

Identify Discriminatory Factors of Traffic Accidental Fatal Subtypes using Machine Learning Techniques

W.Z. Loskor¹

Department of Science and Humanities
Bangladesh Army International University of Science and
Technology, Cumilla-3501, Bangladesh

Sharif Ahamed²

Department of Computer Science and Engineering
Gono Bishwabidyalay, Dhaka-1344
Bangladesh

Abstract—In today's world, traffic accidents are one of the main reasons of mortality and long-term injury. Bangladesh is no exception in this case. Several vehicle accidents each year have become an everyday occurrence in Bangladesh. Bangladesh's largest highway, the Dhaka-Banglabandha National Highway, has a significant number of accidents each year. In this work, we gathered accident data from the Dhaka-Banglabandha highway over an eight-year period and attempted to determine the subtypes present in this dataset. Then we tested with various classification algorithms to see which ones performed the best at classifying accident subtypes. To describe the discriminatory factors among the subtypes, we also used an interpretable model. This experiment gives essential information on traffic accidents and so helps in the development of policies to reduce road traffic collisions on Bangladesh's Dhaka-Banglabandha National Highway.

Keywords—Traffic accident; clustering analysis; machine learning; feature selection; classification; discriminatory factors

I. INTRODUCTION

Traffic accidents have become one of the leading causes of loss of life and property. The likelihood of traffic accidents is increasing as the number of vehicles and roads increases. In 2020, total of 4,891 vehicle accidents in Bangladesh killed 6,686 people and wounded 8,600 others [1]. As a result, 18 individuals died in traffic accidents each day across the country. In its yearly road accident observing report for 2020, Bangladesh Passengers Welfare Association (BPWA) disclosed these data. According to the Accident Research Institute of Bangladesh University of Engineering and Technology (BUET), 56,987 individuals have perished in 58,208 vehicle accidents in Bangladesh in the last two decades. Many researchers have examined road accident datasets and used various machine learning methods to predict the risk of an accident; some of their findings are summarized in the Literature Review section. All of the research efforts on these datasets are aimed at classifying the risk of a car accident. Many organizations exist in many countries around the world to maintain road safety in order to reduce the threat of fatal road traffic accidents. Researchers, more-over, used a variety of techniques, particularly statistics methods, to define the reasons of traffic road accidents through a historical path traffic road dataset. Using various data mining tools and techniques, the data mineworkers investigated different parameters or variables for the reasons of traffic accidents besides diver behaviors. A lot of researchers used to expend a significant

amount of time attempting to find the greatest performing data mining procedure for mining the traffic road accidents dataset.

In this research, we collected traffic accident data in the DBH from the Accident Research Institute (ARI), BUET, from 2007 to 2015, and we only used fatal data records. This research aims to identify only accident fatal subtypes and identify important discriminant features that will assist authorities in better understanding accident risks.

The rest of this paper is as follows. Section II is dedicated to related activity. Section III discusses traffic accident data analysis and methodology. The findings of the experiments are contained in Section IV. Finally, Section V summarizes the work's findings.

II. LITERATURE REVIEW

Several studies have been launched to investigate traffic accident data using various approaches. In 2021, M. Bobermin et al. suggested a novel framework based on Clustering Analysis for the definition of driving simulator experiments [2]. Amir Mohammadi Amiri et al. (2021) used five different hotspot identification algorithms. They are as follows: Getis-Ord G_i^* , Average Nearest Neighbor, kernel density KDE, Global Moran's I, and mean center. Global Moran's I approach outperforms other methods in locating hotspots, according to the findings [3]. In the same year, F. Francis used Hierarchical clustering and K-means clustering the same year to merge the spatially specified groupings into six clusters based on the similarity of their temporal patterns [4]. Dooti Roy et al. (2021) introduced a two-stage clustering-based technique based on SOM followed by neural gas clustering to build a data-driven taxonomy of bus crashes [5]. Rocio Suarez-del Fuego et al. (2021) used unsupervised clustering methods to identify badly injured, belted occupants into groups, bio-mechanical characteristics, and accident severity [6]. The applicability of the k-prototypes clustering method in massive truck-involved crashes was investigated by Syed As-Sadeq Tahfim et al. (2021). To predict the severity of injuries in major truck incidents, four gradients boosted decision trees techniques were used to the dataset and individual clusters [7]. Filbert Francis et al. (2021) found high-risk areas in Dar es Salaam for motorcycle-related injuries. Three distinct motorcycle injury hotspot clusters have been discovered [8].

