Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 17, No. 3, 2026

Traffic Accident Severity Prediction Based on Multi-Model Comparison

Ziyu Cao *

School of Computing and Data Science, Xiamen University Malaysia, Sepang, Selangor 43900, Malaysia

*** Corresponding author Email:** EEE2209276@xmu.edu.my

Abstract: Traffic accident severity prediction is important for improving traffic safety management and supporting risk assessment and prevention. However, current studies have several limitations with respect to the systematic approach for comparing models and the comprehensiveness of evaluation metrics. This paper aims to systematically assess the overall performance of various machine learning models on a unified experimental framework to predict the severity of accidents in a binary classification task. This study randomly sampled 500,000 records from the US Accidents dataset and used them as the experimental sample. Following data cleaning, missing value handling, feature engineering, and categorical variable encoding, a comparison of four representative models was performed: Logistic Regression (LR), Random Forest (RF), XGBoost and Multi-Layer Perceptron (MLP). With the imbalanced nature of the data, this paper evaluates model performance using accuracy, AUC, ROC curves, and confusion matrices to assess the overall classification performance of different models and their ability to identify the severe accident class. The results show that the tree-based ensemble models Random Forest and XGBoost outperform Logistic Regression and MLP in terms of overall predictive performance, with XGBoost exhibiting better overall performance. Variables that contribute significantly to the model's predictions include traffic control facilities, time factors, spatial location, and accident-affected distance. The results indicate that machine learning techniques could be useful in traffic accident severity prediction and could be a reference for traffic safety risk assessment. Meanwhile, class imbalance, feature interpretability and model generalizability are challenges that need further investigation and resolution.

Keywords: Traffic Accident; Severity Prediction; Machine Learning; Model Comparison; Feature Importance.

1. Introduction

Road traffic accidents are one of the serious public safety problems in the world which cause casualties and economic losses. The World Health Organization estimates that around 1.2 million people per year die in road traffic accidents, and that occurrence of accidents is closely associated with a range of factors, such as the driving environment, weather, and traffic control facilities [1,2]. Hence, enhancing traffic safety management and risk prevention is an urgent issue. The traditional accident prediction techniques mainly rely on statistical models, such as regression analysis or Poisson models. These models have advantages in terms of interpretability but have significant limitations in the case of nonlinear relationships among variables in complex data [3]. Machine learning models, in contrast, can handle multidimensional data and can be applied to traffic data with high complexity and heterogeneity [4]. Furthermore, they can learn the underlying patterns within the data and this will help them to improve the predictions about the severity of traffic accidents [3]. At present, machine learning methods have gradually been applied to traffic accident prediction, however existing studies have several limitations, such as limited model types, a lack of systematic comparative analysis, limited data scale, and insufficient feature dimensions [4]. Based on this, logistic regression, random forest, XGBoost and multi-layer perceptrons (MLP) will be employed in this study to systematically compare the predictive performance of different models, while using road traffic accident severity prediction as the main task. Through the comparison of model performance, this study aims to identify more suitable algorithms and provide a reference for future traffic safety management and risk prevention.

2. Literature Review

Road traffic accident severity is influenced by multiple factors. The physical characteristics of intersections (Junction) are strongly correlated with the severity of the injuries [5]. The presence of traffic signals affects collision risk by influencing vehicle–pedestrian interactions, and the lack of crosswalks may increase the incidence of severe accidents [6]. Railway level crossings are generally associated with a higher risk of casualties [7]. Meanwhile, points of interest (POI) like stations and amenities are more appropriate as proxy variables that represent the surrounding land use rather than direct collision-risk factors [8].

Other important factors are time and weather. Time of day (Hour) is related to traffic flow characteristics and may be a traffic safety influencing factor [1]. At night when the illumination is low, visibility is limited and reaction times are extended, which can lead to severe injuries [9]. Day-night prediction models have suggested that lighting affects the mechanisms related to accident severity: In a mountainous area, the same weather conditions can result in opposite effects on the severity of the accidents for both day and night [10,11]. Moreover, the year (Year) is more related to the long-term macro-level safety trends, and the external environmental factors may further increase the accident risk, such as severe weather [12,13].

