Road Accident Severity Analysis and Prediction Using Different Machine Learning Algorithms

Ajoy Sarker[1], Shreya Nag Riya[1], Umme Faria Moon[1], MD. Ahsan Habib Rasel[1], MD. Minhaj Uddin[1], and Dr. Md Musfique Anwar[1]

Department of CSE, Jahangirnagar University, Dhaka, Bangladesh
{ajoy.stu20181, nag.stu2018, moon.stu20181 , ahsanhabib.stu2018, minhaj.stu2018, manwar}@juniv.edu

Abstract. Road accidents in Bangladesh remain an increasingly alarming issue. In this research, our goal is to predict the occurrence and severity of road accidents across all 64 districts of Bangladesh using machine learning techniques, based on data collected from police records in collaboration with the Accident Research Institute (ARI) at Bangladesh University of Engineering and Technology (BUET), covering the years 2001 to 2020. The study addresses both classification and regression tasks: prediction of accident occurrence using classifiers such as Random Forest, XGBoost, and CatBoost, and prediction of severity using regression models including Random Forest Regression, Gradient Boosting, XGBoost, CatBoost, and Artificial Neural Networks (ANN). CatBoost achieved the highest F1 score of 0.6476 among the classifiers, making it the most effective model to handle class imbalance. Gradient Boosting demonstrated the best performance for regression with an R^2 score of 0.0802, which improved to 0.0818 after hyperparameter tuning. The results suggest that machine learning offers a promising approach to understanding and predicting road accident patterns in Bangladesh, providing valuable insights for transportation safety authorities and policymakers.

Keywords: Road Accidents · Severity · Machine Learning · Prediction

1 Introduction

Roadway traffic accidents are an acute problem across the world. According to the report of the World Health Organization (WHO), about 1.9 million people die in road accidents, and other sources calculate that about 20 to 50 million people are injured every year. Here, the injuries caused by car accidents create an irresistible percentage of deaths among children and teenagers between the ages of 15-29 [10].

Unfortunately, we realize that there is an unusually high rate of road accidents in Bangladesh. Road accidents are one of the major concerns in Bangladesh. These incidents affect individuals, families, communities, and countries. The latest observation directed by the Accident Research Institute (ARI) of Bangladesh

University of Engineering and Technology (BUET) found that almost 12,000 people die due to accidents, and around 35,000 people are injured every year [3]. The record shows that on average, every day at least 32 people die in road accidents in Bangladesh [6]. If the roadway traffic difficulties are not controlled immediately, day by day the situation will be more terrible [4]. Though this type of circumstance is never guaranteed, and there are no boundaries of accidents, because accidents can occur with anyone at any time and in any situation. Consequently, the method of understanding and learning about the potential factors that are responsible for road accidents will be more effective if we can determine the geographical areas where the accidents have occurred repeatedly, rather than in other places.

In this research, our motive is to predict the number of road accidents by analyzing the previous report, which can help us identify the places for road accidents as risky areas and thereby make them safer. Here, we have applied different machine learning models in which accident data records are used to understand the properties of all sorts of features. This contributes to determining the safety measures by which risky aspects of accidents can be minimized.

- This research uses Bangladesh's largest road accident dataset (2001–2020) for long-term pattern analysis.
- We apply and compare multiple ML models for classifying and predicting accident occurrence and severity across all 64 districts—a first in this context.

2 Literature Review

This section reviews some of the previous works relevant to this research.

J. Paul et al. proposed a multiclass machine learning model to predict both the occurrence and severity of road accidents in Bangladesh [11]. They analyzed sixty features related to five types of casualties and tested several algorithms, including Decision Tree, Random Forest, Multilayer Perceptron, and Categorical Naive Bayes. The Decision Tree algorithm performed the best, achieving 99.77% accuracy for accident prediction and 99.80% for severity, with F1 scores of 98.68% and 99.80%, respectively. Their work demonstrated that incorporating multiple factors significantly enhances prediction accuracy.

In [12], a voting-based ensemble machine learning model was developed to predict the severity of road accidents, addressing the gap in ensemble methods in previous studies. They combined four algorithms—Random Forest, XGBoost, KNN, and LightGBM—and used SelectKBest and ExtraTreeClassifier for feature selection. The model was tested on accident datasets from both Bangladesh (2017–2020) and the USA (2016–2020), achieving 96% accuracy for binary classification and 71% for multiclass classification in Bangladesh, and 92.1% and 87%, respectively, in the USA. The ensemble model consistently outperformed individual classifiers. Similarly, the study in [16] developed models using neural networks, decision trees, and a hybrid model to predict injury severity in

road traffic accidents. The hybrid model outperformed the individual models, emphasizing the value of data-driven approaches for safety, even with limited resources.

