Urban Crime Trends Analysis and Occurrence Possibility Prediction based on Light Gradient Boosting Machine

Xiangzhi Tong*; Pin Ni*†; Qingge Li‡; QiAo Yuan*; Junru Liu*; Hanzhe Lu*; Gangmin Li*;
* Research Lab for Knowledge and Wisdom, Xi'an Jiaotong-Liverpool University, China
† The University of Auckland, New Zealand ‡ Xi'an Jiaotong-Liverpool University, China

***Abstract*—Big Data and Machine learning have been increasingly used to fight against Urban crimes. Our goal is to discover the connection between crime-related factors and the underlying complex crime pattern. Therefore, to predict the possibility of crime occurrence. Light Gradient Boosting Machine (LightGBM) Model is adopted in our study to predict the crime occurrence possibility based on actual crime information. We found that the prediction results are approximately consistent with an actual variation. We hope this work could help with crime prevention and policing.**

***Index Terms*—Light Gradient Boosting Machine, Crime Forecasting, Data Analysis, Random Forest**

I. Introduction

With the progress of urbanization and growing population density, the crime rate and the number of public disorder cases have been growing tremendously. This creates a great difficulty for the local police department to fight against the crime and threatens the social security and stability of normal life. Increased crime and cases of social disorder have also generated a large number of criminal records. Make use of these huge amounts of data could help crime prevention and win the war of fighting against crime. The challenge is how to effectively use these data and to discover any underlying connections and patterns among the crimes. With the full understanding of the insight of the data, it could help to prevent it happens by taking the initiative and arrange actions in advance.

Data mining is an efficient and popular method to discover the underlying pattern of big data, especially in the field of criminal analysis. Since crimes are neither a systematic nor completely random process, it is extremely difficult to predict whether the crime will happen or what kind of crime will occur. However, recent developments in machine learning and big data shed the light on these difficult problems. Therefore, we apply the Light Gradient Boosting Machine Model (LightGBM) model and big data techniques to predict the crime occurrence possibility by considering involved features. Although the practical prediction accuracy can not reach 100 percent, the outcome of our application could still help in reducing the crime rate to some extent by providing alerting information in crime-sensitive areas [1].

This paper reports our efforts in crime prediction using LightGBM. We firstly train the LightGBM model with optimized parameters base on the crime records released by San Francisco Police Department. Then, we apply the pre-trained model with actual crime data to forecast its re-occurrence. To figure out crime patterns, we tried to combine the LightGBM model with other approaches and data analysis tools, such as Pandas, NumPy to consider more underlying factors such as the geographical distribution of crimes and crime rate variations. It is a hope that our study can also provide some technical application inspirations for law enforcement and community security improvement.

The paper is organized as follows: Section II describes the literature review of related work, section III explains the applied methodology, section IV provides descriptions of the experimental process, section V analyzes the acquired results, and section VI is a general conclusion.

II. Related Work

A. Criminal Feature Analysis

Hu et al. [2] proposed an approach that analyzes spatial-temporal crime patterns base on the near-repeat and social network analysis, such as average computing cluster coefficient computation. Fan et al. [3] construct a model to identify underlying criminal relationships through mining and analyzing the call detail records in the criminal interactive behaviors. Tayebi et al. [4] build a brand-new computational co-offending network analysis model aim to identify organized criminal groups, and extract related information from a large real-life crime dataset.

Isah et al. [5] design a bipartite network model which helps mine underlying relationships among criminals and relationships between organizational structure and discovers the organizational structure of criminal groups. Diviak et al. [6] summarize current applied sophisticated network models and put forward the detail of three main challenges, about social network analysis in the criminal application field, which are data availability, proper formulation of theories, and adequate methods application.

B. Crime Variation Forecasting

Jiang and Barricarte [7] present a new stochastic crime rate forecast and decomposition method, which applies Singular Value Decomposition in matrix theory to predict age-specific crime rates based on time-series analysis. Then the predicted crude rate, age-specific crime rate and age-sex structure can be obtained afterward. Derive from its demonstration method, we choose to apply some similar random tendency prediction models, such as Random Forest, Naive Bayes and LTSM.

