Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh

Md. Farhan Labib, Ahmed Sady Rifat, Md. Mosabbir Hossain, Amit Kumar Das, Faria Nawrine
Department of Computer Science and Engineering
East West University
Dhaka, Bangladesh
Email: farhan.labib4@gmail.com, sadyrifat@gmail.com, mosabbirtarek7@gmail.com, amit.csedu@gmail.com, nawrinefaria@gmail.com

Abstract*— In recent years, the road accident has become a global problem and marked as the ninth prominent cause of death in the world. Due to the enormous number of road accidents every year, it has become a major problem in *Bangladesh*. It is entirely inadmissible and saddening to allow its citizen to kill by road accidents. Consequently, to handle this overwhelmed situation, a precise analysis is required. This research paper has been done to analyze traffic accidents more deeply to determine the intensity of accidents by using machine learning approaches in *Bangladesh*. We also figure out those significant factors that have a clear effect on road accidents and provide some beneficent suggestions regarding this issue. Analysis has been done, by using *Decision Tree, K-Nearest Neighbors (KNN), Naïve Bayes* and *AdaBoost* these four supervised learning techniques, to classify the severity of accidents into Fatal, Grievous, Simple Injury and Motor Collision these four categories. Finally, the best performance is achieved by *AdaBoost.

Keywords- Accident Severity, Machine Learning, Supervised Learning Feature Analysis, Road Accident.

I. INTRODUCTION

Road Accident is the most undesirable and unexpected thing to occur to a road user, though they happen quite often. Unfortunately, we can see a minatory rise of road accidents in Bangladesh, conspicuously highroad accidents over the past few years. It has a massive impact on society as well as in the economy of our country as there is an immense cost of fatalities and injuries. According to a recent inspection, conducted by the Accident Research Institute (ARI) of Bangladesh University of Engineering and Technology (BUET), annually on an average 12,000 lives have been taken by road accidents and lead to almost 35,000 injuries [1]. This record indicates that every day, approximately 32 people were killed in Bangladesh by road accidents and it is quite devastating. Besides this, according to WHO, the economic cost of road accidents to a developing country like us is 2-3% of GDP, which is a significant loss for a country like ours. Moreover, reducing this loss has become a great matter of concern for our country now.

In recent years, traffic accident analysis drew considerable attention to the researchers to determine the factors that significantly affect traffic accidents. But unfortunately, maximum research methods are based on statistical records or by doing some simple survey based on interviews or questionnaires. But, it is not possible to get a better and unerring solution by using these types of primitive approaches. The main puzzle is that behavioral features in traffic accidents are quite difficult to study by these kinds of traditional research methods. Because, accidents are relatively unpredictable and extemporaneous, so direct observation is quite difficult. For that reason, getting 100% accurate data is quite impossible. Implementation of an advanced method that can give better analysis result is a crying need here. Machine learning is one of the most advanced scientific fields of AI that can be applied here to get a better result.

The prime goal of this research paper is to analyze the road accidents and determines the severity of an accident by applying advanced machine learning techniques [2]. There exist so many developed methods in machine learning to examine this sector. In this research paper, the authors perform traffic accident analysis, by applying four advanced and most popular supervised learning techniques of machine learning because of their proven accuracy in this sector. Those approaches are- Decision Tree, K-Nearest Neighbors (KNN), Naïve Bayes and Adaptive Boosting (AdaBoost).

In this paper, the authors classified the intensity of an accident into four categories- Fatal, Grievous, Simple Injury and Motor Collision. To categorize the severity of any traffic accident in these four categories, eleven main factors that affect the maximum number of accidents in Bangladesh have been selected as the feature. Previously occurred, almost 43 thousand traffic accidents data in Bangladesh from 2001 to 2015 have been used as our learning materials. Among these four techniques best performance is achieved by AdaBoost and its accuracy was 80%.

The remnant of this paper is ordered as follows-Section II presents a concise description of some previous works related to this topic. Section III gives a simple overview of our working procedure and provides information regarding dataset preparation. In Section IV, we briefly describe the methodology of these four proposed methods. In Section V, we present our result section by comparing these four methods based on their accuracy. In Section VI, we discuss our future directions and conclude the paper by providing some beneficial suggestions.