A machine learning approach was applied in [13] to predict fatal and non-fatal road accident outcomes from crash data collected in Dhaka city (2017–2022). Several classification algorithms, including Logistic Regression, SVM, Naive Bayes, Random Forest, Decision Tree, Gradient Boosting, LightGBM, and Artificial Neural Network, were tested. Among them, LightGBM achieved the best ROC-AUC score of 0.72. The study used SHAP (Shapley Additive Explanations) to identify the most influential factors contributing to accident fatality, such as casualty class, accident time, location, vehicle type, and road type. These findings can help improve road safety, especially in developing countries.

Labib et al. analyzed road accident severity in Bangladesh using a dataset of about 44,582 traffic accidents from 2001 to 2020, collected by BUET's Accident Research Institute (ARI) [14]. Several classifiers, including Logistic Regression, Decision Tree, Gaussian Naïve Bayes, K-Nearest Neighbors, and AdaBoost, were tested, with AdaBoost yielding the best results. The study aimed to classify accidents into four categories—Fatal, Grievous, Simple Injury, and Motor Collision—and identified key factors affecting severity. It emphasized the importance of predictive modeling in tackling the growing issue of traffic accidents in developing countries.

Another research [15] investigated traffic accident data from Bangladesh's N5 National Highway, a high-risk road. Using 892 records from the Accident Research Institute (ARI), 12 decision tree algorithms were employed to classify accident patterns. The study identified the best-performing classifier and extracted decision rules to prevent future accidents, offering valuable insights for improving safety on this critical highway.

Data mining algorithms predicted high-risk accident locations on the Dhaka-Aricha highway in [17]. Traffic accident data were collected from highway police stations, preprocessed, and analyzed using five classification algorithms: Rotation Forest, NBTree, JRip, Naive Bayes, and Ridor. The accuracies of the models were compared, and the best-performing model was identified. The results can be used to target high-risk locations on the highway to reduce accidents.

3 Methodology

We employed a supervised machine learning approach in this research to predict the occurrence and severity of road accidents. Supervised learning occurs when algorithms learn from labeled data so they can make predictions on new, unseen data after being trained using input-output pairs. By the combination of numerous varied features such as accident location, severity, vehicle type, and road condition, our models attempt to establish patterns influencing both the occurrence and severity of accidents. This enables us to identify statistically

significant factors that can be utilized to assist proactive identification of risky situations [2], thereby contributing towards improving road safety.

The entire ML-based pipeline, from data collection and preprocessing to model training, evaluation, and interpretation for classification and regression tasks, is illustrated in Fig. 1.

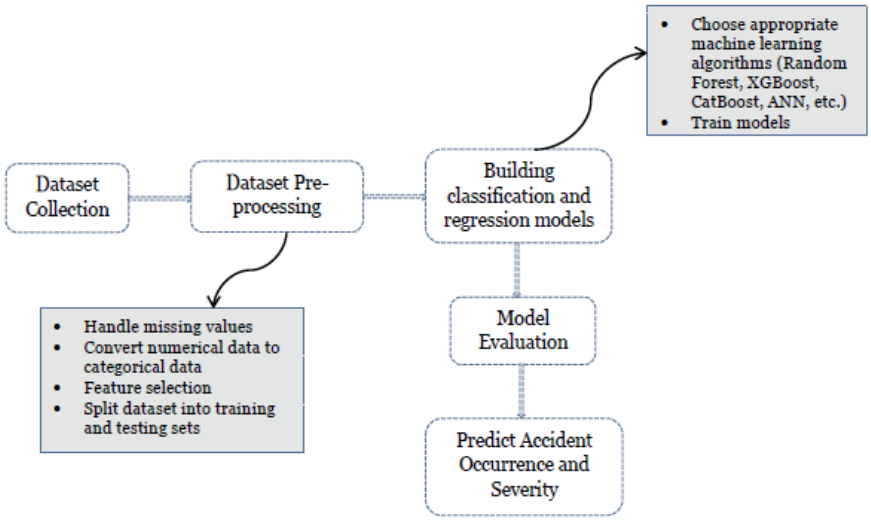

Fig. 1. Proposed machine learning model workflow for road accident prediction

3.1 Dataset Collection

The dataset used for our proposed model is collected from police records in collaboration with the Accident Research Institute (ARI) at Bangladesh University of Engineering and Technology (BUET). We also collected a dataset from the Bangladesh Road Transport Corporation (BRTC). After collecting from these sources, we merged the dataset with all the common columns containing common values. It contains information on all traffic accidents reported between 2001 and 2020.