Vural et al. [8] propose a criminal prediction model which bases on Naive Bayes, Gaussian mixture model, and a parametric model based on the K-means method to help to clarify the incidents with its 83 percent success rate for security forces. Furthermore, another superiority of the model is its ability to take comprehension into the decision-making process. As a consequence, we apply multiple suitable models to generate prediction results, which would help with the investigation and exploration process of model characteristics.

In [9], Awal et al. employed a linear regression prediction model to forecast future trends in the crime of Bangladesh. They classify the crime data of various regions and apply the prediction model to forecast future crime trends of the metropolitan and divisional regions of Bangladesh. And the prediction outcome aims to be helpful and convenient for Bangladesh police and law enforcement agencies to forecast future crime of Bangladesh and release more targeting policies against the criminal trend.

C. Criminal Data Analysis and Visualization

Kasim et al. [10] develop an interactive web-driven geographical system to record crime-related information in a geospatial form and further raise public criminal prevention awareness. If crime-related data are presented with specific icons at the site where the crime happens, it would be more cautionary for citizens to raise awareness.

Chen et al. [11] propose a criminal relationship identification mechanism, which applies scalable visualization tools and the Hyperbolic Tree model to better retrieve and analyze these relationships with superior efficiency.

This is clear that many efforts have been focused on the different technologies to discover the crime patterns and the underlying factors and warning viewers. We argue that crime prediction is not a pure prediction problem, rather it is also an analytical and reporting problem. To demonstrate our opinion, we choose a typical crime dataset and illustrate our prediction model construction, factor analytical process, and the results visualization.

III. Methodology

The applied methodology description is divided into 4 sections as follows: A. Dataset Description, B. Forecast Model Selection, C. LightGBM Description and D. Performance Evaluation Method.

A. Dataset Description

The crime dataset of our work is the online released data of the San Francisco police department, which contains 878,050 different types of categorized crime incidents details from 2001 to 2020 in San Francisco. There are 9 featured attributes contained in the data:

- **Dates**: Detailed date and time when the crime occurs.
- **Category**: Particular type of the crime incident.
- **Description**: Short Summary of the crime incident.
- **DayOfWeek**: Day within the week.
- **PdDistrict**: Police Department District ID that assigned to address the crime.
- **Resolution**: Respective penalization result of the crime.
- **Address**: Detailed position of the crime incident.
- **X**: Longitude of the crime location.
- **Y**: Latitude of the crime location.

B. Forecast Model Selection

After initially observing and analyzing the data properties, it can be concluded that the prediction task here can be seen as a multi-class classification task since each crime record has 9 attributes in total. In order to preliminary evaluate the prediction accuracy of each algorithm, we use the mentioned algorithms to predict the crime type with fewer attributes and apply a log loss evaluation model to evaluate the prediction accuracy. The results are presented in the following table.

TABLE I: Logloss Value of Each Classification Algorithm

Algorithm	Log Loss Quantity
Naive Bayes	2.5816
Random Forest	2.9612
XGBoost	2.9058
Light-GBM	2.9116

The log loss quantity in the table represents the difference extent between the predicted value and actual value, which means a model with a lower log loss value would have higher prediction accuracy. Therefore, the Naive Bayes model should have a better prediction effect than others under such circumstances. However, the characteristics of Naive Bayes can not suit the multi-dimensional prediction requirement in this task. Furthermore, the core conditional independence assumption of Naive Bayes classification theory is that the attributes in the sample are all mutual individuals. In our prediction task, our research purpose also includes studying the influence of each attribute on the final crime occurrence possibility. And the other classification algorithm also has lower prediction accuracy than the LightGBM model. Base on the mentioned aspects, we choose the LightGBM algorithm as the prediction model.

C. LightGBM Description

LightGBM, short for Light Gradient Boosting Machine, is an open-source distributed gradient boosting framework for machine learning, which originally developed by Microsoft. It is derived from the decision tree algorithms and mainly used for ranking, classification tasks [12]. LightGBM speeds up the training process of conventional GBDT by up to over 20 times while achieving almost the same accuracy [13].

