Minority Report: Machine Learning Classification Applied to Crime in Los Angeles

Bryan Cortes
Computer Science
California State University Fullerton
Fullerton, USA
cortes28@csu.fullerton.edu

Alexander Golubow
Computer Science
California State University Fullerton
Fullerton, USA
agolubow@csu.fullerton.edu

Genna Quach
Computer Science
California State University Fullerton
Fullerton, USA
gennaquach@csu.fullerton.edu

Kanika Sood
Computer Science
California State University Fullerton
Fullerton, USA
kasood@fullerton.edu

***Abstract*—Criminal activities in the City of Los Angeles are increasing at a very alarming pace. To assist the police in identifying crucial cases and bringing awareness to the severity of it, Machine learning is a suggested method. This document will go over the machine learning algorithms and techniques that are used to classify crime as violent or non-violent. As opposed to other works which may predict the chances of a crime happening, this algorithm will classify the prediction to bring more awareness to the types of crime being committed; this information will allow the police and citizens to have a better understanding of the nature and severity of criminal activities in an area.**

***Index Terms*—crime, violent, classification, machine learning, severity, awareness**

I. Introduction

Machine learning allows for a system to learn and improve as time goes by when trained with an increasing amount of patterns and data. This would be a useful tool in creating a system that would help society be aware of the crime rate. There are numerous crimes that can occur any day at any time; making it near impossible to prevent every single crime from occurring. Rather than strive for the impossible, the algorithm we are suggesting will be able to pinpoint the possibility of a crime being violent or not, allowing citizens to be more aware of the risks. The City of Los Angeles has a violent crime rate of 445 crimes per 100,000 people, higher than the national rate in 2020 [1]. In an attempt to decrease these rates, we suggest utilizing machine learning to pinpoint violent crimes before they happen. The main aim of this system is to predict how severe the outcome of a crime is; as it may potentially save a life of a victim, preventing them from a life of trauma, and to prevent damages to private property [2]. In this paper, we strive to make changes by comparing the results of multiple machine learning algorithms that we have used to predict the nature of a crime occurring, taking into account the area, the latitude and longitude, the victim's age and sex, and the time of day. Our results indicate that it is in fact possible to predict the severity of a crime with relatively accurate predictions without the need for a substantial amount of features.

A. Existing Works

There are various ways to analyze crime and many approaches to determining patterns from data sources. Some approach it by using Geospatial, a map-based visualization, or Geographical clustering of crime activities to find hot spots [7]. This would be used to get information on where crimes are likely to occur. Most often to get the crime hot spots, the method used would be the K-Nearest Neighbor. This machine learning technique would allow us to get a better understanding of the crime density that would normally not be visible by normal means.

Another way machine learning is used as an attempt to prevent crime is to detect housebreak patterns. The algorithm attempts to predict the chances of a break-in with features of the house, the day of the week, and how close a house is to the other break-ins [9]. This system is always growing as it finds better patterns from the database. In this situation, they used the Series Finder for the pattern detection algorithm. With this method, crime patterns are automatically found, which allows law enforcement to take immediate action to try and prevent them [9]. Prior to this machine learning technique, to be able to find a pattern within the crimes would have taken most likely a few weeks or even years of going through the database, without even a guarantee that a pattern would be found.

There has been a model developed that can predict future crimes one week in advance with an accuracy of about 90% [10]. This illustrates how far machine learning can really go. This algorithm makes these predictions by learning patterns in time and geographic locations from public data on violent and property crimes. Oftentimes, crime prediction methods use an epidemic or seismic approach, to find the hot spots that spread to the surrounding area, however, the issue is the effects of the police enforcement in a high crime area. The model [10] singles out the crime by utilizing the time and spatial

coordinates and then detects the patterns to predict future crimes. Essentially, [10] found importance in city-specific crime patterns rather than relying on traditional neighborhood or political boundaries which may have a bias to them.