978-1-7281-1557-3/19/$31.00 ©2019 IEEE

978-1-7281-1557-3/19/$31.00 ©2019 IEEE

II. RELATED WORK

In the context of Bangladesh traffic, the authors did not find any advanced and manifested research work in this field. There exist a few research works regarding accident analysis for Bangladesh traffic, but unfortunately, these studies were done by applying primitive statistic methods or by doing a simple survey. Lamentably, implementation of advanced machine learning concept is still in the formative stage here.

In paper [3], the authors proposed a method to detect the possibility of accidents on the road by using vision-based techniques in the context of Bangladesh. They have used roadside video data as their learning materials and achieved 85% accuracy in particular situations. The researchers in the paper [4], analyzed 892 traffic accidents in N5 National Highway in Bangladesh by using several decision tree induction algorithms and have tried to figure out the traffic accident patterns. They also draw out the rules for the trees to decrease road accident in this highway. On the contrary, there exist a lot of advanced and beneficial research works regarding this field in other developed countries. In paper [5], the authors have analyzed the status of the road accident occurrence by using machine learning techniques in Istanbul. They have applied the CART algorithm to determine the risk of the accident and achieved above 81.5% of accuracy. Two authors in the paper [6], have applied K-means clustering algorithm and association rule mining these two techniques of data mining to figure out the major factors affiliated with traffic accidents. The authors in the paper [7], also used K-means clustering algorithm and association rule mining these two data mining techniques to identify the most frequent accident-prone area and the main factors linked to those accidents in India. In paper [8], the researchers applied decision tree, Naïve Bayes and KNN these three data mining techniques to find connections the recorded road characteristics to accident intensity in Ethiopia. Based on the founding's they also developed a set of rules to improve road safety in Ethiopia. The authors in the paper [9], used logistic regression based on the vehicle condition after an accident to figure out the impact of road defects on accident severity.

III. PROPOSED MODEL

There are three types of machine learning algorithms-supervised and semi-supervised learning, unsupervised learning, and reinforcement learning [10]. Among these three broad categories of machine learning classification approaches the supervised learning approach has been used in this research paper because of its competency in modeling and regulating dynamic systems. Here, the authors have used the four most popular machine learning techniques for road accident analysis [11]. Those are *Decision Tree, KNN, Naïve Bayes* and *AdaBoost*. Figure 1 presents a simple view of the overall working process.

A. Preparation of dataset

Accurate and extensive accident data records are the most important and prime need to get better performance by applying machine learning approaches. But, getting a perfect and 100% accurate dataset is quite challenging. Therefore, to process data based on the need the authors have followed the following instructions.

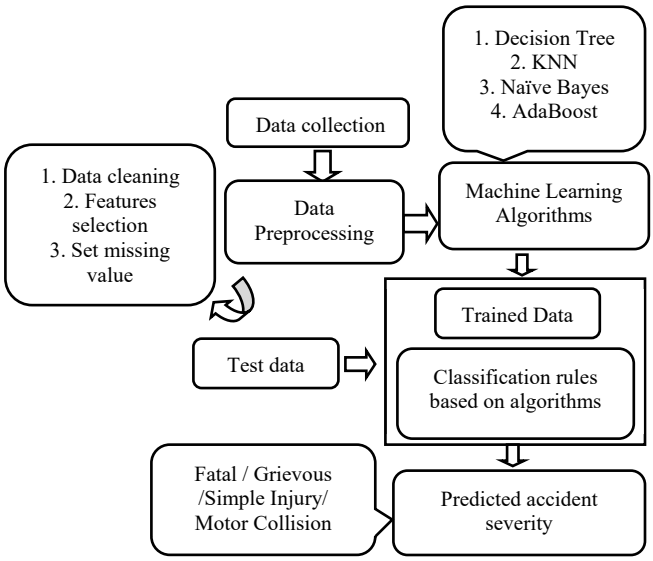

Figure 1. The working mechanism of proposed approaches

1) Data Collection:

For the accurate prediction of the severity of accidents, a considerable number of traffic accident records with full information is required to train by using the proposed approaches. In this research work, the authors have collected a dataset from the ARI of BUET that consists of total 43,089 traffic accidents record from the year 2001-2015 in Bangladesh. We split our entire dataset into two parts-Training Dataset and Test Dataset. 70% of the whole dataset has been chosen randomly by using a python library as a training data set and the remaining 30% has been used as our test dataset. We have used the 70-30 ratio for splitting dataset because of its proven accuracy.