Compared to the pre-sorted algorithm of the XGBoost Model which is another popular boosting algorithm, the LightGBM model innovatively put forward the Histogram-based Algorithm to deal with node division issues. The XGBoost Model commonly sorts the value of each feature to determine the best split point before training, which is quite time-consuming. However, the Histogram-based Algorithm will firstly pack the characteristic values into bins. Especially during the node split procedure, we only need to calculate the number of bins rather than the original size of the data.

Therefore, it is able to save the usage of memory, reduce the amount of calculation of integral gain and accelerate the processing speed of histogram deduction. The Histogram-based Algorithm applied in our experiment is modified from the work of Ke et. al. [13]. The pseudocode of the Histogram-based algorithm is shown below.

```
Algorithm 1: Histogram-based Algorithm
  Input: I: crime data, d: max depth, m: crime features
  Output: output result
nodeSet ← $\{0\}$ tree nodes in current level
rowSet ← $\{\{0, 1, 2, ...\}\}$ data indices in tree nodes
for i=1 to d do
    for n in nodeSet do
        usedRows ← rowSet[n]
        for k=1 to m do
            newH ← new Histogram()
            Build histogram
            for row in usedRows do
                bin ← I.f[k][row].bin
                newH[bin].y ← newH[bin].y + I.y[row]
                newH[bin].n ← newH[bin].n + 1
            end
            Find the best split on histogram newH.
        end
    end
    Update rowSet and nodeSet at the best split points.
end
```

D. Performance Evaluation Method

To evaluate the accuracy degree of the prediction results, Log Loss Model can be applied by comparing the predicted crime rate with the expected value. Log Loss quantifies the accuracy of a classifier by penalizing false classifications. Therefore, the algorithm can evaluate the difference degree between the predicted value and actual value.

In order to calculate the logarithmic losses, the classifier must provide the probability of belonging to each category of input values, not just the most likely category. The calculation form of the Log Loss Model can be expressed as:

$$L(Y, P(Y \mid X)) = -\log P(Y \mid X) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{\bar{i}\bar{j}} \log \left(p_{\bar{i}\bar{j}}\right) \tag{1}$$

Where Y is the output variable, X is the input variable, L stands for the loss function, N is the input sample size, M is a possible number of categories. y_{ij} represents binary indicators, which indicate whether input instance x_i belongs to category y. And P_{ij} is the possibility that input instance x_i belongs to the category j predicted by the model or classifier.

IV. Experimental Work

A. Experiment Environment

The training environment is as follows: CPU: Intel Core i7-9750, GPU: Nvidia GeForce GTX 1650, System: Windows10.

B. Criminal Data Processing

After observing the data, we initially use the pandas to delete duplicated records. There are also some records that do not have exact longitude and latitude. For these cases, their default longitude and latitude are set to -120.5 and 90.0 respectively. Moreover, since there is only 1-time relevant attribute: Dates in the original data, it is not enough to build the model, especially for the time dimension. Therefore, we split the 'Dates' attribute in the original data into 3 attributes: Hour, Minute, and year.

Criminal temporal patterns are complicated since temporal resources could be structured in various intervals like weeks, months, seasons, years and others [14]. Considering the annual variation of crime will increase the complexity in our research, the year attribute and other irrelevant variants, such as Description, Resolution are all been deleted. All the included features build the model are listed in Table II.

C. Feature Construction

In the real world, the location, time, and block are connected with each other and account for a certain proportion to the final prediction output. Initially, we use a label encoder to encode the classification attribute value which is the category of the crime. Furthermore, the weight of each element can also be calculated by the Permutation Importance calculation method and visualized by the show weights method in eli5 [1] package of python.

