บทความวิจัย (Research Article)

Crime Prediction Through Collaborative Analysis of Proximate Police Stations Data

Siraswaya Hongyon and Prabhas Chongstitvatana*

Department of Computer Engineering, Chulalongkorn University
* Corresponding author
prabhas.c@chula.ac.th

Received: 18 Feb 2024
Revised: 30 Apr 2024
Accepted: 30 Apr 2024

Abstract

Crime prediction is a crucial aspect of law enforcement strategies and crime prevention efforts. Machine learning has emerged as a valuable tool in crime prediction, allowing for more accurate and data-driven forecasting. In this study, we focus on forecasting the number of crimes at Pathumwan Police Station in Thailand. Utilizing criminal records from various police stations across Thailand, we employ the K-Means clustering algorithm to group police stations exhibiting similar crime patterns to the Pathumwan Police Station. The clustering results reveal that Wang Thonglang, Nang Loeng, Dusit, Bang Sue, Thung Maha Mek, Samre, Sutthisan, Pak Khlong San, Bangkok Yai, Bangkok Noi, Makkasan, Bang Yi Ruea, and Talat Phlu Police Station are clustered together with Pathumwan Police Station. Subsequently, we apply the Long Short-Term Memory (LSTM) model to forecast crimes at Pathumwan Police Station. The training dataset comprises paired data from police stations within the same cluster as Pathumwan Police Station. Our findings indicate that the optimal results, the value of Pathumwan and Din Daeng pair works best. Although it is a station that is not in the same cluster as Pathumwan station, it has the lowest RMSE and MAE values and the smallest P-value. This indicates the statistical significance of the results. This collaborative approach enhances the accuracy of crime prediction models and contributes to more effective law enforcement strategies.

Keywords: Crime, Machine Learning, K-Means, LSTM, Long-Short Term memory, Prediction, Forecasting

1. Introduction

The issue of crime poses a significant threat to the well-being and tranquility of Thai society. Crimes ranging from property offenses to violent acts against individuals continue to challenge the peace and security of our communities. The causes underlying criminal behavior are multifaceted and complex. Crime stands as a critical social issue that evolves in tandem with the advancement of society and technology. The more our society flourishes and expands, the more

crime tends to thrive. It is a challenge that society cannot entirely evade. In urbanized areas with dense populations, the consequences of the crime problem become even more pronounced. The toll on society is considerable, as it leads to substantial losses and necessitates increased efforts for prevention and law enforcement. Incidents of crime occurring in various regions of Thailand vary significantly, influenced by numerous factors such as economic conditions, unemployment rates, poverty levels, cultural degradation, and more. Consequently, law enforcement agencies face the challenge of allocating limited resources efficiently to areas that may experience a higher prevalence of criminal activities during specific periods.

This research proposes the utilization of Long Short-Term Memory (LSTM) for the analysis and prediction of criminal activities. LSTM is well-suited for time-series data, making it a powerful tool for analyzing and forecasting crime patterns, particularly in areas where crime incidence is expected to vary over time. By incorporating technical and criminological factors as initial variables, this research aims to enhance the efficiency of crime classification and provide insights into the patterns of criminal behavior within different regions. This will enable more effective planning and response to incidents during various time intervals, ultimately contributing to improved public safety and security.

2. Related Work

Numerous studies have been conducted for crime prediction. One of the most used method is machine learning. [1] is a comparative study of machine learning methods for crime prediction and demonstrated the superior performance of LSTM model. [2] uses LSTM, along with IoT data and criminal history data for the city of Chicago, to classify public safety crime incidents through police analytics. [3] uses LSTM with data from Addis Ababa police department, to forecast crime and examine correlations over time, crime types and locations. [4] uses LSTM and Neural Network models to analyze and identify crime patterns and crime incident trends. Police data contains many features. The steps to clean data and select features are studies in [8,0]. The comparison of LSTM method with various other methods such as logistics regression, SVM, Naive Bayes, KNN, decision tree, MLP, random forest, and XGBoost are reported [6, 7, 10, 11, 13, 16, 17, 18, 20, 21]. Many other studies use machine learning for crime prediction and crime data analytic [12,14,15,16,18,19]. These insights help police departments comprehend crime problems and facilitate the prediction of future events.