In terms of spatial factors, geographic coordinates are closely related to risk clustering. For example, geographic information systems (GIS) and spatiotemporal analysis techniques have been used in some studies to identify high-risk accident hotspots in major cities in Ohio and California [14,15]. Time zone (Timezone) can be used as a proxy variable to indicate regional differences in climate and traffic culture [7]. In addition, the accident-affected road distance (Distance) is another important predictive variable which should not be neglected when modeling [3].

In order to address feature complexity, the traditional models like binary logistic regression (LR) are found to have good applicability in analyzing factors associated with accident severity [16]. However machine learning can be more adaptable in the case of high dimensional and imbalanced data, for example, the SMOTE method can be used to solve the problem of class imbalance [17]. Ensemble models like the Random Forest (RF) and XGBoost are well known for their strong performance in capturing complex features and improving classification performance[18]. Meanwhile, Multi-Layer Perceptrons (MLP) have been proven to be successful in learning the main factors that result in severe casualties [19]. For algorithms like SVM, NB, and RNN, they are being mostly studied as supplementary methods for specific tasks [20]. Thus, LR, RF, XGBoost and MLP can be considered as representative models from three different categories: traditional statistical, ensemble learning, and neural network approaches.

3. Research Methodology

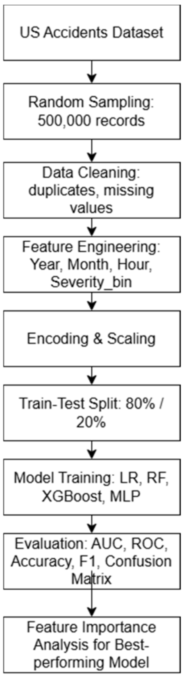

Figure 1. Research workflow

However, there are a number of research gaps in this area. Firstly, there is no standardization in existing datasets and evaluation metrics, which makes direct comparison difficult [13]. Second, existing research adopts different models (LR, RF, XGBoost and MLP) to predict accident severity, and there are differences across studies in terms of data sources, data preprocessing techniques, and evaluation metrics, so it is still necessary to systematically compare the overall performance of these representative models based on a unified experimental procedure [16,18,19]. In addition, some studies are based on a limited set of metrics, such as accuracy, error-related metrics, and confusion matrices to evaluate the models; and fewer studies combine AUC and ROC curves to comprehensively assess the models' ability to identify the severe accident class [12]. Hence, using a unified experimental workflow, the study employs multiple evaluation metrics: Accuracy, AUC, ROC curves, and confusion matrices to perform a comprehensive direct comparison of four models: Logistic Regression, Random Forest, XGBoost and MLP models based on the US Accidents dataset. In addition, the most important predictive factors for the best-performing models are analyzed.

This study's overall workflow is presented in Figure 1, which mainly includes data sampling, data cleaning, feature engineering, model training, and model evaluation. The primary data source was the US Accidents 2016–2023 dataset, and 500,000 accident records were randomly sampled from the full dataset for experimentation [21]. The traffic accident data set includes traffic accident records from different geographic locations in the United States, including information about the time of the accident, the geographic location, weather conditions, traffic control facilities and the accident severity [21–23]. The original target variable was Severity. For this study this variable was converted into a binary variable named Severity_bin: Severity < 3 is coded as 0 (lower-severity accidents) and Severity ≥ 3 is coded as 1 (higher-severity accidents). In this process, the original multi-level severity variable was reduced to a binary classification result, thus establishing a common prediction target for the subsequent model training and performance comparison.

In the data preprocessing phase, the data were preprocessed by first identifying and removing duplicate rows and then counting missing values. In this study, Start_Time was converted into datetime format and three temporal features, Year, Month and Hour were extracted. The missing values of the numerical variables were filled with the median value, and for the categorical variables and the Boolean variables they were filled with the mode; the remaining missing values were imputed with the value 0. After that, irrelevant fields (like ID, description, street, city, ZIP code, and end time) were dropped to simplify the model inputs.