Mert Ersen et al. (2020) used the Kernel Density approach to examine statistical analyses based on accident kinds. The

Kernel Density approach has been found to produce better visual results than other spatial methods [9]. Seyed Mohsen Hosseinian et al. (2020) investigated the effect of different factors on the severity of urban traffic accidents in Rasht metropolis by using frequency analysis of accident data [10]. Qiuru Cai (2020) created the Apriori algorithm to mine the rules that govern the relationship between risk issues and the cause of traffic accidents on urban roads [11]. In 2020, Yunduan Lin et al. used crowdsourcing data to investigate the technique of predicting the complicated behavior of traffic flow evolution after traffic accidents. According to the results, NN outperforms the other models [12].

Sharaf AlKheder et al. (2020), on the other hand, used three data mining algorithms to conduct a thorough investigation of risk factors associated to the severity of traffic accidents. In comparison to previous models, the Bayesian network was more accurate in predicting the variables [13]. Yang Yong Zheng et al. (2020) discovered the elements that influence traffic accidents in undersea tunnels and developed a prediction model for undersea tunnel traffic accidents [14]. Marjana Cubranic-Dobrodolac et al. (2020) suggested a model for assessing and making decisions about a driver's proclivity for traffic accidents that is based on an estimation of the driver's psychological attributes [15].

Based on single-vehicle crashes, Natalia Casado-Sanz et al. (2020) found the contributing factors to a fatal outcome. The most relevant factors related with driver injury severity were identified using a Multinomial Logit model [16]. Human error was highlighted as a major contributory element in road traffic accidents by Asad Iqbal et al. (2020), and the Salt Range was classified as a black spot on account of vehicle braking failure [17]. Minglei Song et al. suggested a road accident prediction model based on joint probability density feature extraction from big data in 2019 [18]. Eight impact factors were chosen by Cheng Zhang et al. (2019), and the Bayesian network was the best model to potentially predict road accident black spots [19].

On accident datasets, Sadiq Hussain et al. (2019) used J48, Multi-layer Perceptron, and BayesNet classifiers. The Multi-layer Perceptron classifier performed well in the study, with an accuracy of 85.33 percent [20]. According to Juan Pineda-Jaramillo et al. (2019), road traffic collisions occur in all clusters, although zones surrounded by landscapes and parks have more run overs than fallen residents [21].

III. MATERIALS AND METHODS

A. Dataset Description

From 2007 to 2015, we collected 1283 data of traffic incidents on the N5NH (N5 National Highway) from MAAP5 of the Accident Research Institute (ARI), BUET Accident report forms have been distributed to various police stations in Bangladesh. There are two parts to the accident report form. One is the main form, while the other is the supplementary form. Each accident record is filled out on the main form, where the top 37 columns depict preliminary information on the severity of the traffic accident. Accidental vehicle information is stored in columns 38-45, whereas driver information is stored in columns 68-52. Furthermore, columns

53-58 and 59-64 contain detailed information about passengers and pedestrians, respectively. 65-67 columns, on the other hand, are utilized to identify the causes of an accident.

B. Proposed Discriminatory Factors of Fatal Subtype Detection Model

Fig. 1 depicts an overview of our process for identifying discriminatory characteristics, which is briefly detailed step by step.

Step 1: Data Preprocessing and Analysis: In this section, all portions of the data, including the route number, have a recurring value. Furthermore, some data, such as XY map, X coordinate, and Y coordinate, have no values. 67 percent values are also missing in the kilometer post and 100-meter attributes. As a result, we decide experimental data eliminate them. We have separated 1002 fatal data from 1283 entries. Numeric and nominal values are blended throughout all records. All nominal values have been converted to numeric values. The features dealing with vehicle details (columns 38-45), driver details (columns 46-52), passenger details (columns 53-58), and pedestrian details (columns 59-64) are deleted from empirical traffic accident data as unusable features. Table I illustrates these characteristics with a brief explanation. Report Number, FIR Number, and Thana are not deemed particularly important and are detached from the empirical data. Formerly, we construct a hit-map to detect linked traits (see Fig. 2). So, we see that the number of vehicles is correlated to the number of driver and pedestrian victims. As a result, we eliminate these two attributes. The remaining attributes are useful in determining the more accurate outcomes in this experiment.