3. Method

3.1 Data Processing

We utilized crime data sourced from the Royal Thai Police, encompassing cases reported across all police stations in Thailand from January 2011 to February 2022. The dataset comprises a total of 478,000 cases distributed among 1,411 police stations. To enhance relevance, we narrowed our focus to 30 police stations in Bangkok, specifically those centrally located in the capital.

Through meticulous selection, we have identified 13 police stations exhibiting crime incidence patterns comparable to that of Pathumwan Police Station. These stations encompass Wang Thonglang, Nang Loeng, Dusit, Bang Sue, Thung Maha Mek, Samre, Sutthisan, Pak Khlong San, Bangkok Yai, Bangkok Noi,

Makkasan, Bang Yi Ruea, and Talat Phlu Police Station. Data from these stations were aggregated to construct a consolidated dataset representing the equivalent level of criminal activity. However, 16 police stations were found not to be in the same cluster, including Lumpini, Din Daeng, Khlong Tan, Huai Khwang, Thonglor, Phaya Thai, Bangrak, Tao Poon, Somdet Chao Phraya, Chana Songkhram, Samranrat, Royal Palace, Yannawa, Tha Phra, Samsen, and Wat Phraya Krai.

In the preprocessing phase, any missing data was addressed, and null values were replaced with zeros. Additionally, to facilitate uniformity and comparability, the data was normalized to a scale ranging from 0 to 1. This standardization process ensures that the various features contribute uniformly to the model, promoting effective training and prediction within LSTM.

3.2 Clustering Police Station

Using the Elbow curve method to select the number of clusters with the optimal k value. It is calculated from the Inertia value. This method is a measure of the error of the sum of the distances between the Object and the Centroid of the Sum of square, called Within-Cluster-Sum-of-Squares (WCSS).

After we determined the K values, we used the K-Means method for clustering. K-means is a method for finding the number of clusters without labels, which is called Unlabeled Data. This method is an Unsupervised Learning method. The main function of K-means is a type of clustering. This type of clustering is based on statistics consisting of two or more groups of data. K-means is an iterative process of assigning each data point to a group. The data points will gradually increase. Grouped together based on similar features to minimize the sum of distances between data points and cluster centroids. to identify the correct group of each data point. After that, we divided the data into K clusters and assigned the mean to each cluster. Data points are placed in the group closest to that group's mean. By calculating the distance between the data point and the centroid, the Euclidean distance metric is used.

$$C_k = \frac{1}{N_k} \sum_{i=1}^{n_k} x_i \tag{1}$$

Equation 1. Equation to find the centroid of k-means.

$$D(x_i, C_k) = \sqrt{\sum_{j=1}^{d} (x_{ij} - C_{kj})^2} \tag{2}$$

Equation 2. Equation to Euclidean distance metric to calculate the distance between the data point and the centroid.

3.3 Forecast Crime Cases

Long Short-Term Memory is a technique that was developed from Recurrent Neural Network (RNN). RNN's working principle is to use the output obtained from calculations from the previous node as input for the next node. Each node of RNN has 2 parts of incoming data: Input data of that node and Output that has been calculated from the previous node. The two sets of data that come into the node will be combined. The results are separated into 2 parts: the results obtained from that node and the results that will be used as input data for the next node. The RNN technique is suitable to be used with data that has Sequence or continuous data such as time series data, audio data, text data, image, and video data, etc.

In our architecture, we employ a multi-layered Long Short-Term Memory (LSTM) network for time series analysis. The input layer is configured as (8, 1), where 8 denotes the number of time steps, and 1

denotes the number of features in each step. The first LSTM layer is comprised of 128 units and utilizes the rectified linear unit (ReLU) activation function, indicating that this layer outputs data sequentially. Subsequently, the second LSTM layer is configured with 50 units and employs the ReLU activation function. To mitigate overfitting. A Dropout layer is incorporated with a dropout rate set at 0.2, and a Dense Layer (output layer) with 1 unit uses the ReLU activation function. The model is trained on the X_train and y_train data over 400 epochs with a batch size of 8.

This architectural configuration is designed to capture temporal dependencies within the data, allowing for a robust representation of sequential patterns. The choice of activation functions and the use of dropout layers contribute to the model's ability to generalize well to unseen data, enhancing its overall performance and reliability.