In terms of feature engineering, this study ensured that numerical features were kept and categorical features with a limited number of values (below 30) were one-hot encoded to avoid excessive dimensionality. The data was then split into an 80% training set and a 20% test set, with stratification to ensure consistent class distributions in both sets. Subsequent analysis of this model will review the model's ability to identify the severe accident class based on confusion matrices, since this study focuses on binary classification of severe and non-severe accidents.

Exploratory data analysis(EDA) was performed to understand data distributions before model construction to facilitate the interpretation of the results. The distribution of the binary severity variable was plotted to get a sense of the sample proportions of severe and non-severe accidents. In addition, a plot showing the proportion of severe accidents under different weather conditions was created, which is used for a preliminary comparison of the relationship between weather conditions and accident severity. This descriptive analysis is mainly used as reference for evaluating and analyzing the results of the model.

In this study, four models were built and compared: Logistic Regression, Random Forest, XGBoost, and MLP. Logistic Regression and MLP were trained with standardized features, whereas Random Forest and XGBoost were trained with encoded but non-standardized features. Finally, the performance of the models was assessed based on Accuracy, AUC, ROC curves, and confusion matrices and feature importance was assessed for the best-performing model(s).

Specifically, Accuracy measures the overall proportion of correct predictions; AUC measures the model's overall ability to distinguish between severe and non-severe accidents; ROC curves show the trade-off between true positive rate and false positive rate under different thresholds; and confusion matrices are used to see the cases where severe accidents are correctly classified or misclassified.

4. Result

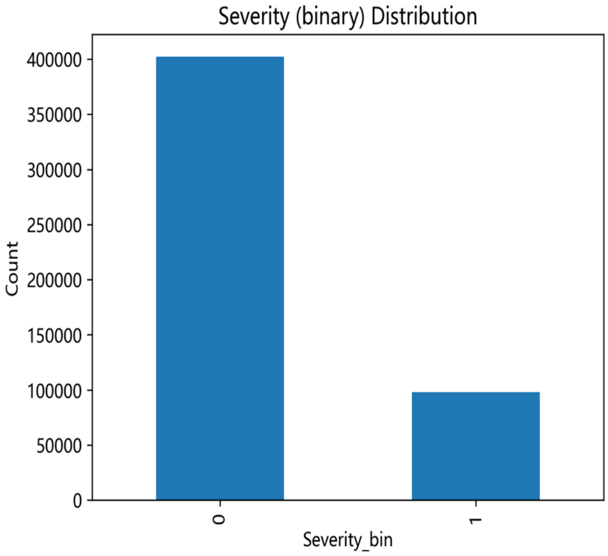

Figure 2. Severity binary distribution

First, the number of samples for non-severe accidents is much larger than that for severe accidents, as shown in the binary severity distribution. The dataset for this study is highly imbalanced with around 400,000 samples from the Severity_bin=0 class and around 100,000 samples from the Severity_bin=1 class. This distribution can lead the model to favor the majority class (in this case, non-severe accidents).