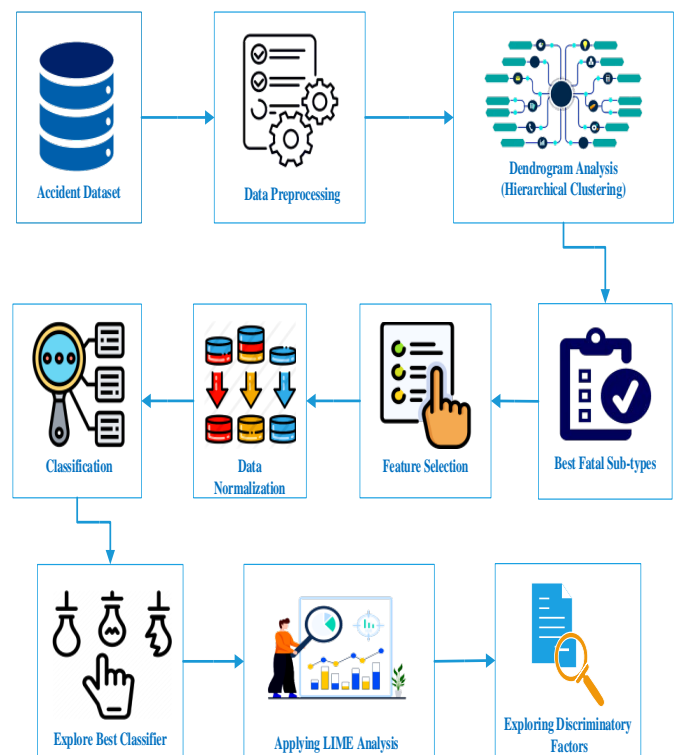


Fig. 1. Proposed Identifying Discriminatory Factors of Fatal Subtype Model.

TABLE I. ACCIDENT DATASET DESCRIPTION

Sl.	Feature Name	Feature Description
1.	Report Number	Accident report number.
2.	FIR Number	Accident FIR number.
3.	Thana	Thana number.
4.	District	District name.
5.	Number of Pedestrian Victims	Pedestrian victims' number.
6.	Number of Passenger Victims	Passenger victims' number.
7.	Number of Driver Victims	Driver victims' number.
8.	Number of Vehicles	Vehicles intricate numbers.
9.	Day of Week	The casualty occurred the day.
10.	Month	The casualty occurred a month.
11.	Date of Month	Casualty date.
12.	Year	The casualty occurred a year.
13.	Accident Type	Collision type of occurred accident.
14.	Movement	Road variety.
15.	Type of Junction	Diversity of junction.
16.	Traffic Control	The behavior of regulating traffic.
17.	Divider	Existsents of the divider.
18.	Weather	Condition of weather when the accident occurred.
19.	Light	Light condition on the road surface.
20.	Road Geometry	The geometry of road surface.
21.	Severity	Type of a casualty.
22.	Type of Surface	Variation of the road surface.
23.	Condition of Surface	Condition of the road surface.
24.	Surface Quality	The road quality.
25.	Type of Location	Accident location.
26.	Read Feature	Variation of road.
27.	Road Class	Road Category.
28.	Responsible Factors 1	An accident-related factor.
29.	Responsible Factors 2	An accident-related factor.
30.	Responsible Factors 3	An accident-related factor.

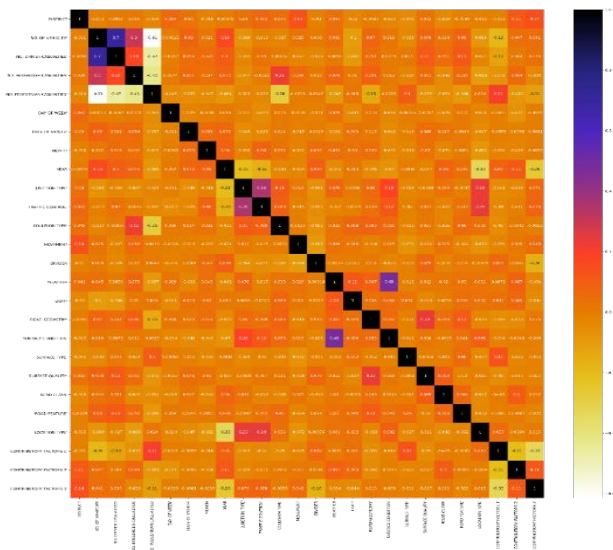


Fig. 2. Hit-Map for Identifying Correlated Features.