3.4 Evaluation Metrics

To assess the performance of LSTM predictions, we use the root mean square error (RMSE) and mean absolute error (MAE) when comparing the forecasted results against actual values.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Actual_i - Predicted_i)^2} \qquad (3)$$

Equation 3. Equation of root mean square error (RMSE)

The Root Mean Squared Error (RMSE) shares the same unit as the actual and predicted values, simplifying their comparison. It gauges the average proximity of predictions to the true values, with a notable characteristic of assigning more significance to larger errors than smaller ones. RMSE stands out as a prevalent metric in regression due to its ease of differentiation and compatibility with gradient-based optimization methods.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Actual_i - Predicted_i| \qquad (4)$$

Equation 4. Equation of mean absolute error (MAE)

The MAE is computed by averaging the absolute differences between the predicted and actual values. It is advantageous over MSE and RMSE as MAE is less sensitive to outliers. Additionally, since MAE calculates based on absolute differences, it gauges how far, on average, the predictions deviate from the actual values. This makes it a useful metric for interpreting the overall effectiveness of a model.

The calculation involves taking the absolute difference between each predicted value and its corresponding actual value, dividing this difference by the actual value, and then averaging the resulting percentages. The overall metrics, RMSE and MAE, are presented by taking the average percentage difference across all values. Both RMSE and MAE are expressed in a way that lower values indicate better performance. During model training, we utilize MAE as the loss function. To assess performance, we employ RMSE and MAE metrics.

4. Result and Discussion

4.1 Clustering Police Station

We selected the K values for the grouping using the elbow method, choosing from a value of 2 to 50. After that, we compared the Inertia values and the percentage differences in the K values obtained from the elbow method, after applying K-Means clustering,

the results indicated that all 13 police stations were grouped into the same cluster as the Pathumwan station. These stations include Wang Thonglang, Nang Loeng, Dusit, Bang Sue, Thung Maha Mek, Samre, Sutthisan, Pak Khlong San, Bangkok Yai, Bangkok Noi, Makkasan, Bang Yi Ruea, and Talat Phlu Police Station, as shown in Fig 1.

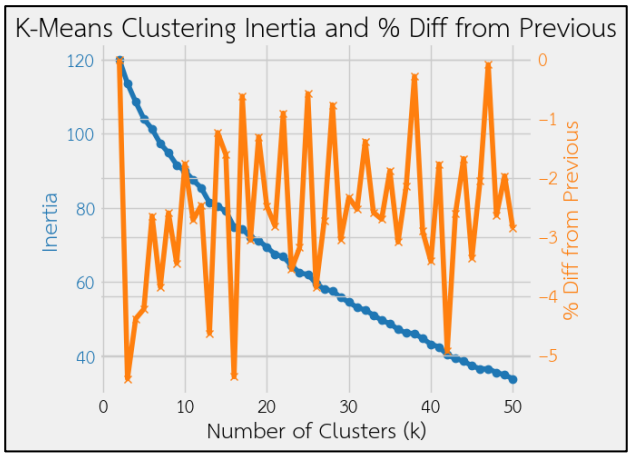

Fig. 1. Number of clusters using elbow method.

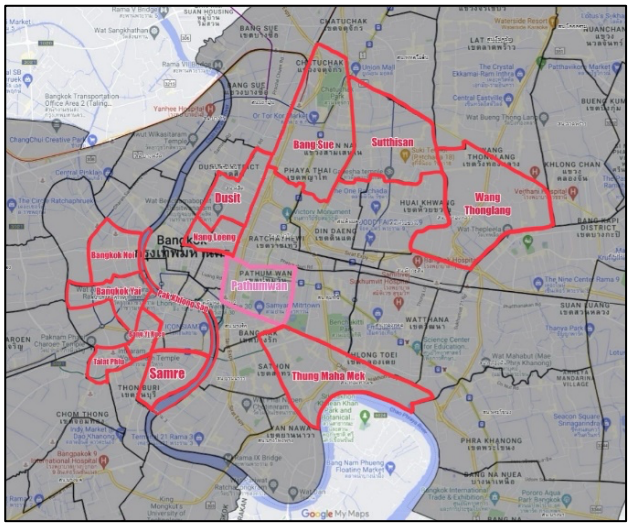

Fig. 2. Map shows police stations selected for this study.