B. Issues

Machine learning allows multiple ways to bring awareness to criminal activities; however, it is not all as good as it seems. As artificial intelligence rises, the new machine learning tools seem to be benefiting us greatly, yet there is a catch. The systems are being fed data that are embedded with bias, which in turn simply reinforces inequality. The situation that arises is that the system will predict crime in a certain area, which in turn sends more police force there, and with more police force, more crime will be reported, thus it ends up in a cycle [11]. In this case, a section may see a constant increase in crime rather than a decrease, simply because there would be more police stationed there, thus more chances of people being caught and reported. On the other end, the sections that did not have such a high-risk crime rate would have fewer police there meaning even less crime would be reported simply because there is less manpower to be enforcing the law. Not only that, but machine learning can only use the data that is reported, meaning it can create biases. For instance, more people of color get reported than white, while the reality is that whites and people of color should all have the same chance of committing a crime, the machine can only analyze the information it is given, meaning it will target the people of color [12]. This is a rising topic on the news, about how cops tend to target people of color and take it easier on the opposite. Even something minor like getting pulled over for a speeding ticket, being a person of color means you have a higher chance of getting a ticket and written off, whereas, white people are more likely to be let off with a warning and nothing documented. This means that in the database, there are more people of color; even though this may not be the case in reality as both categories of people are being pulled over, only that one has a higher chance of being reported.

C. Our Approach

Rather than further developing one of the previous ideas, we decide to make a system that would bring awareness to the people and law enforcement of the severity of a crime; this way we are able to avoid biases and can create a system that simply showcases the results as is. Simply revealing how severe a crime is, would not increase nor decrease the amount of police being sent to a section. Our technique will support an informant to anyone interested in how dangerous a place would be. It does not encourage nor discourage law enforcement or any action to be taken, but merely a suggestion for an individual to be more aware of the cases. This would be a useful tool for individuals who want to be more informed about the safety of certain areas. It is not meant to be a tool just for the government, but a tool that everyone and anyone will benefit from, to increase awareness and promote safety.

II. Background

In this section, we provide the background required for this work.

A. Crime in Los Angeles

A crime can range from minor offenses like traffic violations to more serious cases like murder, assault, or theft. As a big city in California, Los Angeles has a big rising issue of crime occurring. In 2022, Los Angeles had an 11% increase in its overall crime rate, including both violent crimes such as rape, robberies, armed assault, and homicide, as well as property crimes like burglary, arson, and stealing vehicles [3]. Crime prevention is always a popular topic and an area in which governments typically spend large sums of money if they have the means. There are a few different approaches that can be taken to aid the ongoing struggle of crime prevention. There are direct methods such as increased law enforcement, and indirect methods such as raising awareness and grassroots movements that aim to provide better social services in high-crime areas with the hope of reducing crime by preventing people from being in socioeconomic situations in which they are pushed into crime.

B. Machine Learning

In the last decade, the advancement in computational ability has led to a drastic increase in the popularity of machine learning. Machine learning has also seen a rise in popularity due to mass recognition of its powerful capabilities to grant insights into difficult problems that traditional statistics and analysis cannot tackle. When it comes to solving everyday problems such as crime detection and crime prevention, it seems only logical to add machine learning to the tool belt so to speak. One way machine learning is being used to aid in crime prevention is through classification. Classification is an area of machine learning in which data is used to create a model such that new data fed to the model can be determined to belong to a specific class or category. Multiple studies [2,8] have proven the accuracy of common classification algorithms such as Naive Bayes, K Nearest Neighbors, Decision Trees, and Logistic Regression in classifying crime in order to provide accurate future predictions for crime types and affected areas.

III. Methodology

Within this section, we will cover the approach and outline of this work. This includes the pre-processing of the dataset. As well as the approach we develop to solve this problem.

A. Crime Information Dataset

We use the countless crimes that occurred within Los Angeles in California datasets [4]. This data has been updated over time since 2020 regarding data recollected from crime reports that occur within Los Angeles. The data for each entry has a maximum of 28 features with information such as:

- Time and date.

- Geographical information such as latitude, longitude, and street.
- Victim information such as age and sex.
- Crime scene information such as the type of crime.

These categories encompass all of the features within the dataset. Though not all information within the dataset is filled out, certain features within these categories had greater than 80% of its data missing or invalid, such as particular codes that are given from the incident report being missing. In this case, such features are dropped. Each category has its appropriate modification such as normalization or conversion into a binary form.

B. Target Variable

Within the dataset comes documents with information regarding particular data that are given for the crime scene category to help us better understand the meaning of this information. Within the "UCR Manual" document comes information regarding the data given, in addition to the description of the particular crime code provided that is determined for the crime committed [4]. In addition the document known as "UCR-COMPSTAT" has information regarding which of these codes were considered as a "violent", or "non-violent" crime [4]. With the information given in this document, we are able to create our target variable of the particular offense being either a "violent" or "non-violent" crime for each incident report.

C. Features

Once the pre-processing of the dataset has been completed, we settle on 7 features along with our target feature which is compromised of continuous, binary, and categorical data. Now that the new refined feature set has been composed, we are now able to apply the data to machine learning algorithms.