3.2 Dataset Preprocessing

It is one of the most important parts, which is mainly a technique of data mining. In this step, the following operations were performed:

– **Handling missing values:** Initially, our dataset contained numerous missing values across both numerical and categorical variables. To address this,

missing categorical values were replaced with a special label named 'unknown', while the missing numerical values were imputed with the rounded mean value specific to each feature.
- **Mapping:** Our dataset initially included a wide range of numerical values that were taken from the ARI report. We converted these numerical values to the respective categorical variables to improve interpretability and simplify analysis.
- **Data cleaning:** It is the process of removing the noisy and inconsistent data that is not wanted and they are irrelevant for our analysis. These data do not help analyze because they may provide inaccurate results.
- **Feature selection:** Working with a sizable number of features may degrade the performance of the model as the training time increases exponentially with the number of features. Feature selection is mainly used to reduce the dimensionality and prevent overfitting.

3.3 Modeling Techniques

To predict the occurrence and severity of road accidents, we have used a set of supervised machine learning algorithms, particularly developed to operate on two tasks: regression and classification. For the task of predicting the occurrence of an accident (classification task), we used the Random Forest classifier, XGBoost classifier, and CatBoost classifier. For making predictions on the severity of accidents (regression task), we employed Random Forest Regression, Gradient Boosting, Artificial Neural Network (ANN), XGBoost, and CatBoost. We even trained an Ensemble Model, which combines the predictions of multiple regression models to provide greater overall accuracy and robustness. The applied models are described below-

- **Random Forest**: An ensemble method that constructs multiple decision trees and merges their predictions through majority voting for classification and averaging for regression to improve accuracy and avoid overfitting [24].
- **XGBoost**: A very effective gradient boosting framework that incrementally adds models in a series to rectify the mistakes of the previous models through gradient descent. It includes regularization and handles missing values efficiently, making it suitable for classification and regression [18].
- **CatBoost**: A gradient boosting algorithm that is specifically optimized for categorical features. It uses techniques like ordered boosting to prevent overfitting and can be applied to classification as well as regression problems [19].
- **Gradient Boosting**: A boosting technique that creates weak learners in a series, with each new learner minimizing the loss of the previous learners. It is extensively used for regression problems due to its strong predictive power [20].
- **Artificial Neural Network (ANN)**: Multi-layer computational model inspired by the human brain that can learn complex, non-linear relationships in data. ANN is particularly suited for regression tasks with subtle patterns and interactions [21].

– **Ensemble Model**: Combines predictions from multiple base models to enhance generalization and robustness. In our case, ensemble learning was applied to regression by aggregating predictions from individual top-ranked models [25].

3.4 Model Evaluation

To assess the effectiveness of our models, we used different evaluation metrics for regression and classification tasks. These metrics help determine the accuracy of predictions and how well the models generalize to unseen data.

– **For Regression Models:**
  - **Mean Absolute Error (MAE)**: MAE measures the average absolute difference between actual and predicted values.
  - **Mean Squared Error (MSE)**: MSE calculates the average of the squared differences between actual and predicted values, penalizing larger errors more heavily.
  - **R-squared ($\mathbf{R}^2$)**: R^2 shows how well the model explains the data.
  - **Cross-Validation $\mathbf{R}^2$ (CV $\mathbf{R}^2$)**: CV R^2 averages R^2 scores across K-fold cross-validation to assess the model's generalization ability.

The formula for calculating the Mean Absolute Error, Mean Squared Error, R-squared and Cross-Validation R Squared Value are given below:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

$$\text{CV } R^2 = \frac{1}{K}\sum_{k=1}^{K}R_k^2 \tag{4}$$

Where:

– y_i : Actual value
– $\hat{y}_i$: Predicted value
– $\bar{y}$: Mean of actual values
– n : Number of observations
– K : Number of cross-validation folds

Smaller values of MAE and MSE indicate higher accuracy, while R^2 values closer to 1 suggest a better model fit [22].

- **For Classification Models:**
  - **Accuracy**: Accuracy is the percentage of correct predictions.
  - **F1 Score**: F1 Score is the harmonic mean of Precision and Recall. It is particularly useful for imbalanced datasets.