Step 2: Employing Clustering Analysis: Clustering analysis is a technique for categorizing cases into comparable important groups based on their unique characteristics. The agglomerative mode of hierarchical clustering algorithms divides every cluster into small sub clusters or assembles them into super clusters on a regular basis [2]. In a hierarchical architecture known as a dendrogram, the connection between each pair of clusters is determined by the medium of dissimilarity or similarity. However, we apply this strategy to generate numerous fatal sub-types in the accident dataset. To reveal the predictability of the proposed model, these subtypes are considered as separate class labels.

Step 3: Chi-Square Test for Feature Ranking: When two attributes are independent, the executed count is close to the awaited count, resulting in a reduced Chi-Square value. The higher the Chi-Square number, the more dependent the property is on the response. Then it can be chosen for model training. However, in our research, we rank attributes in order to identify the appeasement set of most significant factors that result in the maximum accuracy. After identifying the subtypes, we utilized the Chi-Square test feature ranking technique on the accidental dataset to discover the optimal set of most significant attributes.

Step 4: Normalization: Normalization is a data preparation method used frequently in machine learning. Its major purpose is to use a common scale to adjust the values of numeric columns in the dataset without losing information. In this paper, we use the MinMaxScaler class in Python to normalize fatal sub-types data of the most significant attributes and create a balanced dataset with appropriate structures.

Step 5: Classification Approach: Classification Approach: To compute the class of objects, Classification is a mode of function discovery in which concepts or classes are interpreted and isolated whose label is unfamiliar to the target. On the normalized dataset, we use six machine learning classification algorithms to identify the observed sub-types: Decision Tree (DT), Naive Bayes (NB), K-Nearest Neighbor (KNN), Random Forest (RF), Multi-layer Perceptron (MLP) and Support Vector Machine (SVM). Previous studies of road accidents have used these classifiers extensively [12, 19]. To find the best classifier with the highest accuracy, some evaluation metrics (see Table II) such as Accuracy, F1-Score, and AUROC were used.

Step 6: Exploring Discriminatory Factors Using LIME: Local Interpretable Model-Agnostic Explanations (LIME) is an algorithm that can interpret a model by distracting the data sample input and knowing how the predictions vary. LIME produces a set of interpretations that show how each feature performs against a prediction for a single sample, which is a type of local interpretability. We use LIME to find which characteristics contributed the most to attaining the best result in categorizing the sub-types on the dataset using all the attributes for the best per-forming classifier. As a result, we have discriminatory variables for the classification of subtypes.

TABLE II. EVALUATION MATRIX

Metrics	Details	Formula
Accuracy	Accuracy is the ratio of the number of all correct predictions and the total number of the data.	$Acc. = \frac{TP+TN}{TP+FN+FP+TN}$
F1-Score	F1-Score is a harmonic mean of precision and recall.	$F = \frac{2 * Recall * Precision}{Precision + Recall}$
AUROC	AUC is the summation of all TP rates and TN rates divided by two.	$AUROC = \frac{TP\ rate + TN\ rate}{2}$

IV. RESULT AND DISCUSSION

In our work, we identified the clusters using hierarchical clustering on the dataset. Each cluster is defined as an observed subtype present in the accident dataset. We utilized the Chi-Square test to determine the most significant features after identifying the subtypes. Then, on the selected features, we performed data normalization using Python's MinMaxScaler class. On the datasets, we used various classification algorithms (i.e., DT, KNN, NB, RF, SVM, MLP) to classify the observed subtypes. Classification is accomplished through the use of 10-fold cross-validation. Finally, we used LIME to interpret features for discriminatory factors. Jupyter Notebook version 6.1.4 is used for all of the experiments.

A. The Analogy of Performance of Distinct Classifiers

In this study, Hierarchical clustering yields two subtypes (subtype-1 and subtype-2) (see Fig. 3). The ratio of subtype-1 to subtype-2 is found to be the same. As a result, no data balancing was required. The Chi-Square test result is displayed in Table III. As can be seen in the table, the features are ordered in ascending order depending on their P-Values. The feature with the lower P-Value is more important. We took the different number of features (i.e., 5, 10, 15, 20, and 24) from those significant feature lists and applied different classifiers to the datasets. The experimental results for different classifiers (described in section III.B) utilized to categorize the sub-types are shown in Tables IV, V, VI, VII, and VIII. To explain our findings, we used a variety of evaluation matrices (Accuracy, F1- score, and AUROC). Performance Analysis of All Significant Features is shown in Fig. 4.

