**1.Introduction**

Bangladesh a small country with large population among all the cities Dhaka is most populated city in here. In other word it is one of the most diverse cities in Bangladesh. B.and k.(2018) claim that Crime is one of the biggest and dominating problems in our society and its prevention is an important task . Even though the criminal branches are very active, they are become unable to prevent crime. To made crime prediction possible based on previous data there has been remarkable increase in machine learning algorithms. The aim of this project is to perform analysis and make prediction of crimes in mirpur area using machine learning approach. It focuses on creating a model that can help to detect the pattern of crime by its analyzing previous crime type, location and other necessary attributes. In this project we have used various machine learning models like decision tree, Random forest, naive bayes, logistic regression, K-NN to predict crimes pattern. Geographical location can be used to have better idea about crime pattern. This will help law enforcement agencies to detect and predict crimes with higher accuracy and take precautionary steps. This can help to improve people safety in society as well as will help reduce the rates of crimes indirectly.

The behavioral pattern of criminal activities is studied under crime analysis which is a sub branch of criminology that tries to identify the indicators of such events. Machine learning model work with large amount of data and apply different techniques to find patterns in data making it very useful for predictive analysis. Crime branches takes different patrolling strategies to keep the area secure based on the previous information they gain. A machine learning agent can analyze previous data of crime type, time, geographical location, date and then predict pattern and occurrence of crime based on these factors. It allows predicting nominal class labels and this technique is known as classification. Classification can be used on many different factors like economical background, social status, education level weather forecasting etc.

The following thesis paper describes in brief the research work done before on crime prediction and followed with details of datasets used along with data preprocessing. The above problem made us to conduct a research about how can solve a crime case made easier and ensure people safety. Through we have many previous documentation and cases; these will make our thesis work much easier and faster. we have collected our data from DMP, mirpur police station and from some other source of previous some year. In next section data analysis and supervised classification models like decision tree, random forest is used to predict crime pattern. At the end conclusion and future scope is mentioned as well.

* 1. **Problem Statement**

Population is an asset for a country .Every day a huge number of crime occur all over the world. specially a populated city like Dhaka it’s a common issue. A small statistics (Murder 216, Robbery 83, Women and Child Repression 1782, Theft 1290, Narcotics 16215, and many more only in Dhaka in 2018. Total Cases under DMP - 27150).we need ensure people security by avoiding crime as much as possible. there are so many factors behind occurring crimes here we consider lack of awareness, social status ,educational background, geographical area which can effect on criminal activities directly. by analyzing these attributes we can predict crime pattern ,hotspot of crime so that law enforcement branch can identify crime prone location then take precautionary steps immediately beforehand. Machine learning algorithms like Naïve Bayes classification, Decision Tree classification, random forest, logistic regression are used and tested to get better prediction of accident-prone road locations inside Dhaka metropolitan region.

* 1. **Significance of the problem and Historical Background**

**1.3 Objectives**

**1.3.1 General Objective:** To provide prior assumption about the class of the crime to the enforcement agencies so that it would give them tactical advantages and help resolve cases faster.

**1.3.2 Specific Objectives**

* To identify and analyze major causes behind criminal activities in Bangladesh
* To ensure maximum accuracy from the model
* To identify hot-spot of criminal activities more accurately

**1.4 Research Question**

* What are they major causes behind crime in Dhaka city?
* Which ML model gives the best result to classify imbalanced dataset?
* Just with a geographical location and time, how accurately can we make a prediction?
* What is the accuracy of the designed model?

**1.5 Scope of the study**

* This regional boundary of this report is Dhaka metropolitan region
* Major criminal activities will be considered like theft
* Among different factors causing criminal activities like crime type ,crime time, crime date, education level, geographical location are consider and studied
* Random forest and Decision tree classifier are tested for classification and confusion matrix and cross validation are used to check the accuracy are the dataset.
* Accident data from 2017-2018 has been taken into considerations

**1.6 Limitations of the study**

The study is restricted to Dhaka Metropolitan region only. The results obtained in this study cannot be extended to other cities.

* The dataset we get was not enough to get accurate perdition
* Need more organized and complete data to get more accurate prediction.
* Exact Geographical location will provide more accuracy in result
* Deep Learning & Neural Networks could provide a better prediction
* Imbalanced classes
* Time constraints and lack of some more important factors causing criminal activities, may drop the accuracy of the prediction.