Fig 2 is a map showing the location of police stations obtained from clustering. The pink color shows Pathumwan Station. The red show stations are in the same cluster as Pathumwan Station. And finally, the blue color show stations that are not in the same cluster as Pathumwan Station from the 30 stations selected for this study.

4.2 Compare Police stations

Once we obtain the results from the K-Means algorithm, we find that there are 13 police stations grouped in the same cluster as Pathumwan Police Station. Fig. 3 illustrates a graph comparing the stations within the same cluster as Pathumwan Station, totaling 14 stations.

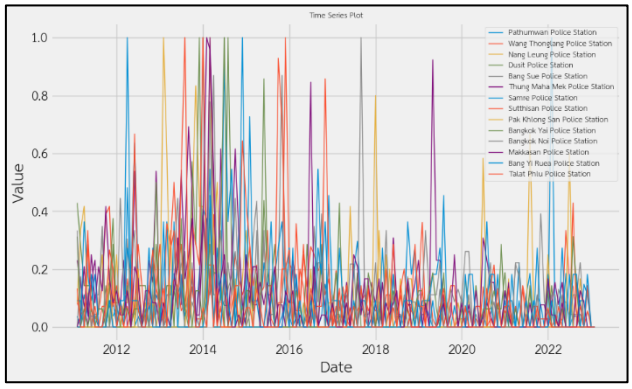

Fig. 3. Graph shows stations in the same cluster as Pathumwan.

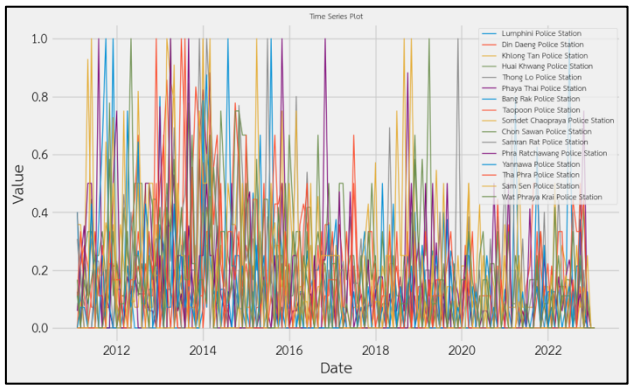

Fig. 4. Graph shows stations not in the same cluster as Pathumwan.

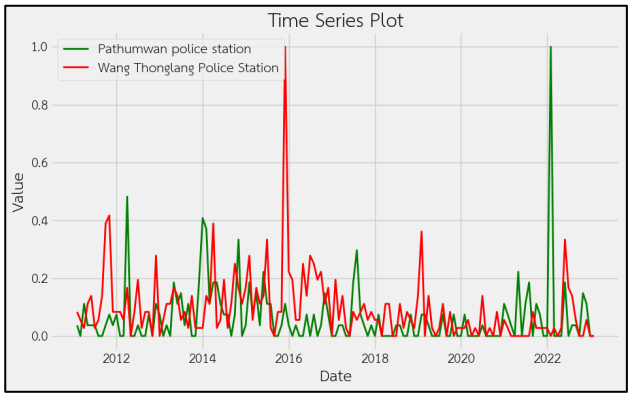

Fig. 5. Data between Pathumwan and in the same cluster - Wang Thonglang and Pathumwan

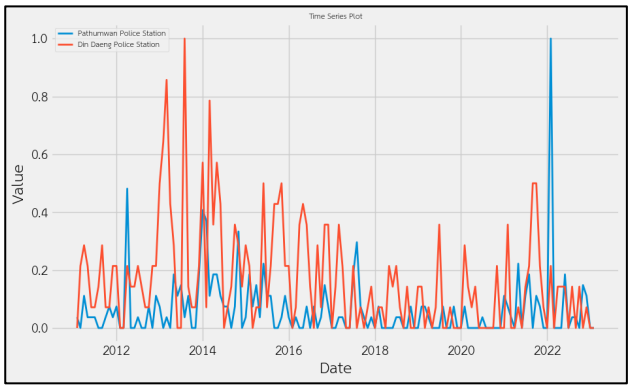

Fig. 6. Data between Pathumwan and not in the same cluster - Din Daeng and Pathumwan