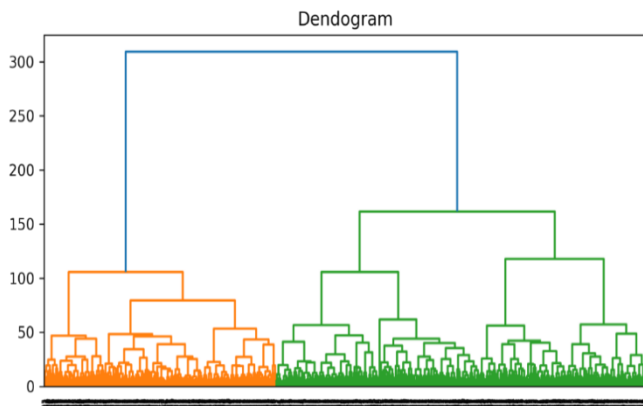


Fig. 3. Dendrogram Analysis to Identify the Fatal Subtypes.

TABLE III. P-VALUES IN ASCENDING ORDERS FOR DIFFERENT FEATURES

Features	P-Value
Date of Month	0.000000
Number of Passenger Victims	0.000017
Month	0.000634
District	0.003480
Surface Quality	0.010259
Type of Junction	0.015621
Light	0.047291
Accident Type	0.059001
Condition of Surface	0.240108
Type of Location	0.252266
Responsible Factors 1	0.419230
Number of Vehicles	0.479777
Day of Week	0.495156
Traffic Control	0.510946
Weather	0.545987
Divider	0.582690
Responsible Factors 2	0.681025
Road Class	0.885795
Year	0.895612
Movement	0.942830
Responsible Factors 3	0.954226
Road Geometry	0.957346
Road Feature	0.988516

TABLE IV. PERFORMANCE ANALYSIS OF DIFFERENT CLASSIFIERS (5 SIGNIFICANT FEATURES)

Classifier	Accuracy	F1-Score	AUROC
DT	99.80	99.80	99.80
KNN	91.91	91.91	97.91
NB	97.00	97.00	98.81
RF	99.90	99.90	100.00
SVM	97.60	97.60	99.77
MLP	99.30	99.50	99.94

TABLE V. PERFORMANCE ANALYSIS OF DIFFERENT CLASSIFIERS (10 SIGNIFICANT FEATURES)

Classifier	Accuracy	F1-Score	AUROC
DT	99.80	99.60	99.80
KNN	77.34	77.34	85.07
NB	95.61	95.61	98.10
RF	99.80	99.80	100.00
SVM	95.21	95.21	99.11
MLP	98.20	98.60	99.78

TABLE VI. PERFORMANCE ANALYSIS OF DIFFERENT CLASSIFIERS (15 SIGNIFICANT FEATURES)

Classifier	Accuracy	F1-Score	AUROC
DT	99.60	99.70	99.50
KNN	65.06	65.06	70.74
NB	94.91	94.91	97.96
RF	99.70	99.70	100.00
SVM	92.32	92.32	98.29
MLP	97.60	97.70	99.76

TABLE VII. PERFORMANCE ANALYSIS OF DIFFERENT CLASSIFIERS (20 SIGNIFICANT FEATURES)

Classifier	Accuracy	F1-Score	AUROC
DT	99.50	99.60	99.50
KNN	62.26	62.26	65.39
NB	93.61	93.61	97.31
RF	99.70	99.60	100.00
SVM	92.31	92.31	98.10
MLP	97.10	96.70	99.65

TABLE VIII. PERFORMANCE ANALYSIS OF DIFFERENT CLASSIFIERS (24 SIGNIFICANT FEATURES)

Classifier	Accuracy	F1-Score	AUROC
DT	99.60	99.60	99.50
KNN	55.38	55.38	57.79
NB	92.71	92.71	97.05
RF	99.80	99.90	100.00
SVM	90.31	90.31	97.62
MLP	96.31	96.31	99.61