**1.7 Definitions**

* **Machine learning**: Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.
* **Classification algorithm**: In machine learning, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation.

Previous researches in crime prediction have found

that factors like education, poverty, employment, and climate

affect the crime rate [2].

**2. Literature Review**

Previous researches in crime prediction have found

that factors like education, poverty, employment, and climate

affect the crime rate [2].

Previous researches in crime prediction have found

that factors like education, poverty, employment, and climate

affect the crime rate [2]. Vanc

Previous researches in crime prediction have found

that factors like education, poverty, employment, and climate

affect the crime rate [2]. Vanc

Previous researches in crime prediction have found

that factors like education, poverty, employment, and climate

affect the crime rate [2]. Vanc

Previous researches in crime prediction have found

that factors like education, poverty, employment, and climate

affect the crime rate [2].

Previous researches in crime prediction have found

that factors like education, poverty, employment, and climate

affect the crime rate [2].

Previous researches in crime prediction have found

that factors like education, poverty, employment, and climate

affect the crime rate [2].

Previous researches in crime prediction have found

that factors like education, poverty, employment, and climate

affect the crime rate [2].

Previous researches in crime prediction have found

that factors like education, poverty, employment, and climate

affect the crime rate [2]

Previous researches in crime prediction have found

that factors like education, poverty, employment, and climate

affect the crime rate [2]

Previous researches in crime prediction have found

that factors like education, poverty, employment, and climate

affect the crime rate [2

Previous researches in crime prediction have found

that factors like education, poverty, employment, and climate

affect the crime rate [2]

Previous researches in crime prediction have found

that factors like education, poverty, employment, and climate

Crime is a socio-economical problem affecting life quality

and economic growth [1]. The specifics of how crime is

conducted changes depending on the type of society and

community. Previous researches in crime prediction have found

that factors like education, poverty, employment, and climate

affect the crime rate [2].

Crime is a socio-economical problem affecting life quality

and economic growth [1]. The specifics of how crime is

conducted changes depending on the type of society and

community. Previous researches in crime prediction have found

that factors like education, poverty, employment, and climate

affect the crime rate [2].

Crime is a socio-economical problem affecting life quality

and economic growth [1]. The specifics of how crime is

conducted changes depending on the type of society and

community. Previous researches in crime prediction have found

that factors like education, poverty, employment, and climate

affect the crime rate [2].

Various researchers have addressed the problems regarding crime control and have proposed different crime-prediction algorithms. The accuracy of prediction depends on the attributes selected and the dataset used as a reference. A. et al(2014) stated that Crime is a socio-economical problem affecting life quality and economic growth. The specifics of how crime is conducted changes depending on the type of society and community. H. et al (2016) stated that Previous researches in crime prediction have found that factors like education, poverty, employment, and climate affect the crime rate .

In, human behavioral data derived from mobile network activity combined with demographic information using real crime data were used to predict crime hotspots in London, UK. In [4], a comparison between two classification algorithms, Decision Tree and Naïve Bayesian, was performed using WEKA, an open-source data mining software, and 10-fold cross-validation. The socio-economic, law-enforcement, and crime datasets for this study were compiled from the 1990 US Census, the 1990 US LEMAS survey, and the 1995 FBI UCR, respectively. The road accident patterns in Ethiopia was studied in [5] considering various circumstantial factors like the driver, weather, car, and road conditions. Three different classification algorithms, KNN, Naïve Bayesian, and Decision tree were used on a dataset of 18,288 accidents. The prediction accuracy for all three algorithms was between 79% to 81%.

A major challenge regarding crime prediction is analyzing large crime datasets accurately and efficiently. Data mining is utilized to find hidden patterns in large crime datasets quickly and efficiently. The increased efficiency and reduced errors in crime data-mining techniques increase the accuracy of crime prediction. A general data-mining framework was developed in [6] based on the experience of the Coplink project, conducted at the University of Arizona. Most research in crime prediction is focused on identifying crime hotspots, which refers to the areas in which the crime rates are above the average level. In [7], authors provided a comparative analysis of Kernel Density Estimation (KDE) and Risk Terrain Modeling (RTM) algorithms for creating hotspot maps and proposed area-specific predictive models using sparse data. In [8], a spatial-temporal model using histogram-based statistical methods, Linear Discriminate Analysis (LDA), and KNN were adopted for crime hotspot prediction. In [9], a crime incidence-scanning algorithm was applied to train Artificial Neural Network (ANN) enhanced by the Gamma test to predict the crime hotspots in Bangladesh. A data-driven machine-learning algorithm based on broken-window theory, spatial analysis, and visualization techniques was used in [10]to analyze drug-related crime data in Taiwan and predict emerging hotspots.

