1 Literature Review

1.1 Summary of existing systems

Predictive policing systems, a statistical approach to help predict criminal activity and policing decision making, [1] have been widely adopted in various regions to enhance law enforcement efficiency. Below are some notable systems and their effectiveness.

PredPol utilizes an ETAS(Epidemic-type aftershock) model integrated with machine learning algorithms, leveraging historical crime records as input features. The model's effectiveness is assessed based on the crime reduction rate and the Prediction Accuracy Index (PAI)[2] [3]. In experiments conducted in Los Angeles, PredPol resulted in a 7.4% decrease in crime[4]. Taking a different approach, RTMDx[4] uses RTM model[5] and focuses on built-environment data, such as streetlight density and land use, in addition to crime records. This system used RRI (Recapture Rate Index) as its evaluation metric[6], and has led to a 35% reduction in gun violence in Newark and a 33% decrease in motor vehicle theft in Colorado Springs[7].

PreCobs[8] (Pre Crime Observation System) is based on Near Repeat Theory, focuses on historical burglary patterns and residential typology data to generates grid maps highlighting high-risk zones based on time, location of past incidents. The system reduced burglaries by 30%-40% in several European cities[9].

Unlike the previous systems, HunchLab[10] integrated Multivariate Regression Models, RTM(Risk Terrain Modeling) with machine learning algorithms. It incorporates a more diverse range of inputs compared with PredPol, including jurisdictional boundaries, crime records, geographical data (such as points of interest), and temporal factors like weather and holidays. It has shown a 31% reduction in property crime in Philadelphia and a notable decrease in violent crime in Chicago.

While all these systems rely on historical crime records as their primary input, they differ in other key aspects such as feature engineer techniques and effectiveness.

1.2 Review of Model methods

From the above systems, it can be seen that the currently used models include four main categories: Statistical, Spatial Analysis and Machine Learning Models.

Statical analyses are based on historical data and statistical methods, suitable for analyzing crime trends and hotspots, including regression analysis and KDE[11] [12](Kernel Density Estimation). While spatial analysis models focus on the geographical distribution characteristics of crime, including RTM[13](Risk Terrain Modeling) and Near Repeat Theory. RTM[14] evaluates environmental factors contributing to crime risks by analyzing spatial correlations between locations and physical or social conditions. As Near Repeat Theory is based on criminological theories that crimes tend to cluster in time and space, these

models predict repeat offenses near prior incidents.
Machine learning model are suitable for big data environments, including random forest, SVM[15] and neural networks. They are used for classification tasks like predicting crime types or high-risk areas.

1.3 Feature engineer techniques

Feature engineering is crucial for machine learning and data analysis of crime data, as it contributes to modifying the data's features (including spatial, temporal etc.) to better represent the nature of the problem and improve the model accuracy[16], including calculation approach, representation and feature importance. For example, when engineering crime hotspots in urban environment, Borges et al.[17] deconstruct "street network" as dead-end density, major road length and so on, and decided to calculate major road length by summing up all roads of this type. For building types, they use binary representation to present if certain buildings are within the area. Finally, they adopt a random forest classifier to determine the importance of features.
In feature engineering, analyze and deconstruct the features, select proper representations and importance are necessary techniques to improve the model.

1.4 Evaluation metrics

To test and adjust the efficiency and efficacy of models, evaluation metrics are introduced to measure the abovementioned systems and models.
Predictive Accuracy Index (PAI) is commonly used in the field of criminal geography. It measures the concentration of crimes captured within predicted high-risk areas relative to their size. Based on it, Levine[18] proposed using RRI (Recapture Rate Index) together with PAI to measure accuracy.
Accuracy and precision are commonly used in classification tasks.
Accuracy represents the proportion of correctly predicted units among all units, which is greatly influenced by the crime level and spatial clustering of incidents in the study area. Precision refers to the proportion of units where crimes actually occurred among the predicted crime occurrence units.
Fairness and Transparency are also qualitative metrics, which are implied in the above systems. Fairness means to evaluate algorithmic bias to ensure equitable policing across different demographic groups. Transparency refers to assess whether predictions are explainable to law enforcement officers and the public.

2 Dataset

In this data analysis project, we will use the open dataset from the Chicago Police Department's Citizen Law Enforcement Analysis and Reporting system, including crime data from 2001 to present (excluding the most recent 7 days) in

Chicago. The dataset contains 22 columns, providing information about temporal aspects (e.g., date, year, updated_on), spatial aspects (e.g., block, location, ward, community area, beat), categorical features (e.g., FBI code, domestic, primary type), case handling (e.g., arrest, domestic), and detailed descriptions of the incidents.
To ensure the training data reflects recent crime trends, we focus only on the data after 2023, which is around 500,000 rows. Within the dataset, some columns exhibit varying degrees of missing data: location_description (0.44%), ward (0.0004%), community_area (0.0004%), location (0.279%, the same to coordinates). Also, some data require type conversion for further analysis.
Thus, data preprocessing for the crime data includes missing data handling, feature extraction, data cleaning, and normalization.

2.1 Data Cleaning

In data cleaning, we address redundancies and inconsistencies in the data format.
Some of the features are not necessary in our future training, either containing redundant inofrmation or are irrelevant for model training, thus we drop the features below:

- Block and location. Block is too granular for prediction and can be replaced by community area and district. Location can be replaced by longitude and latitude as they contain equivalent information.
- Primary_type and description. Almost all primary_type can be inferred from first two digits (including leading zeros) of IUCR code, and description for the last two digits. Although there are exceptions such as case started with 13 can be "criminal trespass" or "criminal damage", but they are trivial and will not significantly affect the model. Therefore, we drop these two columns and replace them with the related IUCR code.
- X-coordinate, Y-coordinate, updated_on.These features are not used in future modeling.

With 138 unique values for location_description, we need to simplify the future training by limiting the number of entries for this column. We group the descriptions with occurrences under 60 to "Others".
There are three kinds of format inconsistencies in Chicago crime data, and should be handled differently.
First, the date column is not in proper datetime format, which we transformed to standard datetime format for better time-based analysis. Longitude and latitude also are transformed to float for future missing value handling.
Also, we standardize the location description by stripping the whitespace and special characters and converting them to lowercase. The up-mentioned IUCR

alternative to description data also include string, we delete those rows to avoid confusing future machine learning.
Lastly, as arrest and domestic columns are originally checkbox data, we convert them to integers (1 for True and 0 for False) for facilitate modeling.
Because there are no duplicates detected in this dataset, we did not perform duplicate removal to clean the data.

2.2 Missing Data Handling

As missing data occurs primarily in spatial and descriptive columns, different strategies are applied: spatial data are inferred by other spatial columns without loss and descriptive column are filled with specific values.
For location coordinates, we use the median coordinates of the cases that happened in the same block or district to impute the missing data. Similarly, we use the most frequent wards and community area data from the same district to suggest the missing ward and community.
For location description, as the description may vary case by case, we choose to fill in "unknown" to handle this missing value.

2.3 Temporal and Spatial Aggregation

Time and space should be two of the most crucial data in the Chicago crime dataset. However, with limited date and location data, the dataset could be too specific and less informal. Thus, we did temporal and spatial aggregation to extract information and take deeper look in the data.
To get more clear information about temporal occurrences, we extract the year, month, day, and hour from the date column. Additionally, a new column is created to distinguish weekdays from weekends, providing insights into weekly crime patterns.
We also aggregate the crime counts per block and district to identify crime density in different areas.

2.4 Feature Standardization

To prepare categorical features for machine learning, we apply one-hot encoding to the location_description column, turning it into numerical features suitable for training.