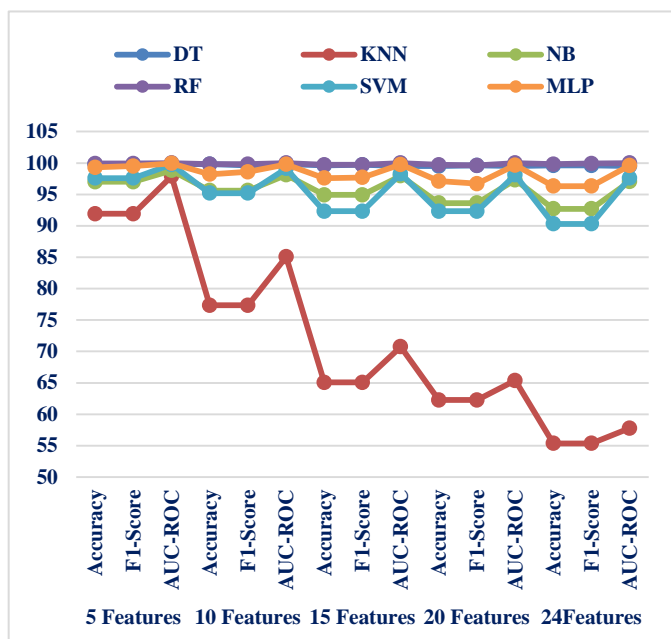


Fig. 4. Performance Analysis of All Significant Features.

From the table, we can see that RF outperforms all other classifiers in terms of accuracy, F1-score, and AUROC (see Table IV to VIII) for all different number of features. For the dataset with only the five most significant features, RF achieves 99.90% accuracy, 99.90% F1-score, and 100.00% AUROC. This is the highest possible score in our study. For other sets of features, RF receives slightly different scores. Furthermore, all other classifiers, with the exception of KNN, achieve high results (i.e., above 90%). It is also worth noting that all classifiers performed best with the most significant 5 feature subset, and their performance degraded as the number of features used increased.

B. Interpretation of Features for Discriminatory Factors

We used LIME on the dataset (with all features) to find the highest performing RF classifier and determine which features contributed the most to correctly categorizing the subtypes. As a result, we obtain discriminatory factors for sub-type classification. The features that contributed the most to identifying distinct sub-types are shown in Fig. 5, which differs significantly from the statistical result we obtained using the Chi-Square test for important features. The most crucial feature identified for subtype classification is 'Road Feature,' as seen in Fig. 5. The relationship between road features and accident subtype is seen in Table IX. The table shows that "Road Feature" - General is the most prevalent cause of accidents and has about the same ratio in both categories. "Road Feature"-Bridge is twice as common in subtype-1 as in subtype-2. Culverts and Speed Breakers are more common in subtypes 1 and 2, respectively. The second most significant attribute is 'Road Class.' The relationship between road classes and accident subtypes is shown in Table X. It is apparent that the most prevalent type of accident is 'Road Class'- Natural and has nearly the same ratio in both subtypes. 'Road Class'- Feeder is twice as common in subtype-1 as it is in subtype-2. 'No. of Vehicles' is the third most essential aspect. The relationship between the 'No. of Vehicles' and the types of accidents is seen in Table XI.

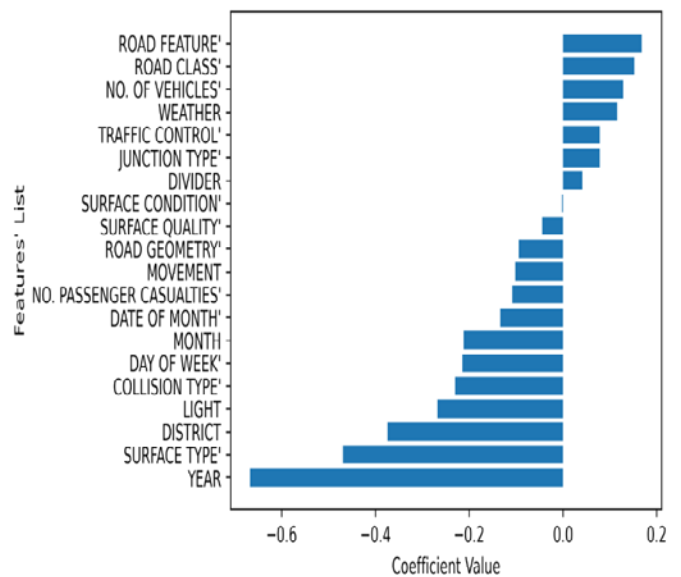


Fig. 5. Interpreting Important Features using LIME.

