Prediction of Road Accident and Severity of Bangladesh Applying Machine Learning Techniques

Joy Paul
Electrical & Computer Engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh
joypaul13848@gmail.com

Zerin Jahan
Electrical & Computer Engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh
zerinjahan.28@gmail.com

Kazi Fahim Lateef
Electrical & Computer Engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh
kazifahim.ruet15@gmail.com

Md. Robiul Islam
Electrical & Computer Engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh
robiulruet00@gmail.com

Sagor Chandro Bakchy
Electrical & Computer Engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh
sagorchandro.10@gmail.com

2020 IEEE 8th R10 Humanitarian Technology Conference (R10-HTC) | 978-1-7281-1110-0/20/$31.00 ©2020 IEEE | DOI: 10.1109/R10-HTC49770.2020.9356987

***Abstract*—Road accidents in Bangladesh have become widespread nowadays. This not only damage our economies but also affect many families as well. Earlier researchers in Bangladesh had suggested a few approaches to machine learning and computer vision, but they had some limitations. Prior work on this issue mainly focused on either the possibility of an accident or the degree of severity. Some works showed low accuracy due to deficiency in record which is a major problem. They either predicted accident with one or two specific factors or only found the factors related to the accident. Through this paper, we proposed a multiclass model in which we combined both the prediction of accidents and their corresponding severity to develop a better model to avoid road collisions. We also merged five accident related casualties to properly interpret the nature of the accident. Analyzing sixty factors of five causalities we used different machine learning algorithms for prediction. Among them Decision Tree, Random Forest, Multilayer Perceptron and Categorical Naive Bayes showed acceptable result, but the best outcome obtained with Decision Tree. This algorithm obtained a strong accuracy of 99.77% for accident prediction and 99.80% for severity prediction With an F1 score of 98.68% and 99.80%.**

***Index Terms*—Road Accident, Severity, Multiclass, Machine Learning, Multilayer Perceptron, Categorical Naive Bayes.**

I. INTODUCTION

Road accidents are increasing daily in Bangladesh by leaps and bounds. It's the most gruesome of everyday occurrences. It happened because of low-skilled and reckless driving, defective vehicles and Road congestion, poor traffic regulation, lack of knowledge of road law and the use of roads by citizens and so on. A survey was drawn up in 2019 based on 11 leading newspapers, the online news portal and TV channels which also showed that over 5000 causalities in road accidents across Bangladesh were reported. There were 788 more civilians killed compared to 2018 [1][2].

Some researchers mainly focused on accident detection, assessment of severity and analysis of road accident patterns. The works are related to road accident analysis like road accident factors and occurrence identification, severity analysis and so on. Previously various research works had been done regarding this issue but that's not enough and more appropriate and accurate automated system is required in the context of Bangladesh.

In [3], M. M. L. Elahi et al. proposed a method that can learn traffic pattern through video data captured from roadside and can track various features of a vehicle by using visionbased technique. They used a probabilistic model Parzen Probabilistic Neural Network (PPNN). For the prediction of road accident, they detected the anomaly of vehicle by their model. Their proposed model got an accuracy of almost 85% for detecting unnatural special occurrence or situation. As features, they just used the velocity of different vehicles and the distance between them. They did not measure the severity of a particular accident with these characteristics.

In [4], Md. S. Satu et al. analyzed N5 national highway specifically Dhaka-Banglabandha highway and found out the traffic accident pattern. For that, they used Decision Tree induction algorithm. Performance of twelve Decision Tree classifiers were compared by them. Among them, the CDT (unpruned) found to be the best classifier. The AUC curve and precision of their findings were fine, but they got a low accuracy and recall. They had operated in a particular region of N5 highway with very few reports of 892 incidents.

In [5], Md. F. Labib et al. analyzed traffic accident by machine learning technique to classify the severity of accident into four categories and compared the performance of different classifier algorithms. The authors conducted an analysis of road accidents using four developed and most common

supervised machine learning algorithms- Decision Tree, K-Nearest Neighbors (KNN), Naive Bayes and Adaptive Boost (AdaBoost). Their best outcome was achieved by Adaboost which showed 80% accuracy. This research paper only aimed primarily to examine severity of road accident. They also used only 15 critical accident related features, which can't predict incident, but only helps explain severity.