TABLE II: Weight of Each Feature in the LightGBM Model

Feature	Weight
Y	0.0614 ± 0.0003
Minute	0.0590 ± 0.0011
X	0.0511 ± 0.0005
PdDistrict	0.0266 ± 0.0004
Block	0.0208 ± 0.0009
Hour	0.0165 ± 0.0008
Day	0.0141 ± 0.0008
Weekday	0.0021 ± 0.0003
Month	0.0014 ± 0.0004

D. LightGBM Model Training

In the training process, there are some parameters required to be specified to control the learning algorithm. The parameter: boosting is set as 'gbdt' (traditional Gradient Boosting Decision Tree) and the parameter objective is set as the 'multi-class' (multi-class classification application) which are

[1] https://eli5.readthedocs.io/en/latest/

the most stable and efficient value to the crime occurrence possibility prediction.

For the learning control parameters such as maximum bin (max number of bins that feature values will be bucketed in), learning rate, minimum data in leaf (Minimum quantity of data in one leaf, used to deal with over-fitting), etc, it is hard to decide the best set of parameters that would achieve the best output. Therefore, we apply the Bayesian Optimization function in the bayesian optimization library [2], which is a classical optimization library in python.

The working principle of Bayesian Optimization is to calculate the log-loss value of each model with different input variables. According to the previous log-loss value, the parameter in the next case will change on a different dimension. This iteration process will continue until the log-loss value will not be improved anymore for 5 rounds. In this experiment, the final best-optimized log-loss value is **2.265498428318** and the most suitable parameters are shown in the table below.

TABLE III: Parameters of the LightGBM Model

boosting	'gbdt'
objective	'multiclass'
learning_rate	0.2355
max_bin	354
max_delta_step	0.6425
min_data_in_leaf	41
num_class	39
num_leaves	48

E. Crime Analysis and Visualization

Finally, to view the geographic distribution of criminal cases more directly, a real geographic map marked with respective crime information will be quite helpful. However, there are 878,050 criminal records in total. If we mark all the criminal cases with related geographic information, the graph will be crowded with dots that can not suit the expectation. Therefore, it would be suitable to apply cluster selection visualization techniques under such circumstances. We divide the data into 200 clusters for every district within the 20 years period and select the place where most crimes happen.

V. EXPERIMENTAL RESULTS

A. Prediction Results Analysis

As indicated in the previous section, each factor will contribute to the final probability result. However, it is ambiguous the extent of the specific feature will affect the prediction output. Then, we apply Partial Dependency Plots [3] to help with this issue and verify the consistency of the model with our expectation. Figure 1 shows each crime type's dependency of appearance possibility on the district, in another way, different districts will have various influences on specific crime types. Here, we choose 3 typical crime types, which are kidnapping (class15), larceny (class 16), and liquor laws (class 17), to show their respective partial dependency plots.

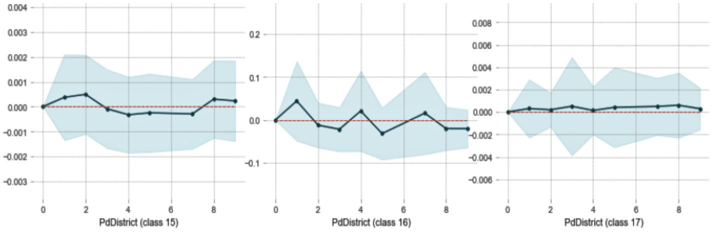

Fig. 1: Partial Dependency Plots of Districts

Consequently, the calculated crime occurrence possibility and the influence of each feature can be visualized by shap [4] in python. The following graph is a typical prediction example, which means the occurrence possibility of theft is approximately 14% on March 29, Sunday at 18:00 in BAYVIEW district with longitude: -122.395251, latitude: 37.755344.

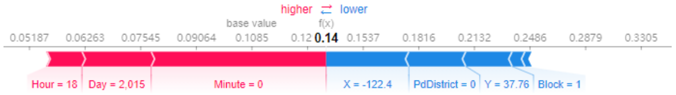

Fig. 2: Occurrence Possibility of Larceny

In order to validate the prediction accuracy, we initially use fbprophet [5] to produce the crime trend and its rational variation range base on actual records. Then, we delete the crime type feature of actual records and use the LightGBM prediction model to generate the base on the resting features.

