Bangladesh Crime Reports Analysis and Prediction

Md Pavel Rahman*, A.K.M Ifranul Hoque*, Md. Faysal Ahmed, Iftekhirul, Ashraful Alam, Nahid Hossain*†
Dept. of Computer Science and Engineering
United International University
Dhaka, Bangladesh
mrahman172017@bscse.uiu.ac.bd, ahoque172189@bscse.uiu.ac.bd, mahmed172030@bscse.uiu.ac.bd,
iftekhirul172006@bscse.uiu.ac.bd, aalam172005@bscse.uiu.ac.bd, nahid@cse.uiu.ac.bd

***Abstract*—The systematic method of detecting crime, evaluating crime patterns, and anticipating crime trends is known as crime analysis and prediction. Crime is inherently unexpected and causes societal disruption. As Bangladesh's population grows, so does the prevalence of crime, which is wreaking havoc on our society in various ways. As a result, analyzing crime data has become crucial for a better understanding of future crime patterns. Machine learning and data mining techniques can be quite useful in predicting future crime trends and patterns in this situation. Various machine learning algorithms are utilized in this study to predict future crime patterns in Bangladesh. The crime statistics are gathered from the Bangladesh Police website to analyse and predict dacoity, robbery, murder, women and child repression, kidnapping, burglary, theft, and other crimes in Bangladesh's various regions. Another dataset from ACLED has been used to predict different kinds of events such as battles, explosions, protests, riots, strategic developments, violence against civilians with geolocations of the events. This research might assist Bangladesh police and law enforcement authorities to predict, prevent, and solve future crimes. The performance and success rate of the project are highly satisfactory. All resources of the project can be found at https://tinyurl.com/297yykmu**

***Index Terms*—Bangladesh, crime prediction, analysis, bangladesh police, ACLED, crime report**

I. Introduction

Bangladesh is the largest delta in the world [1], and it is a small, densely populated country in South Asia [2]. It gained independence in 1971 [3]. Having a relatively small volume, vast population, less GDP, low per capita income, and because of being a third world country, one of the major problems that Bangladesh is currently facing is crime. Having a vast and dense population, Bangladesh is a disaster-prone area with a lack of education and uncertainties of basic needs, which play a key role behind the increasing rate of crime in Bangladesh day by day. Compared to the crime, the population, workforce of Bangladesh police, and other administrative sections of the Bangladesh government, the workforce is limited. Terrorism, child crime, hijack, dacoity, and protest, are the daily events of many villages and cities in Bangladesh.

In a third-world country like Bangladesh, criminal data must be analysed carefully because crimes are not reported most of the time. This makes it harder for us to find the authentic data that is needed for our work. Initially, our work has started with Bangladesh police dataset [4]. There have been some inconsistencies in the Bangladesh police dataset, but still, our work has been done successfully. After doing some work on this dataset, we have moved on to our main working project, which is the ACLED dataset [5].

As the crime rate in Bangladesh is increasing day by day, it is really important to analyze its reason. Research on analyzing and predicting crime is done in many countries around the world but not in Bangladesh. Due to lack of efficient work, we feel the necessity of doing this research. In this project machine learning approach has been used to analyze previous crime data, and based on that, future crime occurrences have been predicted. It will help law enforcement agencies to interact with different crime areas and improve surveillance. Other than that, it will also create a massive opportunity for future research on this topic because it will help to predict crime based on Geolocations. In this paper, two significant datasets have been used. One dataset is from the Bangladesh Police website, and another dataset is from the ACLED website.

For the Bangladesh Police dataset, there have been various police units in Bangladesh like DMP, CMP, KMP, Khulna Range, and Sylhet Range. Then the crime event data held in those areas for events such as Dacoity, Murder, Kidnapping, and Police Assault have been taken. There are 15 crime events like these in the dataset. Data has been collected from 2010 to 2018, and there is some data from 2019, but those are very inaccurate, so it has been left out. In order to understand the various crimes and their consequences, the data has been analyzed by us, and a prediction has been made based on the number of occurrences that might occur. It is done simply by splitting the dataset into a train-test split using the Sklearn [6] library. Then after using various machine learning models like Random Forest, MLP, and Decision Tree to find out the predictions, the best model has been used to predict 2019 since that is the closest unknown data that our model can predict. Also, the entire dataset has been used to predict 2019.