In [6], J. sun et al. presented a hybrid approach combining a Support Vector Machine (SVM) approach with a clustering algorithm for K-means to estimate the probability of crashes. The Random Forest model was used to pick relevant and significant variables which was able to detect crash 5 to 10 minutes before the incident. The incident was calculated using 577 collisions and 5794 non-crash incidents using cross-validation and transfer ability of various models. The results indicate that the accident prediction model, combined with the four most significant variables selected using the Random Forest model, can provide adequate predictive output for crashes. The accuracy of the crash prediction model can be 78.0% with the combination of the clustering algorithm and SVM model. Their accuracy was poor on such a critical topic as an accident. Because of the lack of accurate weather data, their paper failed to consider it. This approach can not be implemented when data records are small. They only predicted the probability of a crash but failed to measure its severity.

In [7], M. Ghadge et al. built a smart phone based system for analyzing road conditions using an Accelerometer and GPS sensors. It is called Bumps Detection System (BDS) that uses Accelerometer to detect potholes and GPS to map potholes position on Google Map. Drivers will therefore be told in advance of the count of potholes on the road. They have used machine learning approaches. For building the model, K-means Clustering algorithm was applied on the training data. The Random Forest Classifier was used to validate this model using test data for better predictability. This paper is not about predicting an accident or its extent, it just predicts the potholes that are one of the causes of a road accident. Many facts relating to injuries are not taken into account here.

In [8], S. Sonal et al. analyzed a framework of road accident. They studied the factors behind the incidents and identified some of the key causes of the incidents. The factors related with it were age, speed, time, road type, weather condition etc. They were also plotted in a graph form. This paper is not prediction based and severity is also not mentioned here. They examine the details of the road accident only by plotting graphs. There are also very few reasons that they find in their research.

Most of the previous works limitations on predicting road accident were either they were focused on predicting accident or some of them only considered severity of the accident. Also most of the previous work was based on predicting accident associate with one or two factors only which will be not enough for a real situation. It will be a complete work if the level of accident as well as its consequence level can be predicted. This is the main goal of our work to remove those limitations of previous works and building a complete accident prediction model. In this paper, 5 main causalities were selected whose impact were huge on accident, they are place, month of the year, time of a day, junction type and vehicles. Each causalities contained several factors. In total 60 facts were considered associate with accident. Throughout this work, these causalities have been merged so that it can be clarified the combine effect of all causalities on accident and severity. Such values directly indicated both the level of accident and the severity which was predicted here. We also showed the comparative results of various machine learning methods. For this prediction, we used 3744 accident records and 2893 severity records of 10 years.

The remaining sections are organized as follows: Section II describes about proposed methodology. Section III presents the results and discussion. Section IV gives the conclusion of the work.

II. PROPOSED METHODOLOGY

Our proposed model included suitable dataset collection and pre-processing of the dataset through some steps. Then the pre-processed data was divided into training and test datasets. The training set carried a known output and the model learned about that data to be generalized later on to other data, also had test set to test prediction of the model on that sub-set. Then training set was fitted by various algorithms which in return gave us various models. Evaluating test dataset repeatedly gave the desired prediction model. Later, this model will also be able to predict output with the help of real time data. The design of proposed method is presented in Figure 1.

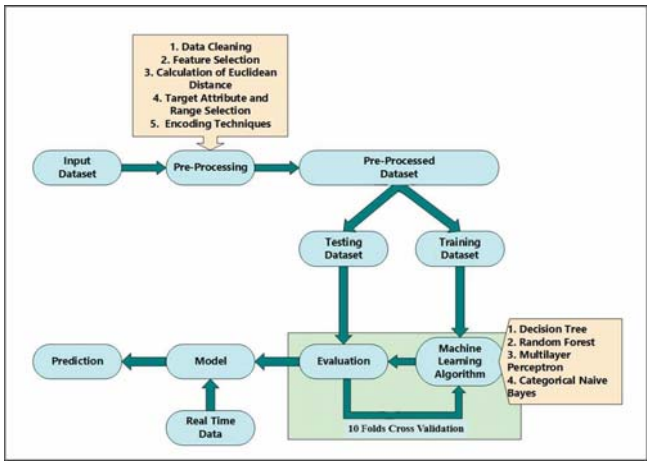

Fig. 1: Block diagram of road accident and severity prediction model.