An approach based on Auto-Regressive Integrated Moving Average model (ARIMA) was utilized in [11] to design a reliable predictive model for forecasting crime trends in urban areas. In [12], authors proposed a probabilistic model of spatial behavior for known offenders based on a random-walk-based approach to model offender activity in the Metro Vancouver area. The random forest algorithm was used in [13] to quantify the role of urban indicators for crime prediction in Brazil. In [14], prospective method, Dempster-Shafer theory of evidence, and the multi-kernel method were used to develop a crime-prediction solution for Chilean large cities. In [15], three algorithms, KNN, Parzen windows, and Neural Networks, were developed, tested, and compared for predicting the crimes in the city of San Francisco. In [16], Gradient Boosting Machine (GBM) technique was applied in a machine-learning prediction model to find hidden links in criminal networks and the weighted page-rank method was used as an effective strategy to weaken and destroy such networks.

Based on the literature, in this research, the classification algorithms random forest and decision tree were used to analyze the mirpur metropolitan area crime dataset.

Crime is a socio-economical problem affecting life quality

and economic growth [1]. The specifics of how crime is

conducted changes depending on the type of society and

community. Previous researches in crime prediction have found

that factors like education, poverty, employment, and climate

affect the crime rate [2].

Crime is a socio-economical problem affecting life quality

and economic growth [1]. The specifics of how crime is

conducted changes depending on the type of society and

community. Previous researches in crime prediction have found

that factors like education, poverty, employment, and climate

affect the crime rate [2].

3.1 Machine Learning

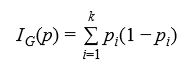
Machine Learning is a type of artificial intelligence which recognizes patterns by data analysis. Through machine learning a computer can learn and make predictions from data without being thoroughly programmed. We can divide Machine learning into three main categories: Supervised Learning, Unsupervised Learning and Reinforcement Learning. In this paper, we will use supervised learning methods in order to predict crime categories.

3.2 Supervised Learning

Supervised learning is a machine learning model that can predict an output from a set of inputs. In supervised learning, the output labels are specifically defined. Input object contains various number of features and usually is represented in a vector form. In the training dataset, each input object is paired with a specific output object. A supervised learning algorithm develops a predictive model using training data and fits new information to the model. Separating training and testing data helps supervised learning models to avoid overfitting. The labels of new information is predicted by the algorithm. Supervised learning models can be implemented on both classification and regression problems. In the crime dataset, the goal is to predict the category of a criminal incident in a given time and place. As the crime categories are discontinuous, this is a supervised classification problem. There are different types of supervised classification models. In this paper Decision Tree, Gaussian Naive Bayes, Linear Regression K-Nearest Neighbor and two Ensemble Methods - Adaboost and Random Forest are used.

3.3.1 Decision Tree Classifier

Decision tree classification model forms a tree structure from dataset. Decision tree is built by dividing a dataset into smaller pieces. At each step in the algorithm, a decision tree node is splitted into two or more branches until it reaches leaf nodes. Leaf nodes indicates the class labels or result. At each step, decision tree chooses a feature that best splits the data with the help of two functions: Gini Impurity and Information Gain. Gini Impurity measures the probability of classifying a random sample incorrectly if the label is picked randomly according to the distribution in a branch



Gini Impurity is computed by summing the probability p times the probability of i mistaking while categorizing an item (1 – pi ) .

While building the tree, Information Gain helps to decide which feature to split next at each step. Information Gain can be calculated using entropy, which is a function to calculate expected value at each step. Entropy is defined as :



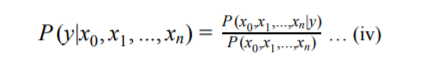
p represents the percentage of each feature being present in the child node after a i split. Sum of pi is always 1. Information Gain can be calculated using the following equation :

IG = Entropy(parent) − Weighted Sum of Entropy(children) … (iii)

At each step, decision tree try to make splits that give the purest child nodes.