TABLE IX. THE RELATION BETWEEN "ROAD FEATURE" AND ACCIDENT SUBTYPE

Road Feature	Cluster	No. of instances
General	Subtype-1	484
	Subtype-2	486
Bridge	Subtype-1	14
	Subtype-2	7
Culvert	Subtype-1	1
	Subtype-2	3
Speed Braker	Subtype-1	2
	Subtype-2	3
Narrow	Subtype-1	1
	Subtype-2	1

TABLE X. THE RELATION BETWEEN "ROAD CLASS" AND ACCIDENT SUBTYPE

Road Class	Cluster	No. of instances
Natural	Subtype-1	488
	Subtype-2	486
Feeder	Subtype-1	4
	Subtype-2	2
Regional	Subtype-1	6
	Subtype-2	8
Rural	Subtype-1	2
	Subtype-2	3
City	Subtype-1	2
	Subtype-2	1

TABLE XI. THE RELATION BETWEEN "NO. OF VEHICLES" AND ACCIDENT SUBTYPE

No. of Vehicles	Cluster	No. of instances
1	Subtype-1	299
	Subtype-2	265
2	Subtype-1	198
	Subtype-2	234
3	Subtype-1	4
	Subtype-2	1
5	Subtype-1	1

TABLE XII. THE RELATION BETWEEN "WEATHER" AND ACCIDENT SUBTYPE

Weather	Cluster	No. of instances
Clean/Fair	Subtype-1	461
	Subtype-2	458
Fog	Subtype-1	18
	Subtype-2	25
Rain	Subtype-1	21
	Subtype-2	17
Wind	Subtype-1	2

The table shows that for subtype-1, 'No. of Vehicles' 1 is more prevalent, while for subtype-2, 'No. of Vehicles' 2 is more common. 'No. of Vehicles' 5 appears only in subtype-1. 'Weather' is the fourth most essential feature. The relationship between weather and accident subtypes is seen in Table XII. The data shows that 'Weather'- Clean/ Fair has a higher risk of accidents. 'Weather'-Rain is exclusively related to the subtype-1. As shown in Fig. 5, 'Traffic Control,' 'Junction Type,' and 'Divider' all play a role in subtype classification. Other features in the list have a negative relationship with classification into subtypes.

C. Relative Studies and Implication

Many researchers have looked into road accident classification, and some of their findings are included in Section 2. We discovered that all of the research efforts on these datasets were focused on classifying the risk of a traffic collision. However, no attempt was made to identify the various forms of road accidents (as far as our knowledge). We used clustering to determine the subtypes in this study, and the appropriate number of clusters for each dataset is justified. Then we used classifiers to find the best classification of subtypes using relevant feature sets. Then, using the explainable AI technique, we showed key features that contributed to the identification of subtypes. Identifying subtypes will assist authorities in better understanding accident risks. We discovered important elements that will assist them in identifying sub-types as well as accident risks.

V. CONCLUSION AND FUTURE WORK

Traffic accidents are viewed as a global issue that results in fatalities and serious injuries. The study of traffic accident data assists the traffic department in identifying the primary persuasive elements of accidents and revealing the relationships between these issues, creating the groundwork for risk control measures to be developed. Discriminatory factors can increase the likelihood of traffic accidents or other factors that contribute to the severity of injuries sustained as a result of traffic accidents. We've compiled a list of 24 features that are thought to be linked to road accidents. The information was obtained from the Accident Research Center (ARI) at BUET. To recognize the clusters, we utilized hierarchical clustering on the dataset. Each cluster represents a perceived subtype found in the accident dataset. To categorize those experimental subtypes, we applied six different classification methods on the datasets. Finally, for the interpretation of features for discriminatory variables, we used the LIME analysis technique. As a result, future work will necessitate a thorough examination of the updated dataset of traffic accidents from across the country, as well as the application of more classification and clustering algorithms, as well as the improvement of the discriminatory factors identification model through additional development and experiments to conduct follow-up traffic accidents.

ACKNOWLEDGMENT

We are grateful to the Accident Research Center (ARI), BUET (<http://ari.buet.ac.bd/>) for allowing us to collect information and data from its Modular Accident Analysis Program, Version-5 database software.