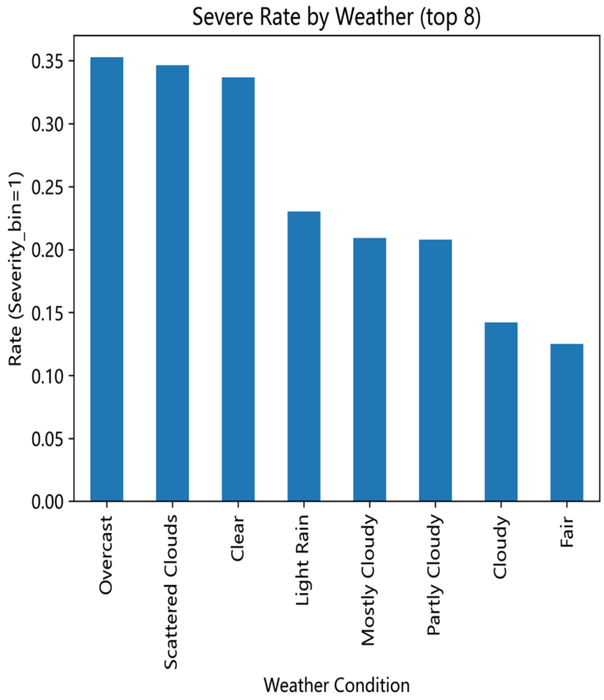

Figure 3. Severe accident rate by weather condition

In terms of weather features, some weather conditions have a higher share of severe accidents than others. The proportion of severe accidents is relatively high for weather types Overcast, Scattered Clouds and Clear, and relatively low for Cloudy and Fair weather types, as illustrated in Figure 3. This suggests a potential association between the weather conditions and the severity of the accident and therefore weather variables can be included as model inputs. However, this result is only descriptive and only indicates differences in severe accident proportions across weather conditions, rather than the weather directly influencing the severity of the accidents.

Table 1. Accuracy comparison of four models

Model	XGB	RF	MLP	LR
Accuracy	0.85958	0.85369	0.80524	0.80291

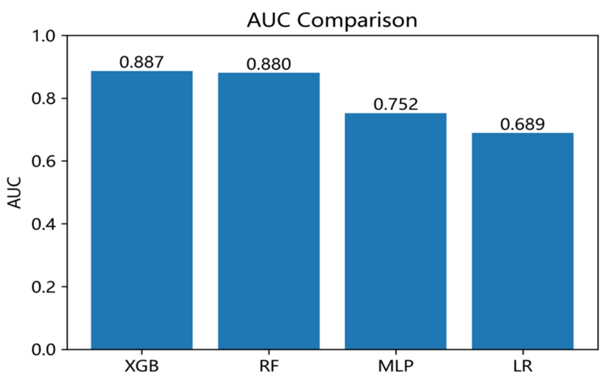

Figure 4. AUC comparison of four models

As for accuracy, XGBoost performed best overall with a score of 0.860 followed closely by Random Forest with a score of 0.854, while MLP and Logistic Regression came in at 0.805 and 0.803, respectively. Overall, XGBoost and Random Forest are more accurate. However, from the earlier mentioned distribution of accident severity, it is clear that there are a lot more samples of non-severe accidents than severe accidents. Thus, a high accuracy score alone does not guarantee the model's ability to detect severe accidents. In this regard, the present paper also provides a detailed analysis of model performance based on AUC, ROC curves, and confusion matrices.

When it comes to the AUC result, XGBoost recorded the highest AUC value of 0.887, and Random Forest's AUC is 0.880, closely following XGBoost. The AUC for MLP was 0.752, which is a moderate level. Logistic Regression had an AUC of 0.689, showing the weakest overall discriminative ability.

The ROC curves also exhibit the same pattern: the curves of the XGBoost and Random Forest are closer to the top-left corner, meaning that both these models have higher classification ability. MLP showed moderate performance and Logistic Regression was closer to random classification.