Higher accuracy and F1 score values indicate better classification performance [23].

4 Data Set

4.1 Dataset Description

In this section, we provide a detailed description of the dataset utilized in this study. The dataset was sourced from police records and collected in collaboration with the Bangladesh University of Engineering and Technology (BUET). It serves as a comprehensive repository of incidents documented by law enforcement authorities, offering valuable insights into various aspects of public safety and law enforcement efforts within the studied region.

4.2 Data Collection Process

The dataset was gathered rigorously, following ethical principles and data protection regulations. Cooperation with law enforcement agencies led to the systematical collection and use of official police records of a specified time-frame. This approach came into existence to give shape to a structured dataset that could be better used in the analysis and modeling.

4.3 Data Characteristics

Our dataset has 44583 records with a number of 26 features, among which 4 are numerical, and 22 are categorical.

- **Thana**: The administrative subdivision within a district.
- **District**: The geographical area or division within a region.
- **Month**: The month in which the accident occurred.
- **Year**: The year in which the accident occurred.
- **Time**: The time of the day when the accident took place.
- **Accident_Severity**: The severity level of the accident, such as 'Fatal', 'Grievous', 'Motor-Collision', or 'Simple'.

- **Junction_Type**: The type of junction where the accident occurred, like 'T-Junction', 'CrossRoad',' No Junction',' Railway ', etc.
- **Traffic_Control**: The type of traffic control present at the accident location, such as 'Centreline',' No Control',' Police Control', and so on.
- **Movement**: The movement direction of the vehicle involved, like '1-way Street' or '2-way Street'.
- **Divider**: Presence of a divider on the road at the accident location(yes or no).
- **Weather**: The weather conditions at the time of the accident, such as 'Fair', 'Fog', 'Rain', and so on.
- **Light**: Lighting conditions, such as 'Dawn', 'Daylight' or 'Night'.
- **Road_Geometry**: The geometry of the road, like 'Straight+Flat', 'Slop Only' etc.
- **Surface_Condition**: Condition of the road surface, such as 'Dry', 'Wet'.
- **Surface_Type**: Type of road surface, such as 'Sealed' or 'Brick'.
- **Surface_Quality**: Quality of the road surface, like 'Good' or 'Rough'.
- **Road_Class**: Types of the road, like 'Regional', 'Feeder', 'National', and so on.
- **Road_Feature**: Features of the road.
- **Location_Type**: Type of location where the accident occurred,d such as 'Rural' or 'Urban'.
- **Vehicle_Type**: Type of vehicle involved in the accident, such as 'Truck', 'Heavy truck', 'Bus', 'Car', 'Pick up', 'Oil tanker', etc.
- **Vehicle_Movement**: The movement of the vehicle involved in the accident, such as 'Going Ahead', 'Parked,' and so on.
- **Vehicle_Loading**: whether the vehicle was loaded within legal limits or not such as 'legal', 'Unsafe', etc.
- **Vehicle_Defect**: Any defects present in the vehicle at the time of the accident.
- **Vehicle_Driver_Age**: The age of the driver.
- **Vehicle_Alcohol**: Indicates whether alcohol consumption was suspected in the accident or not as 'Suspected' or 'Not suspected'.
- **Vehicle_Seat_Belt**: whether seat belts were in use during the accident, like 'Yes' or 'No'.

Every feature offers meaningful information regarding the situations and qualities associated with the ones that were recorded, thus enabling an extensive evaluation and complex modeling.

Fig. 2 illustrates the accident severity levels in the dataset, indicating that fatal accidents are the most frequent while vehicular collisions are the least. This distribution highlights the critical need to focus on reducing the fatal outcomes of road accidents.