D. Classifiers

To create a system that will bring awareness to the severity of crimes, we will utilize four separate learning algorithms. Which are:

- Naive Bayes - Naive Bayes is a type of supervised learning that is based on Bayes' theorem and is an algorithm that calculates the probability of a certain outcome given a set of features [5]. In this case, Naive Bayes is used to classify crimes as violent or non-violent based on the features given and the probability of them leading to violence. While Naive Bayes typically sees its best performance when applied to documents or large sets of text data for sentiment analysis, it can also perform well as a general classifier. This algorithm assumes that all variables are independent of each other, and is highly regarded for its performance speed, which is fast enough to even perform real-time classification.
- K Nearest Neighbors (KNN) - KNN is a type of unsupervised learning that can be used to categorize data that does not have a category or group [5]. In this case, KNN can be used to identify areas that have the most similarities in terms of crime activity patterns. KNN has the advantages of being very simple to implement and comprehend as well as having great flexibility as to how distance is calculated between the test data points and the training data points. While this algorithm is easy to implement, the performance speed of this algorithm negatively scales with the size of the dataset and can be expensive in terms of memory since it is a lazy learner.
- Decision Trees - A decision tree is a type of supervised learning that utilizes branching to showcase every possible outcome of a decision [5]. In this case, Decision Tree can be used to classify crimes as violent or not by the features given. The advantages of decision trees are that they are easy to implement, easy to understand, and can easily be visualized, providing complete transparency as to why each decision was made. A disadvantage of this algorithm is that it can be quite prone to overfitting.
- Logistic Regression - Logistic Regression is a type of supervised learning that strives to estimate the possibility of an event occurring based on the previous data provided [5]. In this case, Logistic Regression can be used to estimate the probability of a crime being violent or non-violent based on the crime's characteristics. Logistic Regression is a simple yet powerful classification algorithm that is also noted for its performance speed.

IV. Experiments

With the selected machine learning algorithms and a processed dataset, we now have the tools to conduct experiments and analysis on those experiments to determine if a desirable result is produced. We will be comparing each of the four discussed algorithms to see which has the highest accuracy in classifying whether a crime is violent or non-violent.

A. Experimental Procedure

To proceed with our experiments, we split to have 80% of our data into the training set and the remaining 20% for the test set. We then conduct the following tests:

1) Select learner: Naive Bayes, KNN, Decision Tree, Logistic Regression;
2) Select feature: 28 total, some of which are: Date, Area name, District, Part, and Weapon.
3) Select feature restrictions: Time, Date, Latitude, Longitude, Street, Victim's Age, Victim's Sex, Type of Crime.

These features are chosen during the preprocessing of our data set as they contain information about the crime committed without being able to explicitly determine whether the crime is violent or not at a glance. For example, both violent and non-violent crimes can be committed in any area to any type of victim, however, features such as "weapon type" inherently provide information as to whether or not a crime is violent, and are therefore not as useful for our goals. Using these selected feature restrictions and the learner, we run tests to record the metric for each. We will then analyze the results of these tests in the upcoming sections.

B. First Model Created

Using the features picked, we attempted to run the machine learning algorithms, however, our first accuracy was very low. We found out that our dataset had a severe class imbalance, the minority class had very poor performance, so to counter this, we implemented Synthetic Minority Oversampling Technique (SMOTE). Which gave us the following results.

C. Model With Smote

Looking first only at the KNN model that was tested with neighbors ranging from 1 to 50. Referring to Fig. 1. The score peaked with it being right over 0.73 with N = 1 at its start, and with the higher KNN values dwindling down to .68. Though we settled on implementing the KNN model at N = 11 as it is where the model is no longer oscillating, and it is not taking a hit in accuracy. We did consider having a lower N value as the standard. However, the model was inconsistent with its scores with low N values. Thus, we are using N = 11 for the KNN model.