For the ACLED dataset, there are various events like Battles, Riots, Violence Against Civilians, Protests, and Strategic Developments in Bangladesh from 2001 to 2020. The dataset also consists of the years and locations along with the location's latitude and longitude in separate columns. Our goal is to use this dataset to conduct various predictions like predicting the number of events that occurred by districts in Bangladesh. It has also been predicted what kind of event might occur in

*equal contribution
†corresponding author

a given location with specific longitudes and latitudes. After finding the best algorithm, the number of events in a given district for 2021 has been predicted. We also predict what kind of crime event might occur in 2021.

The paper is organized as follows: In section II, related work has been described that is done in this field with some background information. The proposed method has been explained in section III. Section IV demonstrates experimental results and analysis. Whereas in section V, a conclusion has been done for the paper mentioning our system's future works and limitations.

II. Related Works

In 2016, M.M.A Hashem et al. proposed an approach to forecast future trends in crime in Bangladesh [7]. For different criminal activities, they used linear regression to train the model using data mining techniques and test with crime data for the year 2016. During the year 2018, Tasmia et al. proposed event and violence recognition from textual news [8]. They extracted data using the NLP tool and triggering word, and a BUH classifier was used for preprocessing the data. They used both MLP and supervised learning algorithms, where supervised learning algorithms provided more accurate results than MLP. In the same year, Mittal et al. projected the impact of the economic crisis [9]. They used four machine learning algorithms, and out of them, linear regression gave the highest accuracy. Nishat suggested a machine learning approach to predict crime using time and location data in 2017 [10]. This model overcame imbalance difficulty by oversampling and undersampling the dataset. AdaBoost decision tree gave 81.93% accuracy among all other machine learning algorithms. Also, in the same year, Waduge presented the term "Modus Operandi' to detect crime patterns in his paper [11]. He discussed several techniques such as K-means, ANN, and Hierarchical clustering. Prabakarar et al. presented an analysis of crime detection techniques [12] where they discussed several data mining and machine learning algorithms for the detection of various crimes.

III. Proposed Method

Our proposed system's outline and development methods have been presented in this section with relevant examples, figures, and tables.

A. Dataset Description

Data is the most unorganized thing in the world. Moreover, when it comes to working with real-world data, furnishing them is a must thing to do. Fortunately, there are many ways to do that. This section reflects dataset related tasks of our project.

After collecting data from Bangladesh Police and ACLED website, two significant datasets have been formed. At first, work related to the Bangladesh police dataset has been done. From there, crime data of the years 2010 to 2018 have been extracted. This dataset has around 15 unique crime events as attributes. The ACLED dataset is considered a location-based dataset, as it provides latitude and longitude for each event. This feature is really helpful for visualization and pinpointing crime events on a map of Bangladesh.

B. Preprocessing

For preparing the data for our model, both datasets have gone through preprocessing. After checking for any missing data or empty cells, unnecessary data have been left out. Two sub-datasets have been created from the ACLED dataset to work with. One sub-dataset has 64 districts that have been labeled for every year with the latitude and longitude to locate the place of a town. For another sub-dataset, event counts have been calculated based on locations, and the count values have been appended in a separate column in a newly created dataset. Techniques such as Standard Scaling and Label Encoding for the datasets have been used. Programming techniques have been used to sort data for ACLED as needed so that better predictions are made. The district names of 1380 locations have been collected using the geolocator API function of Folium [13]. Furthermore, these districts have been visualised using the map function of Tableau [14].