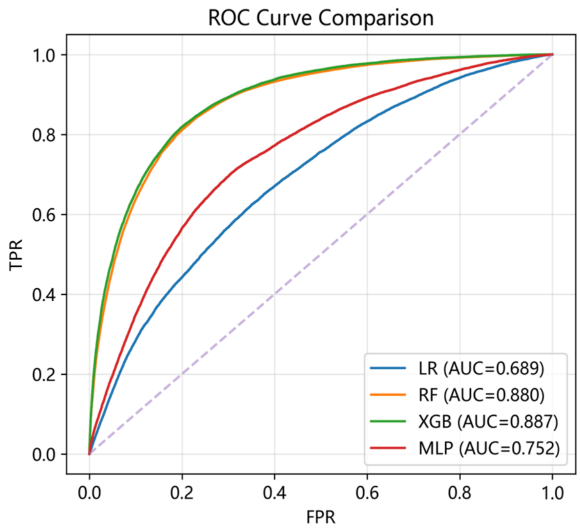

Figure 5. ROC curve comparison of four models

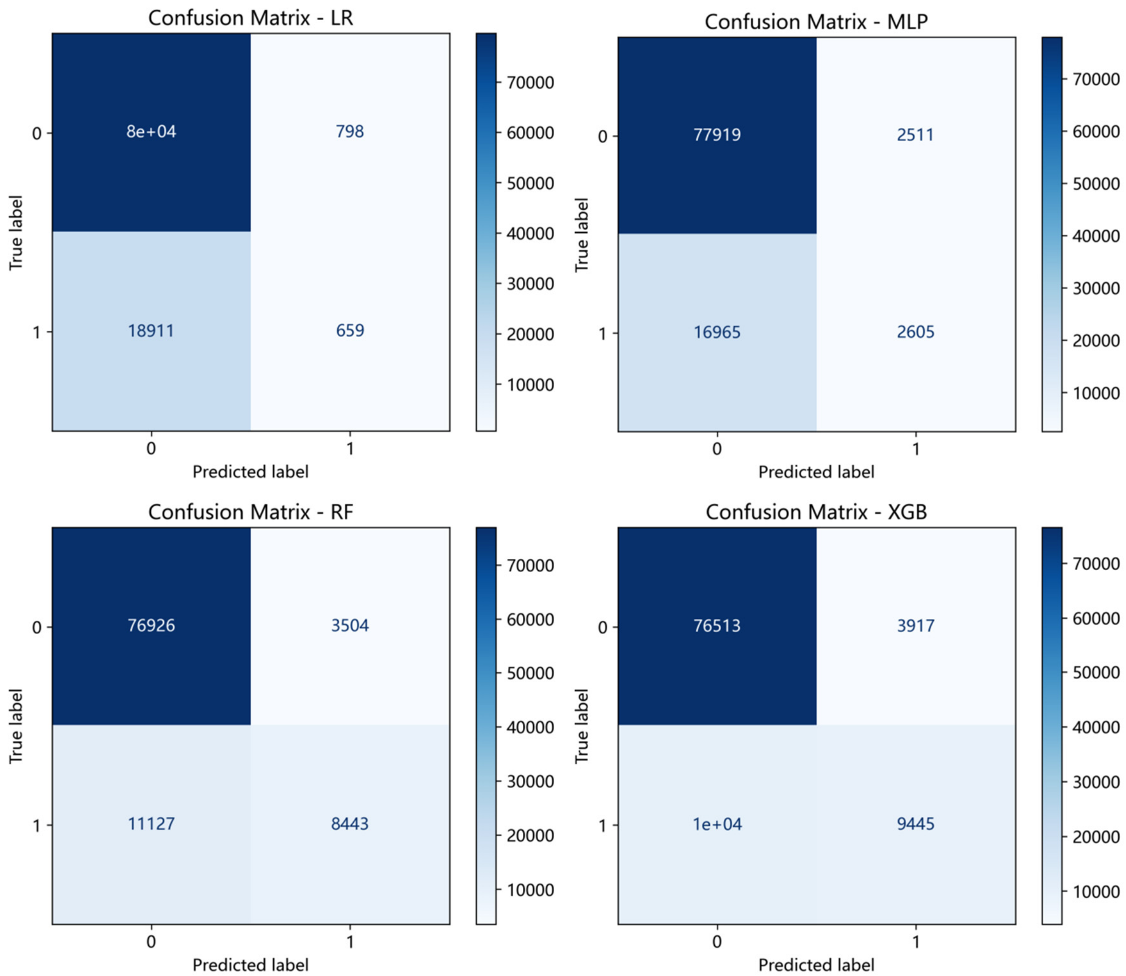

Figure 6. Confusion matrices of Logistic Regression, MLP, Random Forest and XGBoost

The confusion matrices also show that there are differences between the ability of different models in detecting severe accidents. Logistic Regression is able to accurately classify a high proportion of non-severe accidents while incorrectly classifying a significant proportion of severe accidents as non-severe. MLP shows some improvement over Logistic Regression, however it still has a substantial number of false negatives. Random Forest and XGBoost are more capable of identifying severe accidents, suggesting their ability to identify minority-class cases. However, these two models also produce more false positives, indicating that improving severe accident detection may come with an increase in false alarms.