3.3.2 Gaussian Naive Bayes

Gaussian Naive Bayes is a supervised classifier that uses naive assumption that there is no dependency between two features. This classifier is implemented by applying Bayesian Theorem. According to the theorem, class y and a dependent feature vector consisting of x , , .., , has the following relationship:



This probability model, along with a decision rule construct Naive Bayes Classifier. There are different types of Naive Bayes classification algorithms based on data distribution. In this paper, Gaussian Naive Bayes is used, where data is assumed to be distributed according to a Gaussian distribution.

3.3.3 Logistic Regression

Logistic regression uses linear boundaries to classify data into different categories. Logistic regression can work on both binary and multiclass problems. For multiclass dataset, one vs the rest scheme is used. In this method, logistic regression trains separate binary classifiers for each class. Meaning, each class is classified against all other classes, by assuming that all other classes is one category.

3.3.4 K-Nearest Neighbor

Nearest Neighbors method is used in both supervised and unsupervised learning. While testing with new data, KNN looks at k data points in training dataset which are closest to the test data point. k indicates the number of neighbors voting to classify a datapoint. The distance can be measured with various metrics. Euclidean distance is the most common choice.

3.3.5 Ensemble Methods

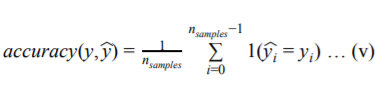
Ensemble learning is a method of combining multiple learning algorithm together to achieve better performance over a single algorithm. Ensemble methods can be divided into two categories: averaging methods and boosting methods. Averaging method independently develops different models and averages their predictions. In the combined model the variance gets reduced, and therefore, it gives better result. Boosting method is built by using separate base models sequentially. The goal of boosting is to reduce biases of the combined model and to build a powerful model from several weak models. In this paper, two ensemble methods are used: Random Forest, which follows the principle of averaging method and Adaboost which is a boosting model.

3.3.5.1 Random Forest

In this ensemble model several decision trees are built using samples drawn with replacement from the training set. The splitting of each node of a tree is not based on the best 19 split of all features, rather the best split among a random set of features. The bias of the tree increases due to randomness, but averaging also helps to decrease variance, hence this model often achieves better result.

3.4.1 Accuracy

Accuracy measures how many predictions are matched exactly with the actual or true label of the testing dataset and returns the percentage of correct results. Accuracy can be calculated using the following equation:



The fraction of correct prediction over n is defined by equation (v), where is sample yi ︿ the predicted value of i th sample and y is the corresponding true value.

3.4.2 Confusion Matrix

Confusion matrix is different than the metrics that have been discussed so far. Confusion matrix returns a table layout that helps to visualize the performance of an algorithm rather than producing a numerical value that indicates the goodness of the algorithm. Confusion matrix is an X × Y matrix where X dimension indicates the true classes and Y dimension indicates the predicted classes. The value of X and Y is equal, and they indicate the number of classes in a dataset.

4.1 Crime dataset and attributes

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

4.2 Classification of crimes

|  |  |
| --- | --- |
| **Category** | Total |
| Robbery | 83 |
| Murder | 216 |
| Riot | 17 |
| Kidnapping | 75 |
| Theft | 1290 |
| Burglary | 613 |
| Woman & Child Repression | 1782 |
| Dacoity | 17 |