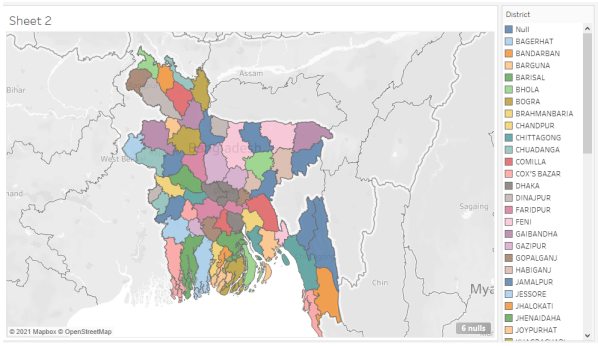

Fig. 1. DISTRICT WISE COLORED MAP OF BANGLADESH

Figure 1 shows a colored map separated by districts for Bangladesh using Tableau.

C. Model Generation

Google Colab [15] has been chosen as our environment as it is user-friendly and easy to adapt. On the other hand, Sklearn has been used as a machine learning library for the different algorithms that it provides. For mathematical problems and data structures, Pandas [16], and Numpy [17] has been used. Various machine learning algorithms such as Decision Tree, Random Forest, Support Vector Regression, and Multilayer Perceptron have been used for our model. For example, Decision Trees are a non-parametric supervised learning method that is useful for tasks such as classification and regression. Random forests are an ensemble learning method where a large number of decision trees are trained for prediction. MLP is a feedforward artificial neural network model that translates data input to a collection of suitable outputs. After splitting the dataset into train and test sets, the training dataset needs to be trained into our model. Here, we have used a basic 75-25 ratio for the train-test splitting of data.

For datasets with a small amount of data like the Bangladesh Police dataset, we have also employed k-fold cross-validation. The procedure works by dividing into k non-overlapping folds using the k-fold cross-validation process. Each of the k-folds is given a chance to be used as a held-back test set, while the rest of the folds are combined to form a training dataset. The best algorithm has been selected depending on accuracy, precision, recall, f1 score for classification tasks. And the R2 score and mean-square error for regression tasks.

For the Bangladesh police dataset, 3 major model predictions have been made. In the first prediction, the year 2018 has been predicted with the rest of the dataset as training data. In the second one, random years have been predicted using train-test splitting of the data. This includes 2018 during the training of the model. In the third prediction, k-fold with 5 folds has been done using the entire dataset.

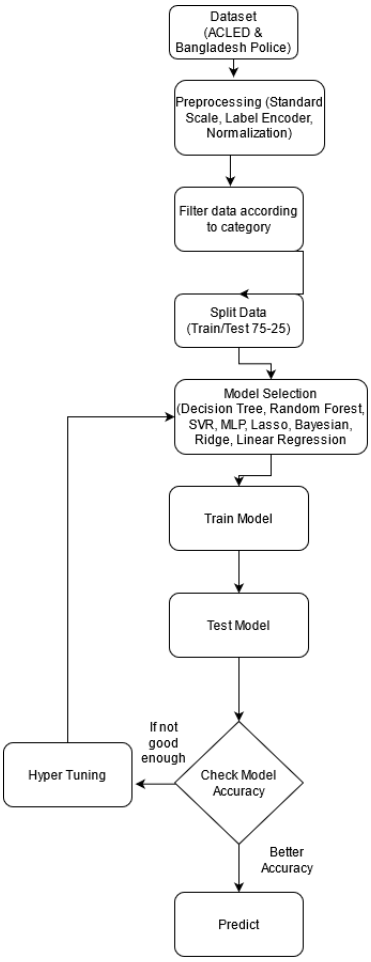

Fig. 2. A SYSTEMATIC APPROACH FOR OUR PROPOSED STUDY

For the ACLED dataset, several predictions have been made. First of all, event types that might occur have been predicted using year, location, longitude, latitude data as attributes. Then we have predicted longitude, latitude, and event types using year and location. Several approaches can be used here. The first method uses a regression model first to predict the coordinates and then uses a classification model to predict the event types separately and finally combine the outputs of the results. The second method is outputting the coordinates first in the first model (regression), then using that output as input in the second model (classification), and then predicting the event type. The third method is using neural networks. The first method has been applied in our prediction. Multi-Output Regression has been used to predict longitude and latitude at once, which is natively supported by Linear Regression, KNN, and Decision Tree models in the Sklearn library. We have predicted total counts of events for specific districts by appending the count in a modified dataset. The metrics for each of the algorithms have been found using the Sklearn library, and the best model has been chosen for our dataset.