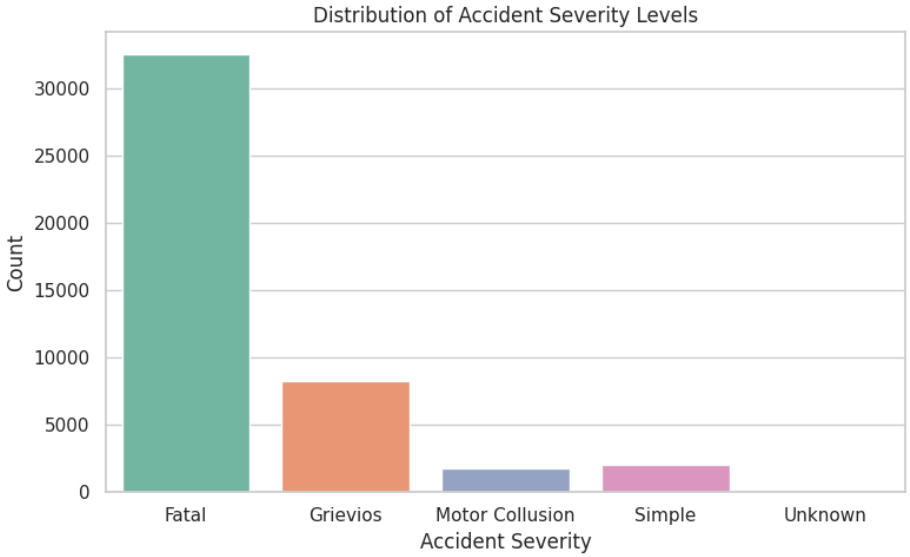

Fig. 2. Distribution of accident severity levels in the dataset

Fig. 3 shows the classification of accident-prone junction types. The highest number of accidents was recorded in the No Junction type area, and the lowest number of accidents was recorded in the staggered junction, while Y-Junctions and railway show no accident records in the dataset.

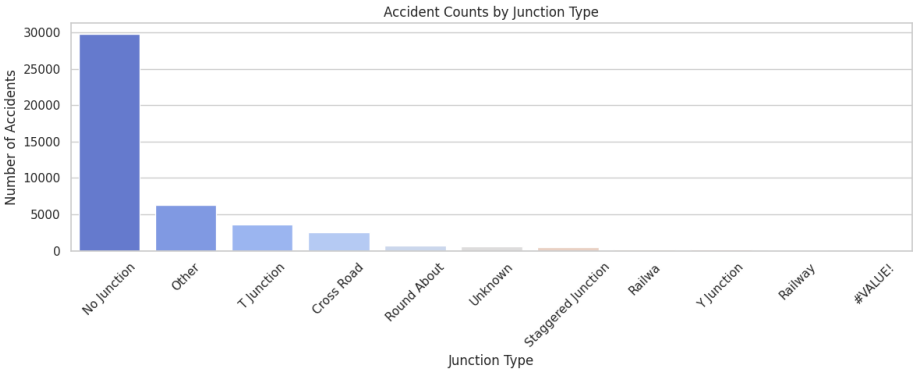

Fig. 3. Classification of accident-prone junction types

5 Results and Analysis

Our dataset has 44,582 records with a number of 26 features, among which 4 are numerical, and 22 are categorical. The features include various aspects such as location, time, vehicle details, road conditions, and accident severity.

5.1 Model Performance

The performance of the regression models used to predict accident severity is presented in Table 1, while Table 2 compares the classification models used for predicting accident occurrence. Among the regression models, Gradient Boosting achieved the highest R^2 score of 0.0802, which improved to 0.0818 after hyperparameter tuning. CatBoost and the ensemble model also performed competitively. In contrast, the Artificial Neural Network (ANN) underperformed, likely due to the dataset's size and structure. For classification, CatBoost achieved the highest F1 score of 0.6476 and handled class imbalance effectively. While Random Forest and XGBoost showed strong accuracy, their F1 scores were slightly lower than that of CatBoost.

Table 1. Regression Model Performance Comparison

Model	**MAE**	**MSE**	**R^2**	**CV R^2**
Random Forest	0.5558	0.5903	0.0298	0.0288
Gradient Boosting	0.5402	0.5596	0.0802	0.0855
Artificial Neural Network	**0.5711**	**0.6428**	-0.0565	-0.0893
XGBoost	0.5379	0.5829	0.0419	0.0443
CatBoost	0.5337	0.5642	0.0726	0.0794
Ensemble	0.5395	0.5630	0.0746	0.0789
Tuned Gradient Boosting	0.5370	0.5586	**0.0818**	**0.0871**

Table 2. Classification Model Performance Comparison

Model	**Accuracy**	**F1 Score**
Random Forest	72.56%	0.6448
XGBoost	72.28%	0.6391
CatBoost	72.10%	**0.6476**

5.2 Key Insights from Data Analysis

- **Accident Severity Distribution:**
  - Fatal accidents are the most common (about 32,000 cases)
  - Followed by Grievous injuries (around 8,000 cases)
  - Motor collisions and simple accidents are less frequent
- **Weather and Lighting Conditions:**
  - Most accidents occur in fair weather conditions
  - Daylight hours see the highest number of accidents
  - Fatal accidents are particularly high during daylight in fair weather
- **Vehicle Types and Severity:**
  - Heavy trucks are involved in the most fatal accidents
  - Followed by buses and pickup trucks
  - Smaller vehicles like bicycles and rickshaws show lower fatality rates
- **Junction Types:**