Table-2: Frequency of top 8 crimes

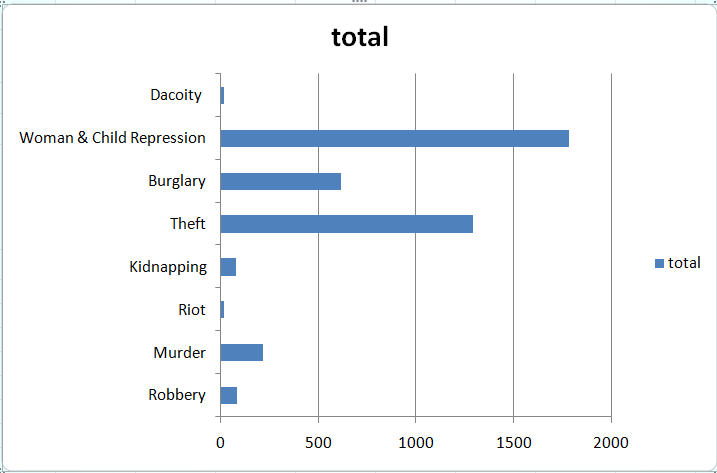


Fig-1:total number of occurrence of of crime categories

4.3 Features

4.3.1 Time of crime incident

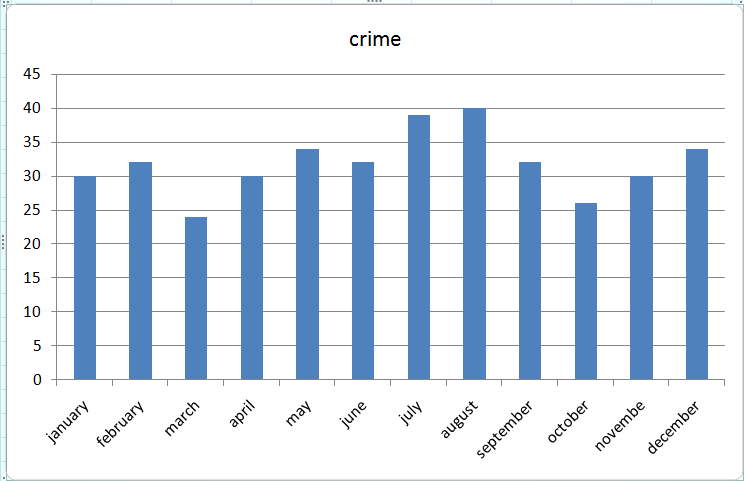


Fig-2: Crimes occurring in different months

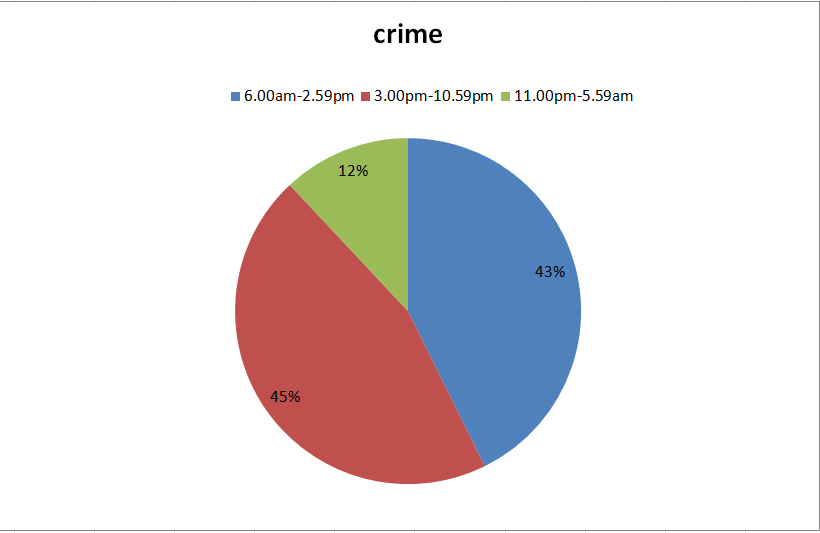


Fig-3: Criminal activities occurring in different hour of day

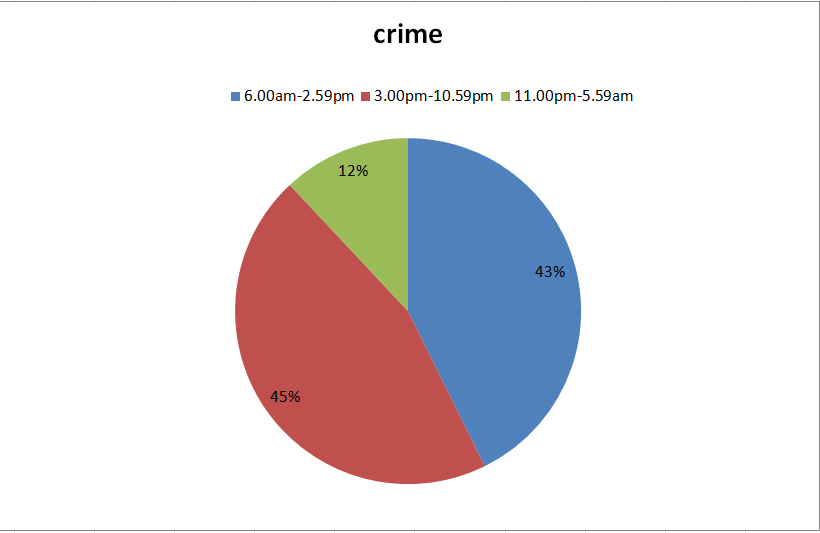
4.3.2 Location of crime incident

5.1 Preprocessing

5.2 Training and Testing Dataset

5.3 Extracting New Features

5.3.1 New Feature from hour



REFERENCE

Alkesh Bharati, Dr Sarvanaguru RA.K ,”Crime Prediction and Analysis Using Machine Learning” in International Research Journal of Engineering and Technology (IRJET) ,Volume: 05 Issue: 09 | Sep 2018

[1] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A.