In Figure 2, a systematic approach has been shown that is used to build, train and test our machine learning models. It is seen that we have first taken the input of the data used for our model. Then preprocessing has been done on this data, such as using a label encoder, standard scaling, or normalization. Then the data has been filtered according to category. If a specific crime category is needed to be picked, then the work is done here. After that, the data has been split into train and test data and used for prediction. Later, various metrics such as accuracy, precision, and recall are checked to see the best model. If the outcome is not satisfactory, then another model is chosen, or the current model settings are tuned to improve prediction performance. After satisfied metrics have been achieved, then the best model is used to predict new unseen data.

D. Evaluation

The evaluation metrics that have been used in our predictions have been shown here. They are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 * TP}{2 * TP + FP + FN} \tag{4}$$

$$MSE = \sum_{i=1}^{D} (x_i - y_i)^2 \tag{5}$$

$$MAE = \sum_{i=1}^{D} |x_i - y_i| \tag{6}$$

$$R^2 = 1 - \frac{RSS}{TSS} \tag{7}$$

IV. Experimental Results and Analysis

We have shown all the experimental analysis and results that have been achieved during our study in this part. This section of the paper goes through the Bangladesh Police dataset and the ACLED Dataset.

A. Bangladesh Police Dataset

Data related to Bangladesh police has been discussed in this subsection. The crime data of the last 10 years from the Bangladesh Police has been collected and used, but there is some inconsistent data in 2019 and incomplete data for 2020. So our dataset consists of data from 2010 to 2018. We have taken various crime data from previous years and have trained the model using algorithms like Random Forest, Decision Tree, SVR, MLP(ADAM), Lasso, Bayesian, Ridge, and Linear Regression. Error metrics have been used, such as Mean Square Error, R2 Score, and Mean Absolute Error on the Bangladesh Police dataset for the whole country. It has been predicted that there will be 262 dacoities, 562 robberies, and 3830 murders for 2019. It is seen that MLP with Adam Optimizer works better than the other algorithms.

Table I shows a sample of metrics calculated for Riot crime. The data of years 2010 to 2017 have been taken as training data and 2018 as the test data for Riot crime values. Then we have compared the predicted values with true values and have calculated the metrics shown. It is seen that the MLP model gives a better prediction compared to other algorithms.

B. ACLED Dataset

In this subsection, we have added several reports that have been done for the ACLED dataset. It is seen that there are over 30629 events that have occurred from 2001 to 2020. Out of these events, there are 2898 battles, 356 explosions remote violence, 8716 protests, 13537 riots, 391 strategic development, and 4731 violence against civilians.

1) ACLED Report 1: We have analyzed entire Bangladesh as a whole and have made a table of total counts of events for the year divided into event types. The highest total events have occurred in Dhaka, which is around 7079. Furthermore, the lowest total events have occurred in Barguna, which is around 79. It has been analyzed by location, the different towns (1380 towns) and the total events per year, and specific events per year. The highest battles occurred in Dhaka, which is around 231, and the lowest is in Panchagargh, with 4 battles only.

Figure 3 shows patterns for some cities and their relation to total crime events. It is seen that Rajshahi has more total crime events than most other cities.

Geolocation Analysis: We have chosen Folium using Nominatim API to show the geolocations of crime activities using the latitude and longitude data in the ACLED dataset.

Figure 4 shows the geo-mapped events in Bangladesh. The blue dots are events that occurred.

3 Prediction outputs have been made. Prediction 1: Here, we have predicted the number of each specific event for 2020 using train test splitting.