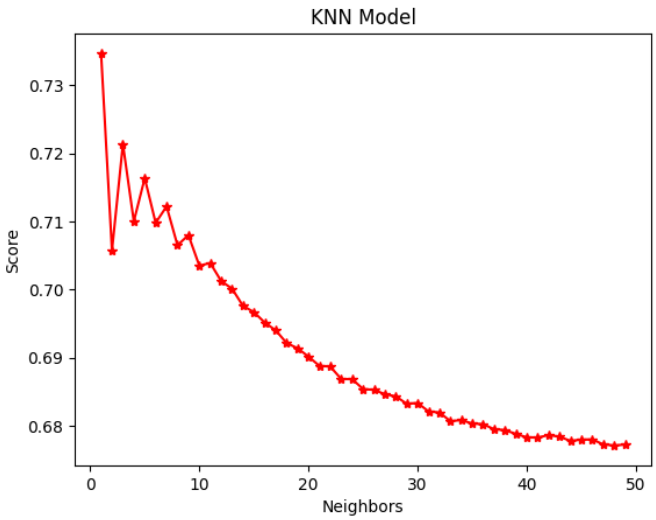
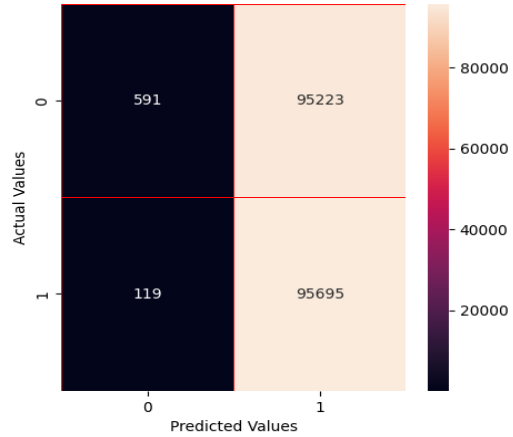


Fig. 1. KNN model with different KNN neighbors.

One of the ways we compare our models is based on the confusion matrix. A confusion matrix is a two by two table for a binary classification problem that has different outcomes of the predictions and the results of a classification problem[6]. This matrix is a great way to measure the performance of a classifier in depth.

Looking at each model and its performance solely from the score we see that the accuracy varies greatly for each model. The lowest was Naive Bayes, scoring around 0.50, and the highest was the decision tree, scoring around 0.79. The accuracy was based on whether the model predicted that based on the given data, would the crime be considered a ”violent”(1) or ”non-violent”(0) crime. Looking at the confusion matrices in figures 2-5, we see that each model aside from Naive Bayes is capable of predicting fairly well when the crime is ”non-violent”(0). This accuracy for predicting ”non-violent” crimes is strong. Looking at when the models are meant to predict the ”violent”(1) crimes, we see that they typically have a similar rate as predicting ”non-violent”, however, the outlier was Naive Bayes. In the most drastic case, as seen in Fig. 2, Naive Bayes was almost unable to classify any ”non-violent” crime. This leads to substantially high recall and precision values while still maintaining an extremely low accuracy score. The rest had a slightly better turnout in terms of being able to predict values for both ”violent” and ”non-violent”. The decision tree has the best overall ability to accurately predict both cases for this class. Though it did take a large hit on being able to predict when it is truly a ”violent” crime, as seen in Fig. 4.



	precision	recall	f1-score	support
0.0	0.83	0.01	0.01	95814
1.0	0.50	1.00	0.67	95814
accuracy			0.50	191628
macro avg	0.67	0.50	0.34	191628
weighted avg	0.67	0.50	0.34	191628

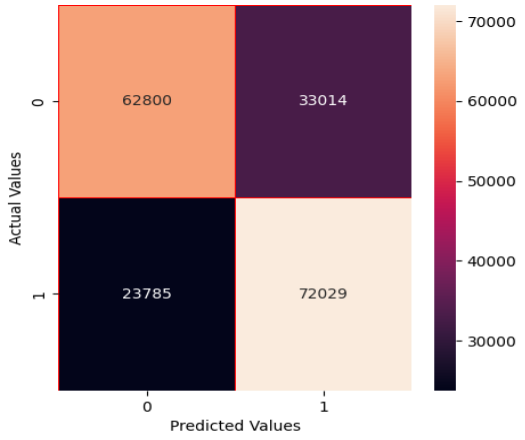
Fig. 2. Confusion Matrix for Naive Bayes.

D. A Closer Look Towards the Best Performing Model

The best-performing model with the dataset that we have provided would be the Decision Tree model. After running multiple metrics such as precision score, accuracy score, recall score, and F1 score, we noted that the Decision Tree model has the best overall scores. The Decision Tree model has precision, recall, accuracy, and F1 scores between 79% and 80% for each class. Our KNN model was also a runner-up, with around 70% to 71% overall for accuracy, precision, recall, and F1 scores. Trailing a bit behind our KNN model is the Logistic Regression model which had an accuracy score of around 60% as well, however after taking a look at the confusion matrix, it is evident that the KNN model is doing a little better in classifying both violent and non-violent acts for what they are.