2) Data Pre-Processing:

In this dataset, all the accident records were written with formal words. We properly organize this total dataset based on the feature. In total, we have found 34 factors that affect previous accidents in some way. Firstly, we methodize all accident records by using these 34 features. After that, for many accident records, we have found 8.7% missing values in the total dataset and the selective 11 features (Table 1) has 1.65% missing value. As these missing values can affect the performance, on account of this, we have applied a method by using the mean value of that feature column to provide an amount where it is required. We use this method as there presents no extreme value which can affect the mean.

a) Feature Selection: Working with a large number of features may affect the performance because training time increases exponentially with the number of features. Even, it has also the risk of overfitting with the increasing number of features. So, for getting a more accurate prediction, feature selection is a critical factor here. Sklearn (A python machine learning library) has been used to diverge the feature with less importance. To obtain the most essential features, we operate

an experiment by applying three different algorithms of feature selection[12]. Those algorithms are *Univariate Feature Selection, Recursive Feature Elimination,* and *Feature Importance.*

1) Univariate Feature Selection explores each feature severally and selects the best features based on univariate statistical tests. There are multiple ways of implementing the univariate feature selection. For this research work, to select the topmost essential features we have applied the Chi-Squared statistical test for non-negative features.
2) Recursive Feature Elimination works by removing the features recursively and uses the accuracy of the model to pick the features that contribute the most to predict the desired variable. Initially, it trained all the features, and by exerting logistic regression, it tried to figure out the significance of each feature. After that, it pruned the features with less importance, and this process gets repeated until it attains the desired number of features.
3) Feature Importance is a trained, supervised classifier to find out the critical feature. It works like a classifier and evaluates each attribute to create splits. By using, information gain approaches it finds which features are affected less and then ranked them according to the measure.

Table I: Top 15 important features gained by using these 3 Algorithms.

Univariate Feature Selection	**Recursive Feature Elimination**	**Feature Importance**
Junction Type	Time	Junction Type
Thana	Traffic control	Thana
District	Weather	District
Time	Light	Time
Traffic control	Road geometry	Traffic control
Weather	Vehicle type	Weather
Light	Movement	Light
Road Geometry	divider	Road Geometry
Vehicle type	Road class	Vehicle type
Movement	Surface condition	Movement
Divider	Surface Type	Divider
Road class	Surface Quality	Road class
Location type	Location Type	Location type
Vehicle defect	Road Feature	Vehicle defect
Surface condition	Vehicle defect	Vehicle loading

We applied these three feature selection methods and obtained top 15 features for each technique (Table I). After that, we try to figure out the standard features among these three experiments and get 11 common features within them (Table I).

IV. The Methodology of Proposed Approaches

a) Decision Tree: For classification problems, the decision tree is extensively used the supervised algorithm. The primary perspective of this algorithm is predicting the value of the desired variable by learning decision rules deduced from the features of the data and create a model of that.

First of all, a root node is designated for the construction of this model based on the best attribute picked by the gain approach and the sub-nodes are then generated on the basis of the decision taken in relation to the status of quality selected at each node. When each node is reduced to a single quality status, the class is determined at the end of the node; it is called a leaf. These courses of action continue recursively until a class is defined at the end of each node. [13]

b) AdaBoost: AdaBoost is mainly a boosting algorithm which is used with short decision trees. Every instance is weighted in the training dataset. At first, the weight is set to,

$$Weight, w_i = \frac{1}{n}$$

Where w_i is the i'th training instance weight and n is the number of training instances. Further, the first tree is created, the performance of the tree on each training instance is used. After that, it evaluates overall errors. Next iteration weights are calculated by the errors. More weight is given where hard to predict, whereas less weight is given where easy to predict. [14]