This study further chose XGBoost for feature importance analysis because it achieved the best overall performance. The outcomes demonstrate that the model mainly relies on variables like traffic control facilities, temporal and spatial factors, and accident-affected distance. Specifically, Traffic_Signal, Stop, Crossing, Junction and Railway are traffic control facilities factors, Year, Month, Hour and nighttime illumination are temporal factors, Start_Lat, Start_Lng and Timezone are spatial factors, and Distance(mi) represents the accident-affected road distance. It should be noted that although these variables are deemed to be of high importance, this does not prove they cause severe accidents.

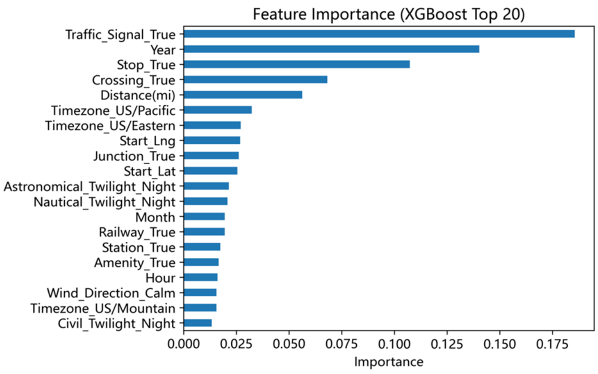

Figure 7. XGBoost feature importance ranking

5. Discussion

The study uses the US Accidents dataset to compare four models: Logistic Regression, Random Forest, XGBoost, and MLP to predict the severity of traffic accidents under a unified experimental workflow. In addition, a feature importance analysis was performed on the XGBoost model, which had the best performance. Results show that there are considerable differences between models when processing complex traffic accident data.

Firstly, in terms of model performance, the two models that performed best, overall, were XGBoost and Random Forest. The main reason is that both are ensemble tree models, which are better suited to handling nonlinear relationships and interactions between features [4]. Usually, the severity of traffic accidents is affected by many factors like weather, time, geographic location, traffic control facilities, and the accident-affected distance, which makes tree-based models more suitable for capturing high-dimensional and nonlinear data structures [3,13,24]. Logistic Regression is a linear model with a limited ability to capture complex variable relationships and shows weaker AUC performance and lower ability to identify severe accidents. It also provides evidence in line with the previous studies for which traditional statistical models have been shown to have limitations, especially in the presence of nonlinear relationships between variables [3]. While MLP can learn nonlinear relationships, it is sensitive to parameter settings, number of training epochs and network structures when learning structured tabular data. In the absence of extensive hyperparameter tuning and class imbalance handling, complex networks do not generally outperform ensemble tree models, and did not in this case, failing to improve on Random Forest or XGBoost [25].

Secondly, the feature importance results show that the most important features in the XGBoost model are mainly Traffic_Signal, Year, Stop, Crossing, and Distance(mi). Of these, Traffic_Signal, Stop, and Crossing are all related to traffic control facilities, which implies that the road environment (traffic lights, stop signs and crosswalks) may be linked to accident severity. This is in line with existing empirical evidence, which shows that changing the physical design of intersections and traffic control devices significantly affects the collision injury outcomes by influencing vehicle–pedestrian interactions [6,12]. The high importance of year may not only be associated with the way the data were collected, but may also be indicative of macro-level, long-term changes in traffic safety environments, policy changes, or even changes in the distribution of accidents in different years [19]. The accident-affected distance (measured in miles) represents the distance affected by the accident, which can suggest whether the size of the accident's impact is related to the severity level of the accident. As mentioned in previous studies, in physical space longer stretches of road are affected in more severe traffic accidents [3]. However these features only show that the model relies on them during prediction and do not prove that they cause severe accidents.