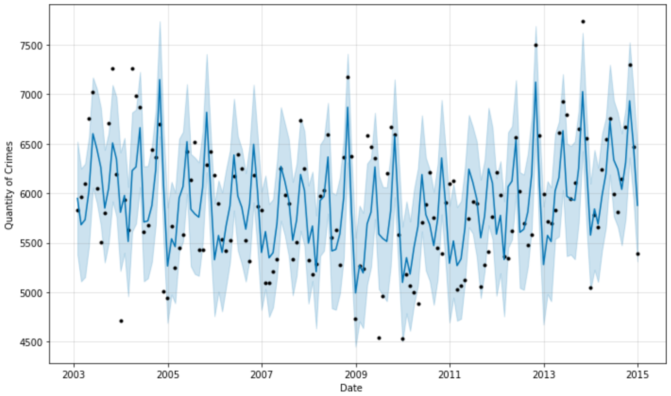

Fig. 3: Predicted and Actual Crime Quantity Variation Graph

If the occurrence possibility of all crime types is below 0.1, no crime will happen under the condition. Otherwise, the crime type with the highest probability will be set as the crime type. After generating the prediction results of all actual records, the

[2] https://github.com/fmfn/BayesianOptimization

[3] https://scikit-learn.org/stable/modules/partial$_{d}ependence.html$

[4] https://shap.readthedocs.io/en/latest/generated/shap.plots.force.html

[5] https://pypi.org/project/fbprophet/

total quantity of monthly crime can be counted and marked on the plot with black dots. Finally, as shown in figure 3, the prediction results mostly locate in the rational range and suit the actual crime variation trend.

However, the black points represent predicted values that are not entirely located in the valid range. There are many reasons that may lead to this error. The typical one is that the histogram algorithm can not find split points accurately and may lead to an over-matching effect. Additionally, the boosting algorithm, which LightGBM belongs to, is a kind of iterative algorithm. Every iteration will adjust the weight of the sample database on the forecast results that come from the last iteration. This iteration process is quite sensitive to the noise, such as feature errors and noisy labels in data, and will finally affect the prediction accuracy.

B. Crime Trend Variation Analysis

After cleaning the data, the variation of crime rate can be visualized daily, weekly, and annually in the line chart. It can be concluded that the general crime variation trend emerges to descend. However, the highest criminal case frequency can be obtained in 2001 and 2002, during the time range when the 911 terrorist attack happens, which reaches the crime rate peak. In order to study the crime frequency variation within a month or a day, we decide to plot the variation trend of each crime trend. However, there are 39 crime types in total, which can not be clearly displayed in one graph. Therefore, we select the top 6 crime type with the most crime amount.

As shown in figure 4, the monthly variation trend of all 6 crimes is almost the same while the assault has the highest amount within all the crime types. They increase slightly from January to May and then decrease to a stable range from June to September. In October, the crime amount reaches the highest within the year.

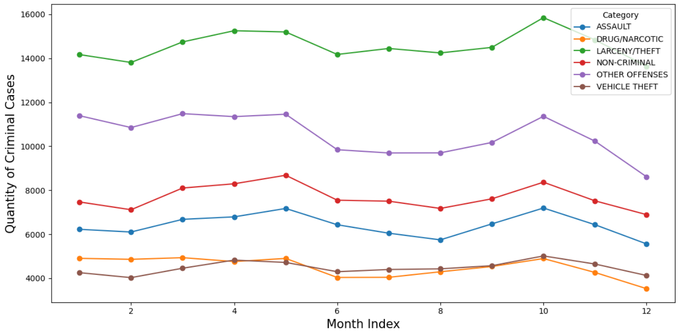

Fig. 4: Monthly Top 6 Crime Count from 2001 to 2020.

Figure 5 shows the hourly amount variation of the top 6 crime type cases within one day. As indicated in the graph, the crime quantity of all the top 6 crime types reaches the lowest point at 5 A.M. when people are sleepy even for the criminals. After the bottom point, the crime trend increases greatly until 12 A.M. when reaches its first peak. Then, after a slight decrease from 12 A.M. to 1 P.M., almost all the crime types present a smooth decrease except assault which has the highest crime amount. The assault keeps going up until 6 P.M. when it reaches its peak point.