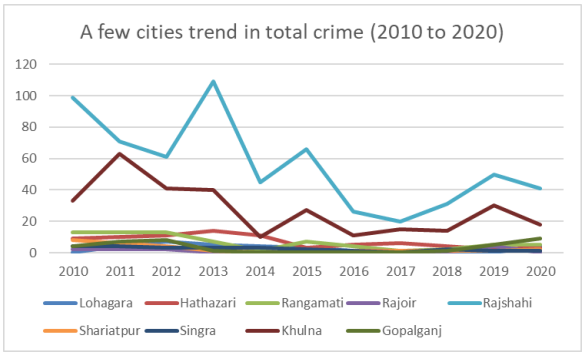

Fig. 3. NUMBER OF BATTLES PER YEAR TREND BY CITIES.

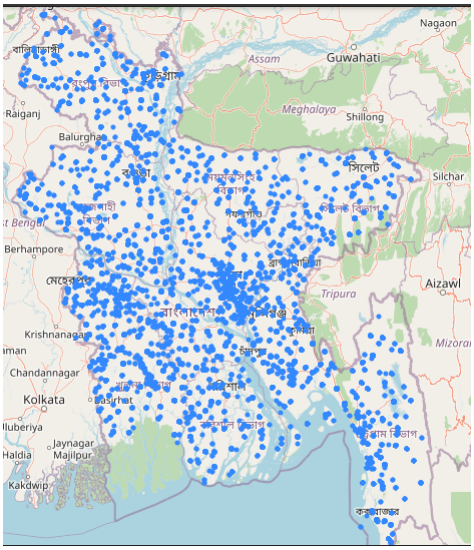

Fig. 4. GEO-MAPPED EVENTS IN BANGLADESH.

Table II shows predicted values and metrics for ExplosionsRemoteViolence (shown in page below). Here it is seen that Decision Tree and MLP have correctly predicted 7 explosion events for the year 2020.

Prediction 2: A prediction has been made for 2021 using the entire dataset.

Year	Event	Total
2021	Battles	21
2021	ExplosionsRemoteViolence	7
2021	Protests	971
2021	Riots	320
2021	StrategicDevelopments	17
2021	ViolenceAgainstCivilians	184

TABLE III
CRIMES PREDICTED FOR 2021 USING ENTIRE ACLED DATASET

Table III shows predictions for 2021.

Prediction 3: The latitude, longitude, and event type have been predicted using Year and Location as inputs.

- Test Data: 2021, Dhaka Predicted Data: 23.88406365, 90.6020547, Protests
- Test Data: 2021, Barisal Predicted Data: 23.62803754, 90.67983387, Riots

Algorithms	Test Years	Test Values	Predicted Values	Mean Square Error(MSE)	Mean Absolute Error(MAE)
Decision Tree	[2018]	[26]	[23.0]	3	3
Random Forest	[2018]	[26]	[36.24]	10.24333333	10.24333333
SVR	[2018]	[26]	[92.35]	66.34686947	66.34686947
MLP(Adam)	[2018]	[26]	[23]	3	3
Lasso	[2018]	[26]	[33.94]	7.942857143	7.942857143
Bayesian	[2018]	[26]	[41.88]	15.87937193	15.87937193
Ridge	[2018]	[26]	[35.26]	9.25872093	9.25872093
Linear Regression	[2018]	[26]	[33.86]	7.857142857	7.857142857

TABLE I
PREDICTION FOR RIOT CRIME FOR WHOLE COUNTRY USING BANGLADESH POLICE DATASET

Algorithms	Test Years	Test Values	Predicted Values	Mean Square Error(MSE)	Mean Absolute Error(MAE)
Decision Tree	2020	7	7.0	0	0
Random Forest	2020	7	12.51	5.508809524	5.508809524
SVR	2020	7	9.34	2.339968012	2.339968012
MLP(Adam)	2020	7	7	0	0
Lasso	2020	7	15.82	8.822807018	8.822807018
Bayesian	2020	7	18.36	11.36188259	11.36188259
Ridge	2020	7	15.79	8.79399023	8.79399023
Linear Regression	2020	7	15.79	8.789473684	8.789473684

TABLE II
PREDICTION FOR EXPLOSIONSREMOTEVIOLENCE EVENT FOR 2020 USING ACLED DATASET

Here it is seen that 2021 and Dhaka have been taken as inputs in the model. The outputs show the longitudes, latitudes, and protests as the event type.