There are some limitations in this study. First, 500,000 samples were taken from the full dataset, which reduced the computational cost, but may not be representative of the full dataset. Secondly, the number of non-severe accidents is much greater than the number of severe accidents. This extreme class imbalance can cause the model to be biased during training and favor the majority class, which can affect its ability to detect severe accidents, as reported in previous research [17]. Third, this study did not conduct extensive hyperparameter optimization; for the MLP, it may not have reached its optimal performance. Larger scale datasets and SMOTE data augmentation combined with SHAP-based feature interpretation framework, and more sophisticated optimization algorithms such as Bayesian optimization, could be used to further improve the model's predictive performance, stability, and the interpretability of the results for the minority class of severe accidents in the future [17,18].

6. Conclusion

This study used the US Accidents 2016–2023 dataset, and developed four machine learning models – Logistic Regression, Random Forest, XGBoost, and MLP – to conduct binary classification of traffic accident severity. Accuracy, AUC, ROC curves, and confusion matrices were used to compare the performance of models. Further, a feature importance analysis was performed using the better-performing XGBoost model.

Experimental results showed that XGBoost achieved the highest AUC value, while Random Forest performed closely, indicating that tree-based models have strong predictive ability for traffic accident severity when processing complex traffic accident data. Logistic Regression had a relatively weaker overall discriminative ability, and MLP showed moderate performance. The feature importance analysis reveals that traffic control facilities, time factors, spatial location and accident-affected distance are important features for the model.

In general, machine learning techniques can effectively assist in predicting the severity of traffic accidents and provide a reference for traffic safety management and risk assessment. However, class imbalance remains an important challenge in identifying severe accidents. Future research can further improve this study from three aspects: class imbalance handling, model hyperparameter tuning, and feature interpretation, to improve the ability to identify severe accidents and enhance model interpretability.

References

[1] Sharafeldin, M., Farid, A., & Ksaibati, K. (2022). A random parameters approach to investigate injury severity of two-

vehicle crashes at intersections. *Sustainability*, 14(21), 13821. https:// doi.org/10.3390/su142113821.

[2] World Health Organization. (2026). *Road traffic injuries*. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

[3] Yan, M., & Shen, Y. (2022). Traffic accident severity prediction based on random forest. *Sustainability*, 14(3), 1729. https:// doi.org/10.3390/su14031729.

[4] Tang, J., Huang, Y., Liu, D., Xiong, L., & Bu, R. (2025). Research on traffic accident severity level prediction model based on improved machine learning. *Systems*, 13(1), 31. https:// doi.org/10.3390/systems13010031.

[5] Wada, Y., Asami, Y., Hino, K., Nishi, H., Shiode, S., & Shiode, N. (2023). Road junction configurations and the severity of traffic accidents in Japan. *Sustainability*, 15(3), 2722. https:// doi. org/10.3390/su15032722.

[6] Haleem, K., Alluri, P., & Gan, A. (2015). Analyzing pedestrian crash injury severity at signalized and non-signalized locations. *Accident Analysis & Prevention*, 81, 14–23. https://doi.org/ 10. 1016/ j.aap.2015.04.018.

[7] Ibtihal, S. A., & Rifaat, S. M. (2026). Crash occurrence and severity at railway level crossings in Bangladesh. *Transportation Research Interdisciplinary Perspectives*, 36, 101840. https://doi.org/10.1016/j.trdip.2026.101840.

[8] Sun, Z., Wang, J., Chen, Y., & Lu, H. (2018). Influence factors on injury severity of traffic accidents and differences in urban functional zones: The empirical analysis of Beijing. *International Journal of Environmental Research and Public Health*, 15(12), 2722. https://doi.org/10.3390/ijerph15122722.

[9] Jafari Anarkooli, A., & Hadji Hosseinlou, M. (2016). Analysis of the injury severity of crashes by considering different lighting conditions on two-lane rural roads. *Journal of Safety Research*, 56, 57–65. https://doi.org/10.1016/j.jsr.2015.12.003.