c) KNN: KNN is a classification algorithm which is based on feature similarity. It analyzes the data and measure the distance and similarities between data and cluster them based on K values. Distance is calculated in many ways, for this research, we used Euclidean distance measurement. The class of new input data is classified by calculating the distance between the clusters and assigned it to the closest one. [15] Euclidean distance is calculated by,

$$d(x, x') = \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2 + \ldots + (x_n - x_n')^2}$$

d) Naïve Byes: Naïve Byes is another classification technique based on Bayes theorem. It predicts the probability of different class based on several attributes and assigns the new class to the highest probability. The algorithm working rule:

$$Posterior\ probability,\ \ P(a|b) = \frac{P(b|a) * P(a)}{P(b)}$$

The posterior probability is mainly the probability of "a" being true given that "b" is true. [16]

V. Result Analysis and Discussion

In this research paper, to evaluate the performance of the proposed approaches, we performed two different

experiments based on the accident severity class. In our first experiment, we have determined the performance of each algorithm, for four accident severity classes (Fatal / Grievous /Simple Injury/ Motor Collision). Naïve Bayes and Ada-Boost both of them, achieve the high accuracy among these four approaches, and their accuracy is 80% (Table II). By overall performance, Ada-Boost gives the best result because of its iterative classification on decision tree.

No. of Class / Algorithms	Precision (%)		Accuracy (%)		F1 Score (%)	
	Four Class	Two Class	Four Class	Two Class	Four Class	Two Class
Decision Tree	68	70	71	73	71	71
KNN	68	70	67	69	67	69
Naïve Bayes	63	63	80	80	71	71
Ada-Boost	68	75	80	80	73	74

Table II: Severity prediction results of algorithms

We observe that most of the accidents in our dataset are Fatal and value for the other three classes is very low. For that reason in our second experiment, we merge Grievous, Simple Injury, Motor Collision these three accident severity classes into one class. Therefore, we have attained the performances of the proposed approaches for two accident severity classes (Fatal / Grievous). In this experiment, we have noticed that the accuracy of Decision Tree and KNN get increased, though the accuracy of Naïve Bayes and AdaBoost remain the same (Table II). But it is also mentionable that, the performance of AdaBoost is much better than the previous experiment as precision and F1 score increased here in a noticeable way.

Besides this, we did experiment with the features in our dataset and have tried to find out their effect on a traffic accident. In Figure 2, statistically we have found that based on the condition of some features the number of accidents gets increased. It's a significant noticeable thing for making proper steps to decrease the number of accidents.

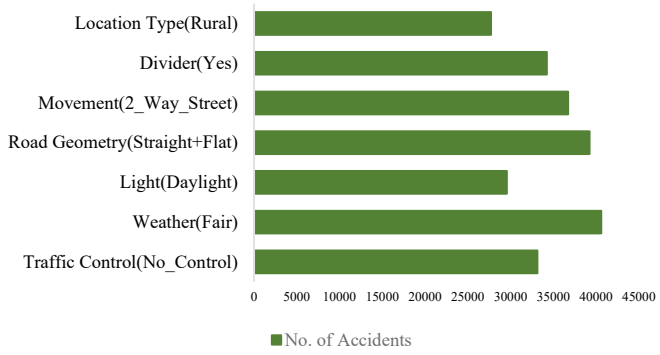

Figure 2. Number of accidents for the condition of features

After that, we analyzed the effect of vehicle types on a traffic accident, and it is seen that the bus is responsible for 39% traffic accident and trucks for 32% (Figure 3). In Figure 4, it is found that most of the accidents have happened on the National road.

Thereafter, we try to figure out, at which time traffic accidents occur more often. Based on the result, it is excogitated that in the rush hour (06-18) accident rate is very high compared to other time (Figure 5).