- Most accidents (around 30,000) occur at locations with no junction
- T-junctions and crossroads show significant accident numbers
- Roundabouts have relatively fewer accidents

– **Road Features and Surface Quality:**

- Regular roads with good surface quality have the highest number of accidents (38,750 cases)
- Bridges and narrowing roads show moderate accident numbers
- Roads under repair show fewer accidents, possibly due to increased caution

– **Driver Age Distribution:**

- Most accidents involve drivers between the ages 25-45
- Fatal accidents show a wider age range
- Simple accidents tend to involve slightly older drivers

Fig. 4 highlights the effect of combinations of traffic control types and road geometry on accident rates.

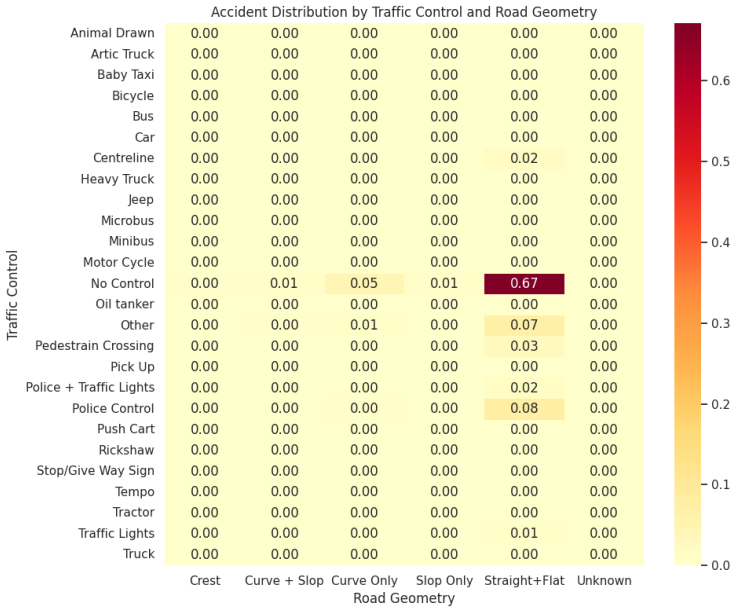

Fig. 4. Heatmap showing accident distribution by traffic control and road geometry

Fig. 5 highlights the relationship between vehicle types and accident severity. It clearly shows that heavy trucks and buses are most frequently involved in fatal accidents, while smaller vehicles like motorcycles and rickshaws tend to be associated with less severe outcomes.

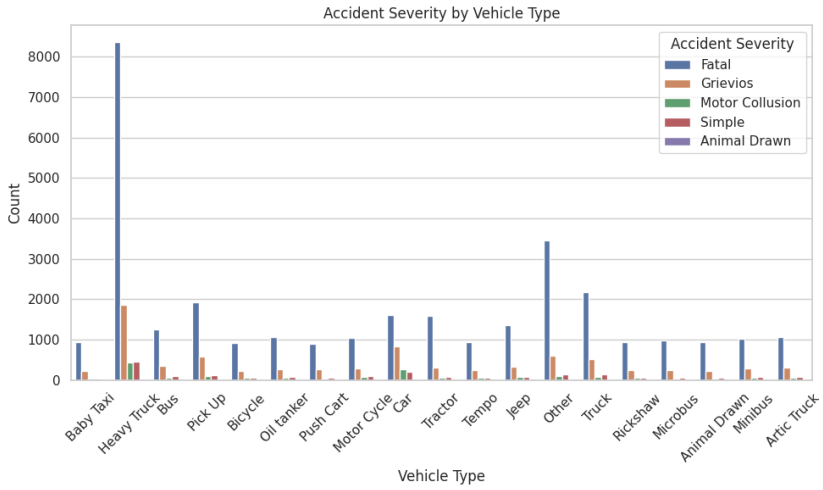

Fig. 5. Accident severity distribution by vehicle type

Fig. 6 shows the variation of accident severity by driver age groups. It indicates that the majority of fatal accidents occur among drivers aged 25-45 years, implying that this age group is perhaps exposed to more danger on the road.