4.3 Forecast Result

Tables 1 and 2 present the LSTM results from all 30 police stations, averaging over 30 trials in RMSE, MAE, and calculate T-Test values to compare the use of Pathumwan Police Station data with all other pairs of police stations. Table 1 shows stations that are in the same cluster and Table 2 shows stations that are different from Pathumwan station. Results marked with an asterisk (*) on the numbers indicate that the mean values are significantly different at a 0.05 significance level compared to using data solely from Pathumwan station for the same metrics, and double asterisks (**) Specifies that the best pair value will be the one with the lowest RMSE and MAE along with the smallest P-value.

Table 1 The result from average of 30 experiments of metric RMSE, MAE and P-Value from LSTM using Pathumwan data and all other pairs in the same cluster.

Police Station	RMSE	P-Value	MAE	P-Value
Pathumwan	7.45047	-	3.26355	-
Pathumwan + Wang Thonglang	5.47886*	0.025	2.57099*	0.011
Pathumwan + Nang Loeng	5.53164*	0.029	2.49437*	0.005
Pathumwan + Dusit	5.38820*	0.019	2.40720*	0.002
Pathumwan + Bang Sue	6.96228	0.766	3.17512	0.865
Pathumwan + Thung Maha Mek	8.30568	0.512	3.76558	0.292
Pathumwan + Samre	8.10871	0.651	3.91762	0.244
Pathumwan + Sutthisan	7.73993	0.862	3.19942	0.898
Pathumwan + Pak Khlong San	5.47109*	0.024	2.41428*	0.002
Pathumwan + Bangkok Yai	6.53950	0.355	3.31786	0.870
Pathumwan + Bangkok Noi	5.47559*	0.027	2.67149*	0.034
Pathumwan + Makkasan	5.60690*	0.036	2.73387	0.051
Pathumwan + Bang Yi Ruea	6.71162	0.405	3.47311	0.467
Pathumwan + Talat Phlu	5.40944*	0.020	2.39431*	0.002

Combining Wang Thonglang and Nang Loeng with Pathumwan yielded better results in terms of both

RMSE and MAE with a P-value of less than 0.05, while combining Dusit and Pak Khlong San with Pathumwan also yielded good results. with a P-value that is less than 0.05, while combining Bang Sue, Thung Maha Mek, Samre, Sutthisan, Bangkok Yai, Bangkok Noi, Makkasan, Bang Yi Ruea, and Talat Phlu with Pathumwan does not increase the RMSE and MAE values were better and has a high P-value.

Table 2 The result from average of 30 experiments of metric RMSE, MAE and P-Value from LSTM using Pathumwan data and all other pairs in the different cluster.

Police Station	RMSE	P-Value	MAE	P-Value
Pathumwan + Lumpini	6.19521	0.166	3.12425	0.645
Pathumwan + Din Daeng	5.05408**	0.007**	2.44854**	0.003**
Pathumwan + Khlong Tan	11.51808	0.057	4.88956	0.012
Pathumwan + Huai Khwang	5.50326*	0.027	2.33035*	< 0.001
Pathumwan + Thonglor	6.82482	0.558	2.96476	0.386
Pathumwan + Phaya Thai	5.47648*	0.025	2.57710*	0.014
Pathumwan + Bangrak	5.42294*	0.023	2.53248*	0.009
Pathumwan + Tao Poon	5.40075	0.2	2.58332	0.013
Pathumwan + Somdet Chao Phraya	5.74073	0.054	2.60970	0.019
Pathumwan + Chana Songkhram	7.26085	0.878	3.33571	0.872
Pathumwan + Samranrat	5.55140*	0.03	2.46593*	0.004
Pathumwan + Royal Palace	5.41842*	0.021	2.25912*	<0.001
Pathumwan + Yannawa	9.25087	0.586	3.46260	0.781
Pathumwan + Tha Phra	5.40776*	0.02	2.37490*	0.001
Pathumwan + Samsen	6.38167	0.258	3.13439	0.719
Pathumwan + Wat Phraya Krai	13.39352	0.046	4.49558	0.108

Combining Din Daeng, Khlong Tan, Huai Khwang, Phaya Thai, Bangrak, Tao Poon, Samranrat, and Royal Palace with Pathumwan had better results with a P-value of less than 0.05, while combining Lumpini, Thonglor, Somdet Chao Phraya, Chana Songkhram, Yannawa, and Tha Phra and Pathumwan did not improve the RMSE and MAE values and had high P-values.