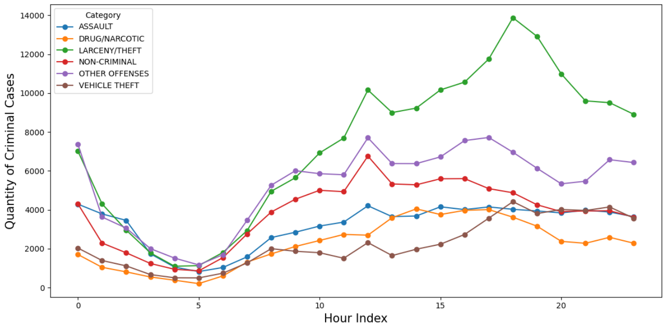

Fig. 5: Hourly Top 6 Crime Count from 2001 to 2020.

C. Geographic Distribution of Crimes

In order to explore the relationship between geographic location and crime frequency, related data has been presented on the geographic map. Firstly, we use pandas [6] to count the respective total crime quantity of each district from 2001 to 2020, which is shown in the table below.

TABLE IV: Total Crimes of each District from 2001 to 2020

District	Total Crime Quantity
BAYVIEW	89,431
CENTRAL	85,460
INGLESIDE	78,845
MISSION	119,908
NORTHERN	105,296
PARK	49,313
RICHMOND	45,209
SOUTHERN	157,182
TARAVAL	65,596
TENDERLOIN	81,809

In figure 6, we mark the selected crime with respective amplitude and longitude on the geographic map. The background map is a real geographic map of San Francisco with a border between districts labeled. Additionally, all the criminal cases indicate on the graph are typical crimes selected within their respective cluster that can typically represent other cases.

In figure 7, the crime frequency of each district has been displayed in the form of a spatial heat map. We created the heat map which could reflect the crime frequency distribution. The basic mechanism of the spatial heat map is that position of a magnitude is forced by its corresponding location in that space. As indicated inside the graph, the darker the color is, the larger the crime occurrence frequency there would be.

It can be concluded from figure 6 and figure 7 that the northeast area of San Francisco has the highest crime rate. The

[6] https://pandas.pydata.org/

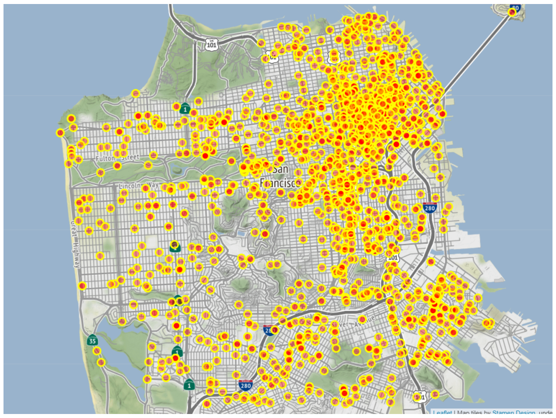

Fig. 6: Mark Crime Occurrence Location

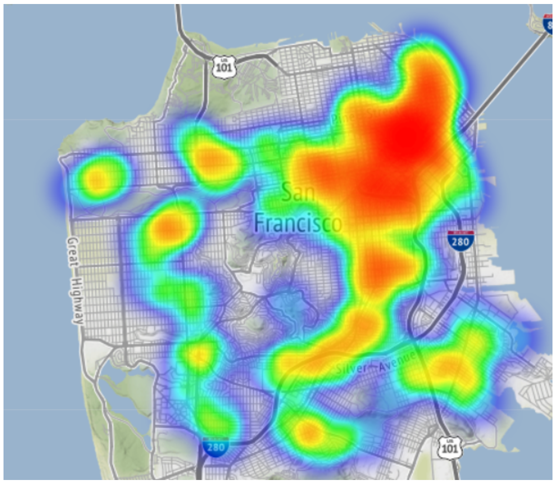

Fig. 7: Geographic Crime Heat Map of San Francisco

districts included are district 6, district 8, and part of District 9, where the downtown, historic scenic spots, and supermarkets locate. Especially district 6, which firstly known as the "Western Addition", is vibrant and filled with shopping, restaurants, and some of the city's oldest homes. Furthermore, according to the economic analysis report released by the Office of Economic and Workforce Development in San Francisco, the three districts with the highest crime quantity also has good economic development status.