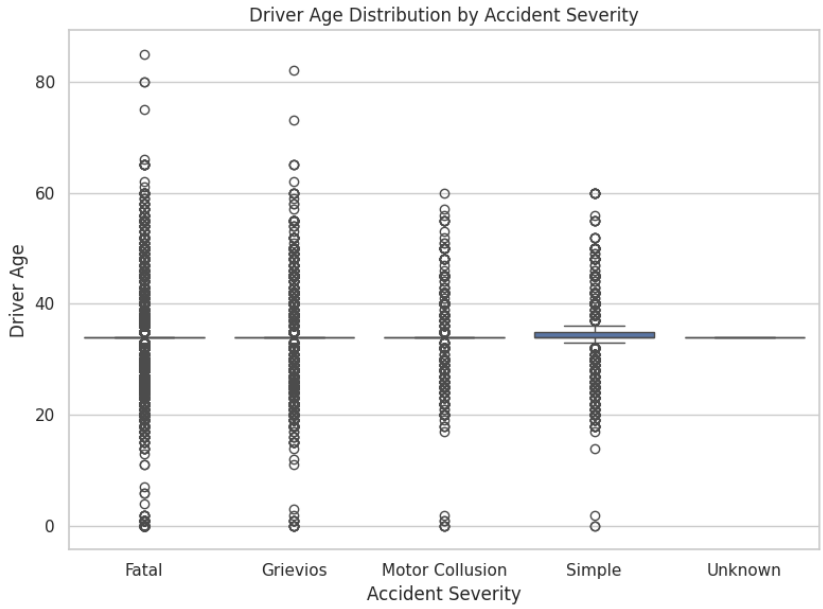

Fig. 6. Driver age distribution across different accident severity levels

Fig. 7 compares the number of accidents on the road based on different conditions and characteristics of the road surface. Surprisingly, roads with good

surface quality and regular features tend to have the most accidents, possibly because they carry more traffic or encourage faster driving. On the other hand, roads that are under repair or have narrowing features see fewer accidents, which may be due to increased driver alertness and caution in such conditions.

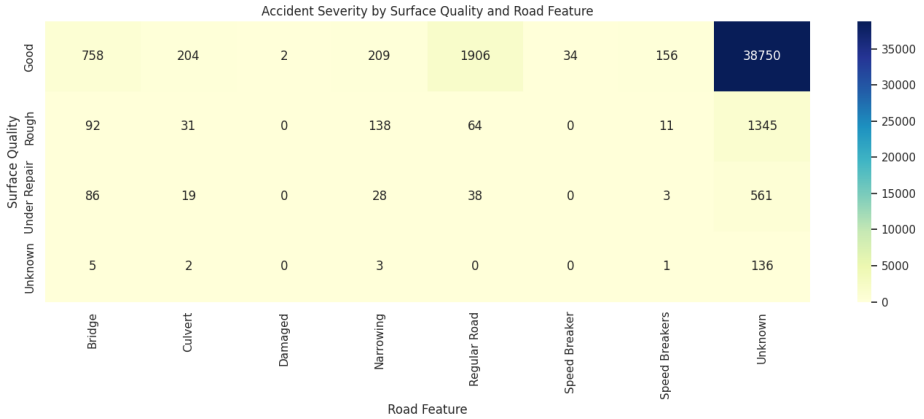

Fig. 7. Accident counts by road surface quality and features

6 Conclusion

Here we are analyzing historical road accident data. We trained our machine learning model with this data to predict the number of casualties in all the districts of Bangladesh. For predicting casualties, we use the Random Forest regression algorithm. Random forest regression algorithm is a supervised machine learning algorithm, which depends on the output of multiple decision trees for obtaining a single conclusion. We also computed the R-squared value and the RMSE; these tell us that the Random Forest algorithm here has done a good job. We realized that the regression model performs well enough on our dataset [5]. In the future, we will concentrate on the number of casualties in each city and also try to find the effect of road accidents on people's lives.

Acknowledgments. We would like to thank the Accident Research Institute (ARI) of Bangladesh University of Engineering and Technology (BUET), and Bangladesh Road Transport Corporation (BRTC) for providing the dataset and support for this research.

Disclosure of Interests. The authors declare that they have no competing interests.