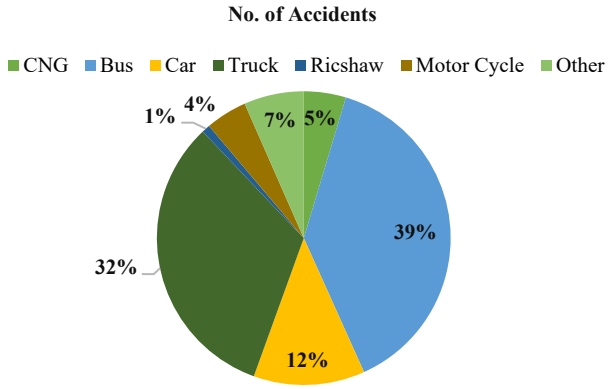

Figure 3. Effect of vehicle type on accident

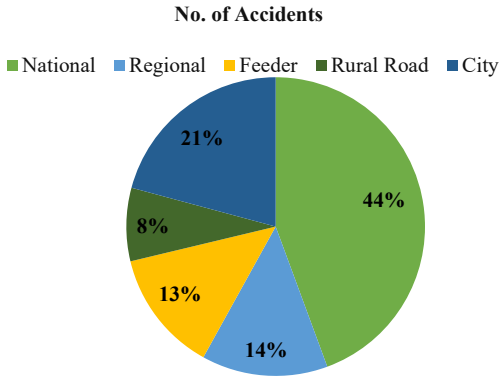

Figure 4. Effect of road class on accident

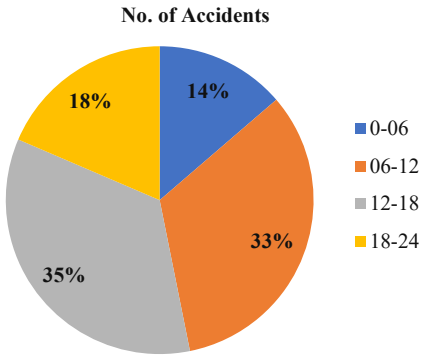

Figure 5. Effect of time on accident

Then we analyzed the accidents based on junction type. From Figure 6, we observed that the accident rate is higher where no junction exists and at the T-Junction. Finally, in Figure 7, we have found that the number of accidents gets increased based on the condition of surface effect features. It has been found that when the surface condition is dry, surface type of sealed and, surface quality is good most of the accident occurred at that time.

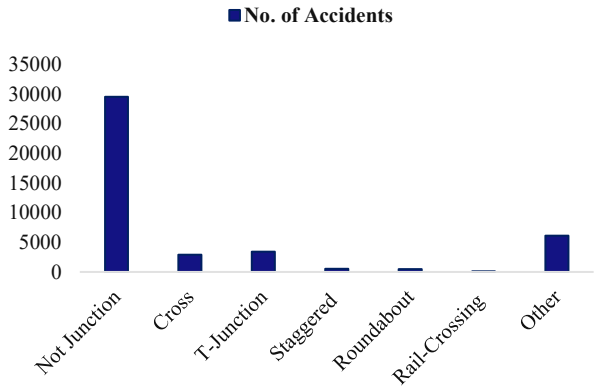

Figure 6. Effect of junction type on accident

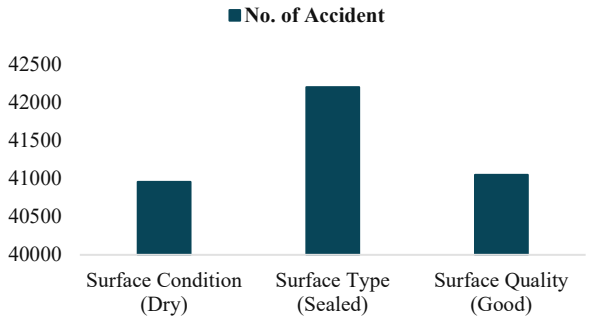

Figure 7. Effect of surface condition on accident

VI. CONCLUSION

Losses in road accidents are unbearable, to the society as well as a developing country like us. So, it has become an essential requirement to control and arrange traffic with an advanced system to decrease the number of road accidents in our country. By taking simple precautions, based on prediction or warnings of a sophisticated system may prevent traffic accidents. Moreover, it's a primary need for our country now, to tackle this situation where every day so many people were killed in a traffic accident and day by day this rate is getting increased. The implementation of machine learning is a functional and a great approach to take an accurate decision with the experience to manage the current situation and the findings of the analysis part (Figure 2- Figure 7) can be suggested to traffic authorities for reducing the number of accidents. We can use proposed approaches to implement machine learning here because of their proven and higher accuracy to predict traffic accident severity.