A. Data collection

The most essential and one of the most challenging tasks was to collect the correct dataset. Datasets for road accidents in Bangladesh are not widely available. There are a few sources from which datasets had been collected. The National Traffic Accident Report 2007, BRTA was used for our prediction [9]. The dataset contains 3744 accounts of injuries and 2893

TABLE I: Collected Data Table

Month	number of accident								
	road environment			road class					
	urban	rural	total	national	regional	feeder	rural road	city	total
January	77	201	278	141	29	60	24	31	285
February	97	163	260	122	31	43	24	44	264
March	106	286	392	171	49	83	40	56	399
April	123	208	331	148	37	60	26	67	338
. .	. .	. .	. .	. .	. .	. .	. .	. .	. .

records of severity. It also covered five main casualties of accident of ten years. They are division, type of junction, vehicle involved, month and time. From PPRC report 2014, 13 years of accidents records of National highways were gathered [10]. Highway and division accident records were combined together as accident records of different places. The demo table from the dataset was given in TABLE I. Similarly, rest of the causalities were collected from the dataset.

B. Data pre-processing

It is the most important part of machine learning modeling. Without pre-processing, raw data cannot be transformed into useful and efficient format for creating machine learning model. The data pre-processing was completed by 5 steps. They are-

1) Data cleaning: Cleaning up data is a crucial move in almost any machine learning project. The main objective of data cleaning is to detect and delete errors and redundant data to establish a reliable dataset. This increases the consistency of the analytics training data and allows accurate decision making. In this paper, we removed some unnecessary data and some outliers. We had used box-plotting technique to detect the outliers and remove them manually.

2) Feature selection: Feature selection technique is the process to define the appropriate features and remove unnecessary, obsolete and spare attributes from the dataset. It is one of the most essential part of data pre-processing. In this work, the selected features are vehicle, place, time, junction type and month. These features have direct impact on road accident as well as its severity.

3) Calculating Euclidean distance: The Euclidean distance is a normal straight line that can measure two points in the Euclidean space. This measuring method is very useful in our work, because several points from different dimensions can easily be calculated. Euclidean distance is more effective than most of the multi-dimensional data calculation method. In this dataset, there were 5 features which represent 5 dimensions. The values of the features were normalized within the range of 0 to 1. These values from each factor were merged by Euclidean distance equation. The equation for n-dimension is-

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2}$$

A point was located on the plane after merging the values of each dimension. The distance from the center to that point was then measured. The measured distance was the final value which was used in our work. The equation looks like-

$$D = [(Place - 0)^2 + (Junction\ Type - 0)^2 + (Month - 0)^2 + (Time - 0)^2 + (Vehicle - 0)^2]^{1/2}$$

From the result of this equation, different combination of different factors can be identified easily.

TABLE II: Pre-processed Data Table

Place	Month	Time (Hr)	Junction Type	Vehicle	Possibility	Level
Dhaka	January	0-1	Cross Junction	Bicycle	45%	High
Dhaka	January	0-1	Cross Junction	Rickshaw	45%	High
Barisal	September	20-21	Tee Junction	Car	13%	Low
Barisal	September	20-21	Tee Junction	Jeep	13%	Low
Chittagong	January	0-1	Cross Junction	Bicycle	21%	Moderate
Chittagong	January	0-1	Cross Junction	Track	21%	Moderate
Chittagong	January	0-1	Cross Junction	Bus	33%	Moderate
. .	. .	. .	. .	. .	. .	. .

4) Target attribute and range selection: The target attribute is the outcome of a machine learning model. This is the dependent attribute of the pre-processed dataset. The values which were found from Euclidean distance calculation were divided into 5 levels. They are- Low (0%-20%), Moderate (21%-40%), High (41%-60%), Very High (61%-80%) and Extreme (81%- 100%). Each level indicates the sharpness of accident and severity possibility. This level column was selected as the target attribute. The final pre-Processed table is shown in Table II.