Pentland, "Once upon a crime: towards crime prediction from

demographics and mobile data," Proc. of the 16th Intl. Conf. on

Multimodal Interaction, pp. 427-434, 2014.

[1] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A.

Pentland, "Once upon a crime: towards crime prediction from

demographics and mobile data," Proc. of the 16th Intl. Conf. on

Multimodal Interaction, pp. 427-4

Multimodal Interaction, pp. 427-434, 2014.

[1] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A.

Pentland, "Once upon a crime: towards crime prediction from

demographics and mobile data," Proc. of the 16th Intl. Conf. on

Multimodal Interaction, pp. 427-434, 2014.

A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland, "Once upon a crime: towards crime prediction from demographics and mobile data," Proc. of the 16th Intl. Conf. on Multimodal Interaction, pp. 427-434, 2014.

H. Adel, M. Salheen, and R. Mahmoud, "Crime in relation to urban design. Case study: the greater Cairo region," Ain Shams Eng. J., vol. 7, no. 3, pp. 925-938, 2016.

R. Iqbal, M. A. A. Murad, A. Mustapha, P. H. Shariat Panahy, and N. Khanahmadliravi, "An experimental study of classification algorithms for crime prediction," Indian J. of Sci. and Technol., vol. 6, no. 3, pp. 4219-4225, Mar. 2013.

H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau, "Crime data mining: a general framework and some examples," IEEE Computer, vol. 37, no. 4, pp. 50-56, Apr. 2004.

T. Beshah and S. Hill, "Mining road traffic accident data to improve safety: role of road-related factors on accident severity in Ethiopia," Proc. of Artificial Intell. for Develop. (AID 2010), pp. 14-19, 2010.

M. Al Boni and M. S. Gerber, "Area-specific crime prediction models," 15th IEEE Intl. Conf. on Mach. Learn. and Appl., Anaheim, CA, USA, Dec. 2016.

N. Mahmud, K. Ibn Zinnah, Y. Ar Rahman, and N. Ahmed, "CRIMECAST: a crime prediction and strategy direction service," IEEE 19th Intl. Conf. on Comput. and Inform. Technol., Dhaka, Bangladesh, Dec. 2016.

Y. L. Lin, L. C. Yu, and T. Y. Chen, "Using machine learning to assist crime prevention," IEEE 6th Intl. Congr. on Advanced Appl. Inform. (IIAI-AAI), Hamamatsu, Japan, Jul. 2017.

E. Cesario, C. Catlett, and D. Talia, "Forecasting crimes using autoregressive models," IEEE 14th Intl. Conf. on Dependable, Auton. and Secure Comput., Auckland, New Zealand, Aug. 2016.

M. A. Tayebi, U. Glässer, and P. L. Brantingham, "Learning where to inspect: location learning for crime prediction," IEEE Intl. Conf. on Intel. and Security Inform. (ISI), Baltimore, MD, USA, May 2015,.

] L. G. A. Alves, H. V. Ribeiro, and F. A. Rodrigues, "Crime prediction through urban metrics and statistical learning," Physica A, vol. 505, pp. 435-443, 2018.

N. Baloian, E. Bassaletti, M. Fernández, O. Figueroa, P. Fuentes. R. Manasevich, M. Orchard. S. Peñafiel, J. A. Pino, and M. Vergara, "Crime prediction using patterns and context," IEEE 21st Intl. Conf. on Comput. Supported Cooperative Work in Des. (CSCWD), Wellington, New Zealand, Apr. 2017.

M. V. Barnadas, Machine learning applied to crime prediction, Thesis, Universitat Politècnica de Catalunya, Barcelona, Spain, Sep. 2016.

E. Budur, S. Lee, and V. S. Kong, "Structural analysis of criminal network and predicting hidden links using machine learning," arXiv:1507.05739, Sep. 2015.