Moreover, to make it more feasible, we will try to make a recommender system by using these approaches that can give a prediction to the traffic accident and can warn the road user. In the future, it will be our try to create a mobile application by implementing this methodology to provide an accurate prediction to the user and make it very useful and beneficial also.

.

REFERENCES

[1] T. Rahman, "Road Accidents in Bangladesh: An Alarming Issue", The World Bank, 2012. [Online]

[2] K. M. Habibullah, A. Alam, S. Saha, A. Amin and A. K. Das, " A Driver-Centric Carpooling: Optimal Route-Finding Model using Heuristic Multi–Objective Search," 2019 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 2019.

[3] M. M. L. Elahi, R. Yasir, M. A. Syrus, M. S. Q. Z. Nine, I. Hossain and N. Ahmed, "Computer vision based road traffic accident and anomaly detection in the context of Bangladesh," 2014 International Conference on Informatics, Electronics & Vision (ICIEV), Dhaka, 2014, pp. 1-6.

[4] M. S. Satu, S. Ahamed, F. Hossain, T. Akter and D. M. Farid, "Mining traffic accident data of N5 national highway in Bangladesh employing decision trees," 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Dhaka, 2017, pp. 722-725.

[5] H. İ. Bülbül, T. Kaya and Y. Tulgar, "Analysis for Status of the Road Accident Occurrence and Determination of the Risk of Accident by Machine Learning in Istanbul," 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, 2016, pp. 426-430.

[6] P. A. Nandurge and N. V. Dharwadkar, "Analyzing road accident data using machine learning paradigms," 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics, and Cloud) (I-SMAC), Palladam, 2017, pp. 604-610.

[7] S. Kumar and D. Toshniwal, "A data mining approach to characterize road accident locations," Journal of Modern Transportation(2016), vol. 24, issue no. 1, pp. 62-72.

[8] Beshah, Tibebe, and Shawndra Hill, "Mining Road Traffic Accident Data to Improve Safety: Role of Road-Related Factors on Accident Severity in Ethiopia," AAAI Spring Symposium: Artificial Intelligence for Development (2010).

[9] A. Esmaeili, M. Khalili and A. Pakgohar, "Determining the road defects impact on accident severity; based on vehicle situation after accident, an approach of logistic regression," 2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE), Langkawi, 2012, pp. 1-4.

[10] M. A. A. Mamun, J. A. Puspo and A. K. Das, "An intelligent smartphone based approach using IoT for ensuring safe driving," 2017 International Conference on Electrical Engineering and Computer Science (ICECOS), Palembang, 2017, pp. 217-223.

[11] T. Adhikary, A. K. Das, M. A. Razzaque, A. Almogren, M. Alrubaian, and M. M. Hassan, "Quality of Service Aware Reliable Task Scheduling in Vehicular Cloud Computing," Mobile Networks and Applications, Volume 21, Issue 3, pp 482-493, June 2016.

[12] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research (2011), vol. 12, pp. 2825-2830.

[13] H. I. Bulbul and Ö. Unsal, "Comparison of Classification Techniques used in Machine Learning as Applied on Vocational Guidance Data," 2011 10th International Conference on Machine Learning and Applications and Workshops, Honolulu, HI, 2011, pp. 298-301.

[14] R. Wang, "AdaBoost for Feature Selection, Classification and Its Relation with SVM", A Review. Physics Procedia (2012), vol. 25, pp. 800-807.

[15] Z. Zhang, "Introduction to machine learning: k-nearest neighbors," Annals of Translational Medicine (2016), vol 4, issue no. 11, pp.218-218.

[16] I. Rish, "An Empirical Study of the Naïve Bayes Classifier," IJCAI 2001 Work Empir Methods Artif Intell (2001), vol 3.