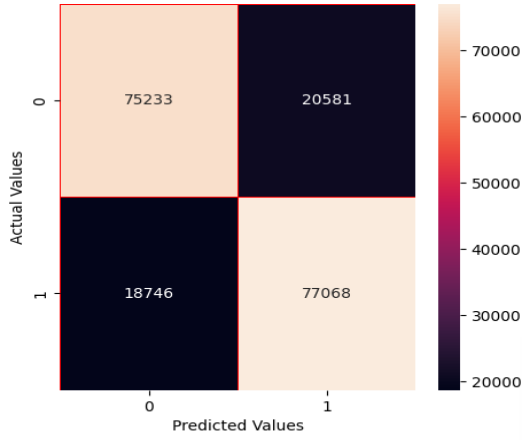
E. A Closer Look Towards the Worst Performing Model

From our confusion matrix in 2, we can see that Naive Bayes is by far our worst-performing model. This model went rogue so to speak, classifying almost every data point it came



	precision	recall	f1-score	support
0.0	0.73	0.66	0.69	95814
1.0	0.69	0.75	0.72	95814
accuracy			0.70	191628
macro avg	0.71	0.70	0.70	191628
weighted avg	0.71	0.70	0.70	191628

Fig. 3. Confusion Matrix for KNN.



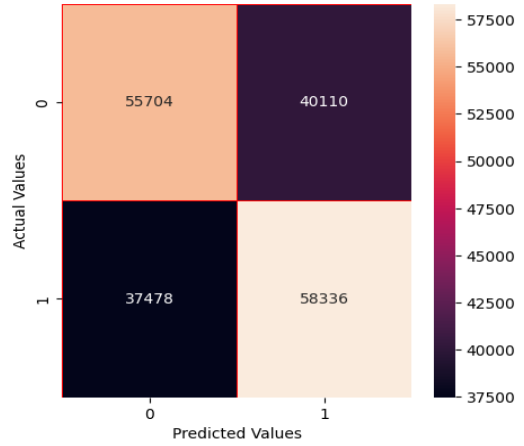
	precision	recall	f1-score	support
0.0	0.80	0.79	0.79	95814
1.0	0.79	0.80	0.80	95814
accuracy			0.79	191628
macro avg	0.79	0.79	0.79	191628
weighted avg	0.79	0.79	0.79	191628

Fig. 4. Confusion Matrix for Decision Tree.

across as ”violent”. This is made evident by the extremely high recall score compared to its other performance metrics. This algorithm is clearly not suited for this type of problem, as it had a similar issue even before dataset balancing techniques were performed in order to obtain more accurate results.

F. A Side By Side Comparison

In 6, we have a side-by-side of four performance metrics for each of our classifiers. These metrics - Accuracy, precision, recall, and F1 Score - are each indicative of the success of the model regarding true positives, true negatives, false positives, and false negatives. ideally, the goal would be for all scores



	precision	recall	f1-score	support
0.0	0.60	0.58	0.59	95814
1.0	0.59	0.61	0.60	95814
accuracy			0.60	191628
macro avg	0.60	0.60	0.60	191628
weighted avg	0.60	0.60	0.60	191628

Fig. 5. Confusion Matrix for Logistic Regression.

to be high, which would be representative of a model that is overall accurate for predicting both ”violent” and ”non-violent” crimes in this instance. By comparing the scores of this model, we can see that Naive Bayes has the largest variance in scores due to it really only having the ability to predict ”violent” for both ”violent” and ”non-violent” crimes. With the decision tree model, we see a relatively high accuracy score with similar scores for precision, recall, and F1. Once the dataset was balanced, the model performed consistently for both violent and non-violent crime predictions.

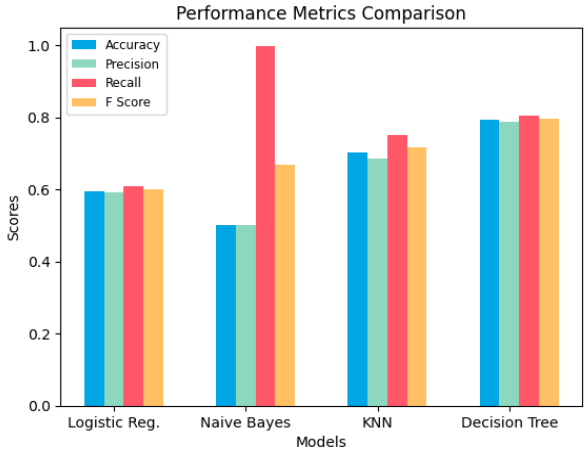


Fig. 6. Comparison of Performance Metrics Between Models

V. Conclusion

To create a system that will help bring awareness to the severity of potential crimes, we utilized different machine learning techniques with the hope of identifying specific variables that can aid in predicting whether a potential crime