REFERENCES

- [1] N. F. Antara, "Dhaka Tribune," E-News Paper, 9 1 2021. [Online]. Available: <https://www.dhakatribune.com/bangladesh/2021/01/09/report-18-people-killed-every-day-on-average-in-road-accidents-in-2020>. [Accessed 20 10 2021].
- [2] Bobermin, Mariane, and Sara Ferreira. "A novel approach to set driving simulator experiments based on traffic crash data." *Accident Analysis & Prevention* 150 (2021): 105938.
- [3] Amiri, Amir Mohammadian, et al. "GIS-based crash hotspot identification: a comparison among mapping clusters and spatial analysis techniques." *International journal of injury control and safety promotion* (2021): 1-14.
- [4] Li, Yu, et al. "Crash report data analysis for creating scenario-wise, spatio-temporal attention guidance to support computer vision-based perception of fatal crash risks." *Accident Analysis & Prevention* 151 (2021): 105962.
- [5] Roy, Dooti, Ved Deshpande, and M. Henry Linder. "A cluster-based taxonomy of bus crashes in the United States." *Computational Statistics* (2021): 1-18.
- [6] Suarez-del Fuego, Rocio, et al. "Cluster analysis of seriously injured occupants in motor vehicle crashes." *Accident Analysis & Prevention* 151 (2021): 105787.
- [7] Tahfim, Syed As-Sadeq, and Chen Yan. "Analysis of Severe Injuries in Crashes Involving Large Trucks Using K-Prototypes Clustering-Based GBDT Model." *Safety* 7.2 (2021): 32.
- [8] Francis, Filbert, et al. "Investigation of road infrastructure and traffic density attributes at high-risk locations for motorcycle-related injuries using multiple correspondence and cluster analysis in urban Tanzania." *International journal of injury control and safety promotion* (2021): 1-11.
- [9] Ersen, Mert, Ali Hakan Büyüklü, and Semra Erpolat Taşabat. "Analysis of Fatal and Injury Traffic Accidents in Istanbul Sariyer District with Spatial Statistics Methods." *Sustainability* 13.19 (2021): 11039.
- [10] Hosseinian, Seyed Mohsen, and Vahid Najafi Moghaddam Gilani. "Analysis of factors affecting urban road accidents in rasht metropolis." *Eng Transactions* 1 (2020): 1-4.
- [11] Cai, Qiuru. "Cause Analysis of Traffic Accidents on Urban Roads Based on an Improved Association Rule Mining Algorithm." *IEEE Access* 8 (2020): 75607-75615.
- [12] Lin, Yunduan, and Ruimin Li. "Real-time traffic accidents post-impact prediction: Based on crowdsourcing data." *Accident Analysis & Prevention* 145 (2020): 105696.
- [13] AlKheder, Sharaf, Fahad AlRukaibi, and Ahmad Aiash. "Risk analysis of traffic accidents' severities: An application of three data mining models." *ISA transactions* 106 (2020): 213-220.
- [14] Yong-Zheng, Yang, and Y. Mei. "Analysis of Influencing Factors of Traffic Accidents in Undersea Tunnel." *Journal of Engineering Research and Reports* (2020): 37-49.
- [15] Čubranić-Dobrodolac, Marjana, et al. "Modelling driver propensity for traffic accidents: a comparison of multiple regression analysis and fuzzy approach." *International journal of injury control and safety promotion* 27.2 (2020): 156-167.
- [16] Casado-Sanz, Natalia, Begoña Guirao, and Maria Attard. "Analysis of the risk factors affecting the severity of traffic accidents on Spanish crosstown roads: the driver's perspective." *Sustainability* 12.6 (2020): 2237.
- [17] Iqbal, Asad, et al. "Road Traffic Accident Analysis and Identification of Black Spot Locations on Highway." *Civil Engineering Journal* 6.12 (2020): 2448-2456.
- [18] Song, Minglei, Rongrong Li, and Binghua Wu. "A novel prediction model of traffic accidents based on big data." *International Journal of Modeling, Simulation, and Scientific Computing* 10.04 (2019): 1950022.
- [19] Zhang, Cheng, Yue Shu, and Lixin Yan. "A Novel Identification Model for Road Traffic Accident Black Spots: A Case Study in Ningbo, China." *IEEE Access* 7 (2019): 140197-140205.
- [20] Hussain, Sadiq, et al. "Performance evaluation of various data mining algorithms on road traffic accident dataset." *Information and Communication Technology for Intelligent Systems*. Springer, Singapore, 2019. 67-78.
- [21] Pineda-Jaramillo, Juan, and Óscar Arbeláez-Arenas. "Modelling road traffic collisions using clustered zones based on Foursquare data in Medellín." *Case Studies on Transport Policy* 9.2 (2021): 958-964.