Conclusion and Future Work

In this study, the crime data of San Francisco from 2001 to 2015 has been used for the prediction and analysis. We have carried geographical distribution analysis, annual trend analysis, and occurrence possibility prediction related to crime. Especially for the appearance possibility, Light Gradient Boosting Machine Model was trained to forecast crime possibility in the future and validate by forecasting crime type of validation dataset. In terms of exploring the amount that each feature contributes to the prediction crime possibility, LightGBM is quite effective and accurate compared to other classification models. Because the dimension of the current dataset is not quite enough, the prediction result can not be further improved. In the future, we will try to consider more dimensions relate to crime, such as the local education condition, population constitution and public security policy.

Acknowledgment

This study is partially supported by the AI University Research Centre (AI-URC) through the XJTLU Key Program Special Fund (KSF-P-02) and KSF-A-17.

References

[1] S. Sathyadevan, et al., Crime analysis and prediction using data mining, in: 2014 First International Conference on Networks & Soft Computing (ICNSC2014), IEEE, 2014, pp. 406–412.

[2] T. Hu, X. Ye, L. Duan, X. Zhu, Integrating near repeat and social network approaches to analyze crime patterns, in: 2017 25th International Conference on Geoinformatics, IEEE, 2017, pp. 1–4.

[3] Y. FAN, T. YANG, G. JIANG, L. ZHU, R. PENG, Identifying criminals' interactive behavior and social relations through data mining on call detail records, DEStech Transactions on Computer Science and Engineering (aiea) (2017).

[4] M. A. Tayebi, U. Glasser, Investigating organized crime groups: A social network analysis perspective, in: 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, IEEE, 2012, pp. 565–572.

[5] H. Isah, D. Neagu, P. Trundle, Bipartite network model for inferring hidden ties in crime data, in: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, 2015, pp. 994–1001.

[6] T. Diviák, et al., Sinister connections: How to analyse organised crime with social network analysis?, Acta Universitatis Carolinae Philosophica et Historica 24 (2) (2018) 115–135.

[7] Q. Jiang, J. J. S. Barricarte, A crime rate forecast and decomposition method, International Journal of Criminology and Sociological Theory 4 (2) (2011).

[8] M. S. Vural, M. Gök, Criminal prediction using naive bayes theory, Neural Computing and Applications 28 (9) (2017) 2581–2592.

[9] M. A. Awal, J. Rabbi, S. I. Hossain, M. Hashem, Using linear regression to forecast future trends in crime of bangladesh, in: 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), IEEE, 2016, pp. 333–338.

[10] S. Kasim, H. Hafit, N. P. Yee, R. Hashim, H. Ruslai, K. Jahidin, M. S. Arshad, Cmis: Crime map information system for safety environment, in: IOP Conference Series: Materials Science and Engineering, Vol. 160, IOP Publishing, 2016, p. 012096.

[11] H. Chen, H. Atabakhsh, T. Petersen, J. Schroeder, T. Buetow, L. Chaboya, C. O'Toole, M. Chau, T. Cushna, D. Casey, et al., Coplink: Visualization for crime analysis, in: Proceedings of the 2003 annual national conference on Digital government research, 2003, pp. 1–6.

[12] Z. Chu, J. Yu, A. Hamdulla, Lpg-model: A novel model for throughput prediction in stream processing, using a light gradient boosting machine, incremental principal component analysis, and deep gated recurrent unit network, Information Sciences 535 (2020) 107–129.

[13] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, Advances in neural information processing systems 30 (2017) 3146–3154.

[14] K. Leong, A. Sung, A review of spatio-temporal pattern analysis approaches on crime analysis, International E-Journal of Criminal Sciences 9 (2015) 1–33.