The LSTM model outcomes suggest that integrating data from police stations paired with Pathumwan stations enhances prediction accuracy, as evidenced by reductions in RMSE and MAE, accompanied by statistically significant P-values. The optimal pairing is determined by the pair exhibiting the lowest RMSE and MAE, along with the smallest P-value. According to the dataset provided, the most favorable pairing is between Pathumwan and Din Daeng, yielding an RMSE of 5.05409 and MAE of 2.44855, with corresponding P-values of 0.007 and 0.003 for RMSE and MAE, respectively. Despite Din Daeng not belonging to the same cluster as Pathumwan Station, this pairing yields the most favorable outcomes, as evidenced by the lowest RMSE and MAE values and the smallest P-value, signifying the statistical significance of the results.

When assessing the outcomes across different police stations, certain combinations of data yield statistically significant improvements, with p-values below 0.05, while others result in mixed or less favorable outcomes. Such discrepancies may stem from limitations in the available data, particularly regarding the diversity of case types and the specificity of crime scene coordinates. To overcome these limitations and enhance the model's efficacy, comprehensive guidelines for data development are imperative.

To enhance the model's capabilities, Additional information must be requested from other agencies. This effort aims to acquire more detailed information regarding the types of cases and specific coordinates of crime scenes. Such additional data would enable the model to better learn and understand the unique characteristics and distinctions associated with each case, as well as the locations where incidents are likely to occur.

This approach is crucial for refining the model's comprehension of crime patterns, ultimately leading to improvements in predictive accuracy. The incorporation of diverse and comprehensive datasets, obtained through collaboration with various agencies, is anticipated to contribute to the robustness and effectiveness of the crime prediction model.

5. Conclusion

We utilized K-means clustering to identify clusters of police stations with similar patterns in criminal cases, specifically selecting those grouped with Pathumwan Station. The data from these selected stations was then combined with that of the Pathumwan Station using LSTM as a machine learning method. The analysis revealed that the optimal results, the value of Pathumwan and Din Daeng pair works best as it has the lowest RMSE and MAE values and the smallest P-value. This indicates the statistical significance of the results. Even though it is a station that is not in the same cluster as Pathumwan Station.

6. References

[1] Zhang, Xu, Lin Liu, Luzi Xiao, and Jiakai Ji. "Comparison of machine learning algorithms for predicting crime hotspots," *IEEE Access*., Vol. 8, pp. 181302-181310, 2020. doi: 10.1109/ACCESS.2020.3028420.

[2] Mohammadi, M., and A. Al-Fuqaha. "'Predicting incidents of crime through LSTM neural networks in smart city domain". *International Conference on Smart Cities, Systems, Devices and Technologies. 28 Jul-2 Aug. Nice, France* : pp. 32-37. 2019.

[3] Meskela, Tsion Eshetu, Yidnekachew Kibru Afework, Nigus Asres Ayele, Muluken Wendwosen Teferi, and Tagele Berihun Mengist. "Designing time series crime prediction model using long short-term memory recurrent neural network," *International Journal of Recent Technology and Engineering*., Vol. 9, pp. 402-405, 2020.

[4] Feng, Mingchen, Jiangbin Zheng, Jinchang Ren, Amir Hussain, Xiuxiu Li, Yue Xi, and Qiaoyuan Liu. "Big data analytics and mining for effective visualization and trends forecasting of crime data," *IEEE Access*., Vol. 7, pp. 106111-106123, 2019. doi: 10.1109/ACCESS.2019.2930410.

[5] Gowri, J., and S. Padmaja. "A Survey on Prediction of Risk Related to Theft Activities in Municipal Areas using Deep Learning". *International Conference on Electronics and Renewable Systems. 2-4 Mar.*

Tuticorin India : pp. 1321-1326, 2023. doi: 10.1109/ICEARS56392.2023.10085123.

[6] Greff, Klaus, Rupesh K. Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber. "LSTM: A Search Space Odyssey," *IEEE transactions on Neural Networks and Learning Systems*., Vol. 28 (No. 10), pp. 2222-2232, 2017. doi: 10.1109/TNNLS.2016.2582924.