5) Encoding technique: Several machine learning algorithms are unable to operate directly on label data. They need numeric value for both input and output variables. Categorical data have to be converted into a numerical form. As both the independent and dependent column has categorical values, we used One Hot Encoding for converting the independent variables and Label Encoding for dependent variable.

C. Proposed system

After pre-processing, the final dataset contains 1,93,536 records which has the combination of every possible factors of each causalities. These records were our key concern as these can provoke an accident in certain ways.

Then the pre-processed dataset was divided into two parts. One part was used as training dataset and its amount was 80% of the main dataset. Another part was used as testing dataset which was 20% of the main dataset. Here, both the training and testing datasets were made by taking data randomly. Then the training set was fitted into different machine learning algorithms. To avoid overfitting situation, the process was repeated by applying k-fold cross validation function where the value of k was 10. After final evaluation, our desired model was ready for prediction. The model was finally ready to take real time data as prediction input. The algorithms used in this work is described below-

1) Decision Tree: Decision Tree is a supervised learning method that can be used for problems of classification and regression, but is preferred to solve problems of classification. This is a tree-structured classifier, with internal nodes representing a data set's features, decision rules are represented by branches of tree and each leaf node reflects the result. 'Entropy' is a function that is used to calculate the splitting output which we have used in Decision Tree. They are also used to measure a dataset's impurity or randomness. The equation of entropy is,

$$Entropy = \sum_{i=1}^{c} -p_i * \log_2(p_i)$$

The technique for choosing the split at each node is called splitter. Here 'best' is used as splitter. Another function used here is 'random state'. It monitors estimator randomness. The characteristics are often permuted randomly on every split. So it is assigned to 'None' parameter.

2) Random Forest: A Random Forest is an ensemble method that can perform both regression and classification tasks using multiple Decision Trees, a strategy called Bootstrap and Aggregation, usually referred to as bagging. The core idea behind this is the combination of many Decision Trees to determine the final product rather than depending on individual Decision Trees. This type of algorithm helps develop the manner in which complex data is analyzed by technologies. When conducting Random Forests based on classification data, the Gini index is sometimes used, or the formula used to determine how nodes are on a branch of a Decision Tree. For better outcome, Entropy is used in this work. which is,

$$Entropy = \sum_{i=1}^{c} -p_i * \log_2(p_i)$$

This approach takes the log base 2 of the probabilities, rather than using simple probabilities. Entropy determines how a Tree chooses to divide the data. In fact, it affects how the Decision Tree draws its boundaries. The major advantage is that it offers the additive property.

For building trees, bootstrap samples were used. N estimators is just the number of trees that are built before the maximum vote is taken. Higher number of trees will give better results, but slow down code. That's why we have taken 30 trees, offering better results as well as faster performance.

3) Multilayer Perceptrons: Multilayer Perceptrons is an artificial neural network. They consist of an input layer for receiving the signal, an output layer for determining or predicting the output, between these two, an unspecified number of hidden layers, which are the real Multilayer Perceptron computing engine. This algorithm used frequently in supervised learning problems. The equation is,

$$y = \varphi \sum_{i=1}^{n} (w_i x_i) + b$$

$$y = \varphi(w^T x + b)$$

here, w is weights of vector, x is inputs, b is bias and φ is activaion function. In this work, hidden layer size is 100 units with single hidden layer, activation function used is 'relu'. Solver function is a weight optimizer, here 'adam' is used which works well in large dataset. Here, batch size is 200, learning rate is 0.1 and max iteration is taken 500.

4) Categorical Naive Bayes: The Naive Bayes algorithm is a supervised classifying machine learning algorithm. It learns the likelihood of an object having those features belonging to a specific category. It is based on the Bayes Theorem. It gives us a method for calculating the conditional probability. The Naive Bayesian model is simple to construct, with no complex iterative parameter estimation, making it especially useful for very large datasets. Among different types of Naive Bayes algorithm, Categorical Naive Bayes has been used in this work. It implies that every attribute that the index represents has its own categorical distribution. The probability of category t in feature i given class c is,

$$P(x_i = t|y = c; \alpha) = \frac{N_{tic} + \alpha}{N_c + \alpha n_i}$$

here, N_{tic} is the number of times category t appears in the samples x_i, which belongs to class c. N_c is the number of samples and α is the smoothing parameter which is taken 1. n_i is the number of available categories of feature i.