[10] Safari, M., Effati, M., & Arabani, M. (2024). Spatial analysis of daytime and nighttime crash severity on horizontal curves of mountainous rural highways: A case study in Northern Iran. *Results in Engineering*, 24, 103344. https://doi.org/10. 1016/j. rineng. 2024.103344.

[11] Zhang, K., & Hassan, M. (2019). Crash severity analysis of nighttime and daytime highway work zone crashes. *PLOS ONE*, 14(8), e0221128. https://doi.org/10. 1371/ journal. pone. 0221128.

[12] Tanishita, M., & Sekiguchi, Y. (2023). Impact analysis of road infrastructure and traffic control on injury severity of single- and multi-vehicle crashes. *Sustainability*, 15(17), 13191. https:// doi.org/10.3390/su151713191.

[13] Shaffiee Haghshenas, S., Guido, G., Vitale, A., & Astarita, V. (2023). Assessment of the level of road crash severity: Comparison of intelligence studies. *Expert Systems with Applications*, 234, 121118. https://doi.org/10. 1016/j. eswa. 2023. 121118.

[14] Alam, M. S., & Tabassum, N. J. (2023). Spatial pattern identification and crash severity analysis of road traffic crash hot spots in Ohio. *Heliyon*, 9(5), e16303. https://doi.org/ 10. 1016/ j. heliyon. 2023. e16303.

[15] Alsahfi, T. (2024). Spatial and temporal analysis of road traffic accidents in major Californian cities using a geographic information system. *ISPRS International Journal of Geo-Information*, 13(4), 157. https://doi.org/10.3390/ijgi13040157.

[16] Al-Ghamdi, A. S. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. *Accident Analysis & Prevention*, 34(6), 729–741. https://doi.org/ 10. 1016/ s0001-4575(01)00073-9.

[17] Chen, Y. (2026). Research and analysis of traffic accident severity prediction based on data augmentation and feature interpretation. In *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)* (pp. 685–692).

[18] Zhou, W. (2024). Traffic accident severity prediction based on data cleaning and machine learning (Random Forest / XGBoost). *Highlights in Science, Engineering and Technology*, 85, 376–388.

[19] Ruangkanjanases, A., Sivarak, O., Weng, Z.-J., Khan, A., & Chen, S.-C. (2024). Using multilayer perceptron neural network to assess the critical factors of traffic accidents. *HighTech and Innovation Journal*, 5(3), 157–169. https://doi. org/ 10.28991/hij-2024-05-03-01.

[20] Arhin, S. A., & Gatiba, A. (2020). Predicting crash injury severity at unsignalized intersections using support vector machines and naïve Bayes classifiers. *Transportation Safety and Environment*, 2(2), 120–132. https://doi.org/ 10. 1093/ tse/ tdaa011.

[21] Moosavi, S. (n.d.). *US Accidents (2016–2023)* [Dataset]. Kaggle. https://doi.org/10.34740/kaggle/ds/199387.

[22] Moosavi, S., Samavatian, M. H., Parthasarathy, S., & Ramnath, R. (2019). A countrywide traffic accident dataset. arXiv. https: // doi.org/10.48550/arXiv.1906.05409.

[23] Moosavi, S., Samavatian, M. H., Parthasarathy, S., Teodorescu, R., & Ramnath, R. (2019). Accident risk prediction based on heterogeneous sparse data: New dataset and insights. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 33–42). https://doi.org/10.1145/3347146.3359078.

[24] Pourroostaei Ardakani, S., Liang, X., Mengistu, K. T., So, R. S., Wei, X., He, B., & Cheshmehzangi, A. (2023). Road car accident prediction using a machine-learning-enabled data analysis. *Sustainability*, 15(7), 5939. https://doi.org/ 10. 3390/ su15075939.

[25] Sameen, M., & Pradhan, B. (2017). Severity prediction of traffic accidents with recurrent neural networks. *Applied Sciences*, 7(6), 476. https://doi.org/10.3390/app7060476.