feature. Then, we also joined the sentencing data set, which has around 300,000 records, with key features disposition charged offense title and commitment term. Once all the data was joined, we split the data 80-20, with 80% of the data allocated to the training set and 20% allocated to the test set. In order to evaluate the models, we used multiple goodness-of-fit metrics such as mean absolute error, root mean squared error, and R-squared. This holistic approach to evaluation helped us in understanding different aspects of each model's performance (like error magnitude and variance explained). Ultimately, the best gradient boosting machines model was outperformed by the best random forest regressor model, which had the following goodness-of-fit results: mean absolute error = 3.49608153392831, root mean squared error = 9.2550046978414, R-squared = 0.9221434305350965 (please note that we get these goodness-of-fit values when predicting the next month of unknown data - as we move farther into the future, the predictions naturally become less accurate). Therefore, the final model we chose to deploy was the random forest regressor model.

In order to achieve this level of goodness-of-fit, we took the following steps. We normalized crime scores by population, which provided a more accurate reflection of the crime impact in different postcodes. Then, we incorporated lag features (past values of the crime scores) into the model. This allows the model to consider historical data, which is often predictive of future trends in crime rates. We also included temporal features, which means that the model includes the month and year as features, which can capture seasonal and yearly trends in crime rates. This temporal aspect is crucial in forecasting models. We also made sure to use the most recent 20% of the data as the test set. In doing this, we are effectively evaluating the model on the most recent data. This is a realistic approach for time series forecasting and can provide a more accurate assessment of the model's predictive power. Finally, we performed hyperparameter tuning using GridSearchCV in order to further optimize the model.

After all of this experimenting and tuning, our final random forest regressor model was integrated into our platform, enabling users to assess whether crime (via crime score) is expected to increase or decrease over time and by how much. For all of these models and tests, we used Python as the programming language.

6 Conclusion and Discussion

In summary, our platform streamlines the process of accessing crucial crime data for informed decision-making. It features an interactive map, comparative bar chart, novel crime index, and crime score forecasts, all managed by a single set of filters. Users can effortlessly gather essential crime information and compare between two addresses with a few clicks. This platform addresses the time-consuming and often conflicting manual searches individuals undertake when moving to, traveling to, or starting a business in Chicago. While currently limited to Chicago, it has the potential to scale to include cities worldwide with publicly accessible crime data, revolutionizing the way people seek crime insights globally.

We were successful in implementing new techniques for helping citizens access crime data to help them make more informed decisions around crime. However, there are a few things we could add onto this comprehensive application. We could have improved the UI/UX design of this application to be more enticing to users. We could have also incorporated a comment section that would allow more information be added to each crime. For example, if a shooting occurred users could report about what time the shots happened and how many there were. This would help police officers with information gathering as they are often at the mercy of one eyewitness report. Along these same lines, an anonymous reporting section could help with crime reporting as some individuals are afraid to report suspicious activity directly to the police. Another feature that we could have implemented in production would be a decision tree for users to avoid crime. With this, we could tell users to take certain streets or go out at certain times depending on the area in which they live.

The crime index is not perfect and certainly is biased. For example, since we only associated a punishment with the most severe crime a person was convicted of, there are cases where an individual with 15 minor counts would be weighted the same as a person with no additional charges. We believe this bias to be small as our data set was quite large and crime indexes were calculated using a weighted average. Also when combining our sentencing and crime data, we were not able to perfectly match all of the crime types. There may have been more we could have done to alleviate this but we were surprised to see such a mismatch between police and district attorney. If we were to extend this application in the future, we would call for federal legislation that would normalize crime codes, reporting, and dissemination. This would allow for a much smoother and likely more accurate joining of our sentencing and crime data, which would thus improve the accuracy and value of our application.

All team members have contributed a similar amount of effort.

References

[1] Ayisheshim Almaw and Kalyani Kadam. Crime data analysis and prediction using ensemble learning. In *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1918–1923. IEEE, 2018.

[2] Natalia Andrienko, Gennady Andrienko, Silvia Miksch, Heidrun Schumann, and Stefan Wrobel. A theoretical model for pattern discovery in visual analytics. *Visual Informatics*, 5(1):23–42, 2021.

[3] Fatima Dakalbab, Manar Abu Talib, Omnia Abu Waraga, Ali Bou Nassif, Sohail Abbas, and Qassim Nasir. Artificial intelligence & crime prediction: A systematic literature review. *Social Sciences & Humanities Open*, 6(1):100342, 2022.

[4] Zikun Deng, Di Weng, Shuhan Liu, Yuan Tian, Mingliang Xu, and Yingcai Wu. A survey of urban visual analytics: Advances and future directions. *Computational Visual Media*, 9(1):3–39, 2023.

[5] Mingchen Feng, Jiangbin Zheng, Jinchang Ren, Amir Hussain, Xiuxiu Li, Yue Xi, and Qiaoyuan Liu. Big data analytics and mining for effective visualization and trends forecasting of crime data. *IEEE Access*, 7:106111–106123, 2019.

[6] Germain Garcia-Zanabria, Erick Gomez-Nieto, Jaqueline Silveira, Jorge Poco, Marcelo Nery, Sergio Adorno, and Luis G Nonato. Mirante: A visualization tool for analyzing urban crimes. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 148–155. IEEE, 2020.

[7] David F Greenberg. Time series analysis of crime rates. *Journal of quantitative criminology*, 17:291–327, 2001.

[8] Suhong Kim, Param Joshi, Parminder Singh Kalsi, and Pooya Taheri. Crime analysis through machine learning. In *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 415–420. IEEE, 2018.

[9] George Mohler. Marked point process hotspot maps for homicide and gun crime prediction in chicago. *International Journal of Forecasting*, 30(3):491–497, 2014.

[10] Gnaneswara Rao Nitta, B Yogeshwara Rao, T Sravani, N Ramakrishiah, and Muthu Balaanand. Lasso-based feature selection and naïve bayes classifier for crime prediction and its type. *Service Oriented Computing and Applications*, 13:187–197, 2019.

[11] 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.

[12] 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.

[13] DN Varshitha, KP Vidyashree, P Aishwarya, TS Janya, Dhananjay Gupta KR, and R Sahana. Paper on different approaches for crime prediction system. *International Journal of Engineering Research and Technology (IJERT)*, 2017.

[14] Hongjian Wang, Huaxiu Yao, Daniel Kifer, Corina Graif, and Zhenhui Li. Non-stationary model for crime rate inference using modern urban data. *IEEE transactions on big data*, 5(2):180–194, 2017.

[15] Xu Zhang, Lin Liu, Minxuan Lan, Guangwen Song, Luzi Xiao, and Jianguo Chen. Interpretable machine learning models for crime prediction. *Computers, Environment and Urban Systems*, 94:101789, 2022.