III. RESULT AND DISCUSSION

Different algorithms were applied for our work. Among them, 4 algorithms were found preferable for this prediction work which are Decision Tree, Random Forest, Multilayer Perceptron (MLP) and Categorical Naive Bayes. The accuracy, precision, recall, F1 score and standard deviation of each algorithm were shown in Table III. It can be seen from the table that the accuracy of the Decision Tree and MLP is almost equal and higher than the other two. But F1 score of Decision Tree is greater than MLP. Also the run time of Decision Tree is very fast compare to MLP. Decision Tree also showed better result than Random Forest. For large dataset, validation on a complete set of Decision Trees is higher than a subset

TABLE III: Result of Road Accident and Severity Prediction Model

Algorithm	Type	Accuracy	Precision	Recall	F1 Score	Standard Deviation
Decision Tree	**Accident**	99.77%	98.68%	98.67%	98.68%	0.03%
	Severity	99.80%	99.83%	99.77%	99.80%	0.02%
Random Forest	**Accident**	99.55%	99.74%	97.94%	98.80%	0.05%
	Severity	99.56%	99.64%	99.61%	99.63%	0.05%
Multilayer Perceptron	**Accident**	99.77%	99.19%	97.49%	98.30%	0.05%
	Severity	99.82%	99.77%	99.82%	99.79%	0.03%
Categorical Naive Bayes	**Accident**	93.85%	91.69%	85.79%	88.24%	0.16%
	Severity	97.84%	98.34%	97.01%	97.59%	0.06%

of Decision Tree for the score estimation. Due to bootstrap samples in Random Forest, the out of bag (OOB) score showed some error that effects on prediction accuracy. So in overall it can be concluded that Decision Tree is best suited for the prediction of road accident and its severity. Its accuracy is 99.77% for accident and 99.80% for severity prediction with an F1 score 98.64% and 99.80%. The standard deviation for Decision Tree is 0.03% and 0.02% respectively.

Receiver operating characteristics (ROC) curve of Decision Tree algorithm has shown in Figure 2 and Figure 3.

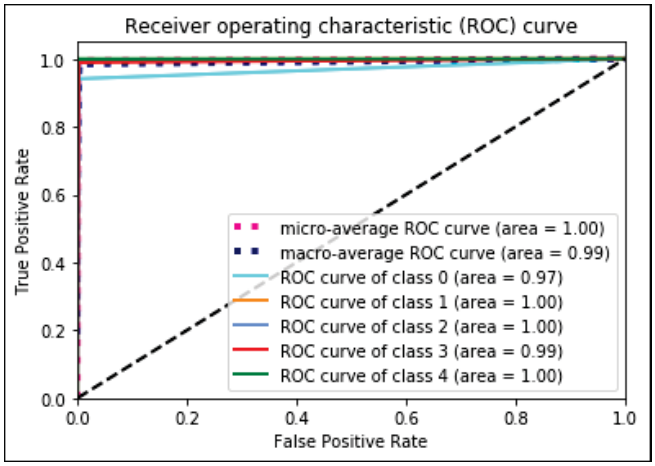

Fig. 2: ROC curve of Decision Tree algorithm for accident prediction.

It has two parameters where one is true positive rate (TPR) or sensitivity and another is false positive rate (FPR) or 1-specificity at different classification thresholds. The equations are-

$$TPR = \frac{TP}{TP + FN}$$

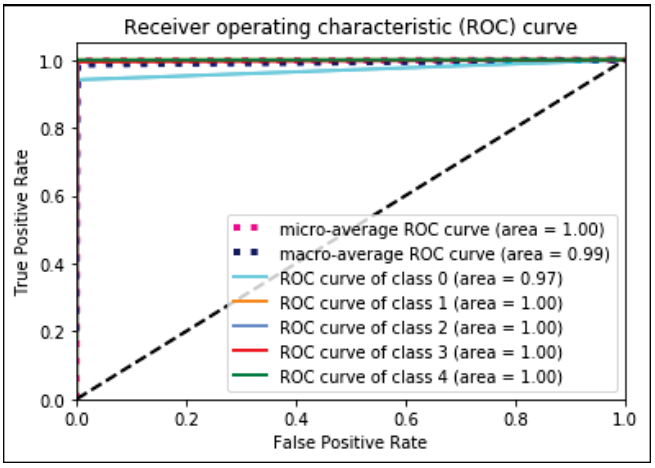

Fig. 3: ROC curve of Decision Tree algorithm for severity prediction.