2) ACLED Report 2: The following figure has mapped Battle Density by District for years 2001 to 2020:

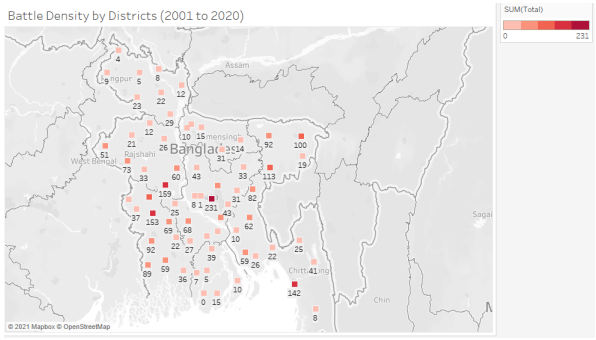

Fig. 5. BATTLES DENSITY BY DISTRICT FOR BANGLADESH

Figure 5 shows battle density by districts. It is seen that Dhaka has the highest number of battles which is around 231.

Further improvements to the dual prediction (Prediction 3 in ACLED Report 1) have been made by adding probability for the classification.

Model	R2 Score	MSE	Precision	Recall	F1 Score
Decision Tree	0.929	0.256	0.486	0.513	0.493
MLP	0.060	0.937	0.513	0.71	0.588
Random Forest	0.937	0.241	0.489	0.518	0.497
SVM	0.051	0.946	0.502	0.686	0.557
KNN	0.543	0.654	0.448	0.478	0.454

TABLE IV
DUAL PREDICTION (GEOLOCATION AND EVENT TYPES) METRICS FOR ACLED DATASET

Table IV shows performance metrics of dual predictions. It is seen that Random Forest works best. An example of input-output is shown here. Test Data: 2021, Dhaka. Predicted Data: 23.7104, 90.4074, [0. 0.00568298 0.79119182 0.13058522 0.02704326 0.04549672]. Here the input is year and location, and prediction is latitude, longitude, and event type. The event types are based on the probability of each event happening. The probability classes are battles, explosions, remote violence, protests, riots, strategic developments, violence against civilians. So for 2021 and Dhaka as input, we get the longitude, latitude, and event type probability as output. Here it is seen that Protests have the highest likely probability of 0.79119182.

3) ACLED Report 3: K Means Clustering has been used to cluster the locations by the district.

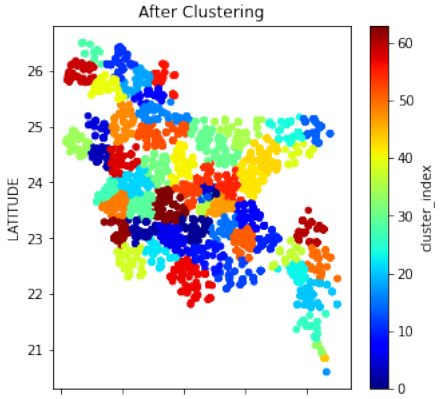

Fig. 6. DISTRICT WISE CLUSTERING USING K-MEANS CLUSTERING

In Figure 6 we have used K-Means Clustering to cluster the districts. There are around 64 clusters shown here.

Specific event totals have been predicted for 2021, such as there will be 21.0 Battles, 7.0 ExplosionsRemoteViolence, 971.0 Protests, 320.0 Riots, 17.0 Strategic Developments, and 184.0 Violence against civilians. The highest battle totals for 2021 will be in Rangamati Hill at 6.55 battles, and the lowest will be in Cox's Bazaar at 1.22 battles.