[7] Jozefowicz, Rafal, Wojciech Zaremba, and Ilya Sutskever. "An empirical exploration of recurrent network architectures". *International Conference on Machine Learning. 6-11 Jul. Lille France* : pp. 2342-2350, 2015.

[8] Bank of Thailand. (1 January 2023). *Financial Institutions and Special Financial Institutions' Holidays*. [Online] Available : https://www.bot.or.th/

[9] Roya Thai Survey Department. (1 February 2023). *Location of the police station*. [Online] Available : https://geoportal.rtsd.mi.th/

[10] Siami-Namini, S. & Siami Namin, A. *Forecasting Economics and Financial Time Series: ARIMA vs. LSTM*. arXiv:1803.06386, 2018.

[11] Siami-Namini, Sima, and Akbar Siami Namin. *Forecasting economics and financial time series: ARIMA vs. LSTM*. arXiv preprint arXiv:1803.06386, 2018.

[12] Wang, Shihuai, and Kunxiaojia Yuan. "Spatiotemporal Analysis and Prediction of Crime Events in Atlanta Using Deep Learning". *International Conference on Image, Vision and Computing. 5-7 Jul. Xiamen China* : pp. 346-350, 2019. doi: 10.1109/ICIVC47709.2019.8981090.

[13] Ullah, Waseem, Amin Ullah, Tanveer Hussain, Zulfiqar Ahmad Khan, and Sung Wook Baik. "An efficient anomaly recognition framework using an attention residual LSTM in surveillance videos," *Sensors.,* Vol. 21 (No. 8), pp. 1-17, 2021.

[14] Himanshi, Himanshi. *Analysing Crime Patterns using Machine Learning: A case study in Chicago*. MSc Research Project. National College of Ireland, 2022.

[15] Mehta, Jay, Vaidehi Vatsaraj, Jainam Jain, and Anant V. Nimkar. "Categorical Crime Rate Analysis and Prediction". *International Conference on Computing Communication and Networking Technologies*. *3-5 Oct. Kharagpur India* : pp. 1-7, 2022. doi: 10.1109/ICCCNT54827.2022.9984368.

[16] Devi, J. Vimala, and K. S. Kavitha. "Automating Time Series Forecasting on Crime Data using RNN-LSTM," *International Journal of Advanced Computer Science and Applications*., Vol. 12 (No. 10), pp. 458-463, 2021. doi: 10.14569/IJACSA.2021.0121051.

[17] Shao, XiuLi, Doudou Ma, Yiwei Liu, and Quan Yin. "Short-term forecast of stock price of multi-branch LSTM based on K-means". *International Conference on Systems and Informatics. 11-13 Nov. Hangzhou China* : pp. 1546-1551, 2017. doi: 10.1109/ICSAI.2017.8248530.

[18] Manengadan, Mufeeda, Silpa Nandanan, and Neethu Subash. "Crime Data Analysis, Visualization and Prediction Using LSTM," *International Journal of Data Science and Analysis*., Vol. 7 (No. 3), pp. 51-59, 2021. doi: 10.11648/j.ijdsa.20210703.11.

[19] Safat, Wajiha, Sohail Asghar, and Saira Andleeb Gillani. "Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques," *IEEE Access*., Vol. 9, pp. 70080-70094, 2021. doi: 10.1109/ACCESS.

[20] Iha, Takahiro, Ibuki Kawamitsu, Ayako Ohshiro, and Morikazu Nakamura. "An LSTM-based Multivariate Time Series Predicting Method for Number of Restaurant Customers in Tourism Resorts". *International Technical Conference on Circuits/Systems, Computers and Communications. 27-30 Jun. Jeju Korea (South)* : pp. 1-4, 2021. doi: 10.1109/ITC-CSCC52171.2021.9501432.

[21] Alhirmizy, Shaheen, and Banaz Qader. "Multivariate time series forecasting with LSTM for Madrid, Spain pollution". *International Conference on Computing and Information Science and Technology and Their Applications. 3-5 Mar. Kirkuk Iraq* : pp. 1-5, 2019. doi: 10.1109/ICCISTA.2019.8830667.