$$FPR = \frac{FP}{FP + TN}$$

A clear idea of the micro and macro average of our prediction model can be obtained from the ROC. From the curve micro average of the model is 1.00 and its macro average is 0.99. Here micro average is greater than micro average. So it can be said that larger labels are accurately classified, whereas small ones are classified with an accuracy of 99%. The accuracy of each class is also reflected in the ROC curve.

CONCLUSION

In Bangladesh, there is no road accident prediction or severity checking system available now. In terms of reducing accidents, this prediction model can be very useful for improving road management, road design and road safety. Building a system for a country like ours is also expensive but it can save thousands of valuable lives that could be a great strength. We evaluated the prediction model by modifying previous years' statistical datasets showing 99.77% accuracy in accident prediction and 99.80% accuracy in severity prediction. For this task, few algorithms worked nicely, but the best results have been achieved with Decision Tree algorithm with an F1 score of 98.68% and 99.80% respectively. People can also have an overview of the factors affecting the accident and its impacts. Initially it can be integrated into mobile apps that are available in the hands of every person. Various devices such as GPS, speedometer or other can be used to get the real time data. It can be implemented in traffic system and autonomous car. Further features can be added to identify the road accident and severity more specifically. Current ride hailing services in Bangladesh can be combined with this prediction model.

REFERENCES

[1] S. O. Report, "Road crashes kill 7,855 people in 2019," The Daily Star, 11-Jan-2020. [Online]. Available:

https:// www.thedailystar.net/ country/ road-accidents-kill-7855-people-in-2019 -in-bangladesh -1852684. [Accessed: 24-Feb-2020].
[2] "Bangladesh: Alarming rise in road crashes," Anadolu Ajansı. [Online]. Available: https://www.aa.com.tr/ en/ asia-pacific/ bangladesh-alarmingrise- in-road-crashes/ [Accessed: 24-Feb-2020].
[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, doi: 10.1109/ICIEV.2014.6850780
[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, doi: 10.1109/R10- HTC.2017.8289059
[5] M. F. Labib, A. S. Rifat, M. M. Hossain, A. K. Das and F. Nawrine, "Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh," 2019 7th International Conference on Smart Computing Communications (ICSCC), Sarawak, Malaysia, Malaysia, 2019, pp. 1-5, doi: 10.1109/ICSCC.2019.8843640
[6] J. Sun and J. Sun, "Real-time crash prediction on urban expressways: identification of key variables and a hybrid support vector machine model," in IET Intelligent Transport Systems, vol. 10, no. 5, pp. 331-337, 6 2016, doi: 10.1049/iet-its.2014.0288.
[7] M. Ghadge, D. Pandey and D. Kalbande, "Machine learning approach for predicting bumps on road," 2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Davangere, 2015, pp. 481-485, doi: 10.1109/ICATCCT.2015.7456932.
[8] S. Sonal and S. Suman, "A Framework for Analysis of Road Accidents," 2018 International Conference on Emerging Trends and Innovations In Engineering And Technological Research (ICETIETR), Ernakulam, 2018, pp. 1-5, doi: 10.1109/ICETIETR.2018.8529088.
[9] "National Road Traffic Accident, 2007," [online]. Available: http:// www.rhd.gov.bd/ Documents/ RoadDesignAndSafety/ NationalRoadTrafficAccidentReport2007/ NationalRoadTrafficAccidentReport2007. pdf. [Accessed: 22 –Dec-2019].
[10] "Road Safety in Bangladesh Ground Realities and Action Imperatives," [online]. Available: http:// www.brac.net/ images/reports/[Accessed: 22- Dec-2019]
[11] M. Zheng et al., "Traffic Accident's Severity Prediction: A Deep-Learning Approach-Based CNN Network," in IEEE Access, vol. 7, pp. 39897-39910, 2019, doi: 10.1109/ACCESS.2019.2903319.