will be violent. After comparing four classification algorithms through multiple tests, we observe that the decision tree classifier and KNN achieve the best accuracy with 0.79 and 0.70 respectively, while maintaining similar precision, recall, and F1 scores to their accuracy scores. We conclude that more work can be invested into this area to have a more in-depth analysis. There is potential for this system to become a crime prevention assistance.

A. Future Work

Our future work would involve the expansion of this classification to include the type of crime in addition to whether or not the crime is violent. Additionally, it would be helpful to use this data to identify crime hot spots that correspond to the time of day in order to better inform the citizens of Los Angeles regarding when and where they should consider being extra aware of their surroundings. We will also be trying to implement different data balancing techniques such as stratification in tandem with SMOTE to see if that will yield better results.

REFERENCES

[1] "Los Angeles, CA Crime Rate & Safety - US News best places," U.S. News A World Report. [Online]. Available: https://realestate.usnews.com/places/california/los-angeles/crime (accessed April 26, 2023).

[2] N. Shah, N. Bhagat, and M. Shah, "Crime forecasting: A machine learning and computer vision approach to crime prediction and prevention - Visual Computing for Industry, biomedicine, and art," SpringerOpen, April 29, 2021. [Online]. Available: https://vciba.springeropen.com/articles/10.1186/s42492-021-00075-z (accessed April 26, 2023).

[3] B. Pallarp and T. Abdollah, "Which Los Angeles neighborhoods are safest?," USA Today, March 20, 2023/ [Online]. Available: https://www.usatoday.com/story/news/2023/03/20/los-angeles-crime-rates-map-safest-neighborhoods/11032761002 (accessed April 26, 2023).

[4] LAPD, "Crime Data from 2020 to Present," lacity, February 10, 2020. [Online]. Available: https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8 (Accessed April 26, 2023).

[5] K. Wakefield, "A guide to the types of machien learning algorithms," A guide to the types of machine learning algorithms- SAS UK. [Online]. Available: https://www.sas.com/en_gb/insights/articles/analytics/machine-learning-algorithms.html (Accessed April 23, 2023).

[6] Simplilearn, "What is a confusion matrix in machine learning?," Simplilearn.com, Feb. 16, 2023. [Online]. Available: https://www.simplilearn.com/tutorials/machine-learning-tutorial/confusion-matrix-machine-learning (Accessed April 23, 2023).

[7] L. Buczak and M. Gifford, "Fuzzy Association RUle Mining for Community Crime Pattern Discovery, ACM SIGKDD Workshop on Intelligence and Security Informatics held in conjuction with KDD-2010, (2010).

[8] S. Albahli, A. Alsaqabi, F. Aldhubayi, H. T. Rauf, M. Arif, M. Muhammad, and H. Tayyab, "Predicting the type of crime: Intelligence gathering and crime analysis," ResearchGate. [Online]. Available: https://www.researchgate.net/publication/347947011_Predicting_the_Type_of_Crime_Intelligence_Gathering_and_Crime_Analysis. (accessed April 24, 2023).

[9] C. Rudin, "Predictive policing: Using machine learning to detect patterns of crime," Wired, Aug. 07, 2015. [Online]. Available: https://www.wired.com/insights/2013/08/predictive-policing-using-machine-learning-to-detect-patterns-of-crime/ (Accessed April 25, 2023).

[10] M. Wood, "Algorithm predicts crime a week in advance, but reveals bias in police response," Biological Sciences Division - The University of Chicago, June 30, 2022. [Online]. Available: https://biologicalsciences.uchicago.edu/news/features/algorithm-predicts-crime-police-bias (Accessed April 26, 2023).

[11] L. Jany, "Researchers use AI to predict crime, biased policing in major U.S. cities like L.A.," Los Angeles Times, July 04, 2022. [Online]. Available: https://www.latimes.com/california/story/2022-07-04/researchers-use-ai-to-predict-crime-biased-policing (Accessed April 26, 2023).

[12] H. Reese, "What happens when police use AI to predict and prevent crime?," [Online]. Available: https://daily.jstor.org/what-happens-when-police-use-ai-to-predict-and-prevent-crime/ (Accessed April 26, 2023).