References

1. GeeksforGeeks: Gradient Boosting in ML. https://www.geeksforgeeks.org/ml-gradient-boosting/. [Accessed 04-05-2024]

2. Archana Jalke, Kirti Suryavanshi, Akanksha Jadhav, Shruti Jadhav: Road accident analysis and prediction of accident severity using machine learning. https://www.irjet.net/archives/V7/i12/IRJET-V7I12129.pdf
3. Ashok S. Patil, A. N. Bindu, Y. R. Nikitha: Road Accident Analysis and Prediction. https://www.ijresm.com/Vol.32020/Vol3Iss2February20/IJRESMV3I255.pdf. [Accessed 05-2024]
4. Md Sifat Bin Siraj, Fazle Rabbi, Mossa. Soniya Asma, Md. Jubayadul Islam: Road accident analysis: A case study dhaka metropolitan area. 6:7, 11 2021
5. Al Amin Biswas, Md. Jueal Mia, Anup Majumder: Forecasting the number of road accidents and casualties using random forest regression in the context of bangladesh. In: 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pages 1-5, 2019
6. Md. Farhan Labib, Ahmed Sady Rifat, Md. Mosabbir Hossain, Amit Kumar Das, Faria Nawrine: Road accident analysis and prediction of accident severity by using machine learning in bangladesh. In: 2019 7th International Conference on Smart Computing Communications (ICSCC), pages 1-5, 2019
7. Srivignesh R: A Walk-through of Regression Analysis Using Artificial Neural Networks in Tensorflow. https://www.analyticsvidhya.com/blog/2021/08/a-walk-through-of-regression-analysis-using-artificial-neural-networks-in-tensorflow/. [Accessed 04-05-2024]
8. Setty Rao: Ritchie's predictive model of accidents using a machine learning algorithm. 09 2020
9. Shweta, J Yadav, K Batra, A K Goel: A framework for analyzing road accidents using machine learning paradigms. Journal of Physics: Conference Series, 1950(1):012072, aug 2021
10. World Health Organization: Road traffic injuries. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries. [Accessed: April 20, 2024]
11. J. Paul, Z. Jahan, K. Lateef, M. Islam, and S. Bakchy, Prediction of road accident and severity of Bangladesh applying machine learning techniques, in Proc. IEEE Region 10 Humanitarian Technology Conference (R10-HTC), 2020, pp. 1–6, doi: 10.1109/R10-HTC49770.2020.9356987.
12. K. Fahad, M. Joarder, M. Nahid, and T. Tasnim, Road accidents severity prediction using a voting-based ensemble ML model, 2024
13. M. A. K. Rifat, A. Kabir, and A. S. Huq, An explainable machine learning approach to traffic accident fatality prediction, arXiv preprint arXiv:2409.11929, 2024. [Online]. Available: https://arxiv.org/abs/2409.11929
14. F. Labib, A. Rifat, M. Hossain, A. Das, and F. Nawrine, Road accident analysis and prediction of accident severity by using machine learning in Bangladesh, in Proc. IEEE Int. Conf. on Systems, Computation, and Control (ICSCC), 2019, pp. 1–5, doi: 10.1109/ICSCC.2019.8843640.
15. M. Satu, S. Ahamed, F. Hossain, T. Akter, and D. Farid, Mining traffic accident data of N5 National Highway in Bangladesh employing decision trees, in Proc. IEEE Region 10 Humanitarian Technology Conference (R10-HTC), 2017, doi: 10.1109/R10-HTC.2017.8289059.
16. M. Chong, A. Abraham, and M. Paprzycki, Traffic accident analysis using decision trees and neural networks, arXiv preprint arXiv: cs/0405050, 2004, doi: 10.48550/arXiv.cs/0405050.
17. M. Satu, T. Akter, M. Arifen, and M. Mia, Predicting Accidental Locations of Dhaka-Aricha Highway in Bangladesh Using Different Data Mining Techniques, Int. J. Comput. Appl, vol. 165, 2017, doi: 10.5120/ijca2017914096.

18. T. Chen and C. Guestrin, XGBoost: A scalable tree boosting system, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
19. L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, CatBoost: Unbiased boosting with categorical features, in Advances in Neural Information Processing Systems (NeurIPS), vol. 31, pp. 6638–6648, 2018.
20. J. H. Friedman, Greedy function approximation: A gradient boosting machine, Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001. doi: 10.1214/aos/1013203451.
21. S. Haykin, Neural Networks: A Comprehensive Foundation, 2nd ed. Upper Saddle River, NJ: Prentice Hall, 1998.
22. C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
23. D. Chicco and G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genomics, vol. 21, no. 1, p. 6, 2020.
24. L. Breiman, Random forests, Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, doi:10.1023/A:1010933404324.
25. Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms, Boca Raton, FL, USA: CRC Press, 2012.