Iran University of Science and Technology

International Journal of

Railway Research

ISSN: 2423-3838

Data-Driven Techniques for Identifying Factors Affecting the Severity of Driver Injuries in Highway-Railway Grade Crossing Accidents: A Comparative Analysis Using Random Forest, XGBoost,and Multinomial Logistic Regression

Rayehe sadat Mousavi[1], Behnam Bagherian[1], Zahra sadat sanagostar[1], Mohammadali Zayandehroodi[2], Zahra Saghian[3], Morteza Bagheri *[1]

[1] School of Railway Engineering, Iran University of Science and Technology, Tehran, Iran

[2] School of Civil Engineering, Iran University of Science and Technology, Tehran, Iran

[3] School of Industry Engineering, Iran University of Science and Technology, Tehran, Iran

ARTICLE INFO

Article history:

Received: 12.11.2024

Accepted: 31.05.2025

Published: 09.06.2025

Keywords:

grade crossings

severity of the incidents

machine learning

random forest

XGBoost

A B S T R A C T

This study investigates the factors influencing the severity of accidents at highway-rail grade crossings in the United States and explores strategies to mitigate the risks to road vehicle drivers. Two approaches are employed for modeling accident severity: statistical methods, such as multinomial logistic regression, and machine learning techniques, including Extreme Gradient Boosting (XGBoost) and random forest algorithms. The analysis is based on data from the Federal Railroad Administration's database, covering a twelve-year period (2010-2022). The results identify several key controllable factors that significantly impact accident severity, including vehicle speed, the position of road users, visibility obstructions, the number of cars in the train, and the speed of the train. Among the models tested, XGBoost demonstrated superior accuracy in predicting accident severity compared to multinomial logistic regression and random forest. Based on the findings, several recommendations are proposed to reduce accident risk at grade crossings, such as lowering train speeds, implementing advanced speed control systems, enhancing lighting at crossings, improving barrier inspections, and optimizing train scheduling. These measures aim to enhance safety and minimize collision severity at highway-rail crossings.

1. Introduction

In April 2024, a train hit a van on a road at a railroad crossing in southwestern Idaho in the western United States. Four people in the van died at the scene [1]. According to the National Safety Council tracked of America, quoting the FRA, there were 274 deaths in grade crossing incidents in 2022[2]. Based on the FRA data, there were 2,578 incidents at grade crossings in the United States between 2013 and 2023; 467 of the incidents were fatal, while 1,139 of the incidents led to injuries. Data from the FRA reveal that between 2002 and 2011, the driver's behavior contributed to the level of incidents at grade crossings [3][4]. Referring to Figure 1, it is evident that driver behavior has impacted the incidents at grade crossings in the United States that caused fatalities and

*Corresponding author
Email address: morteza_bagheri@iust.ac.ir

injuries; this calls for research on driver behavior.

The transport system significantly relies on grade crossings that facilitate the interchange of traffic between railways and roads. The National Safety Council report of 2022 reveals that a train or vehicle hits a train every four hours in the United States, which means that these crossings are important for safety and productivity [5]. As per the statistics of 2023, 247 individuals died in grade crossing incidents in the United States, and over 766 individuals were injured. Thus, grade crossings are known to be points that may be improved [6]. The FRA report for the year 2023 shows that 225 incidents occurred at grade crossing in the United States. The following statistics reveal that while there are many researches conducted on the safety of these points, the incidence of incidents at these crossings is high. However, in addition to the frequency of incidents at these points, one should focus on the severity of the incidents. In this regard, many characteristics and parameters influence the level of risk at the grade crossing. However, it is necessary to identify which are more critical and have the highest impact on the level of risk at the grade crossing. According to the FRA and the National Highway Traffic Safety Administration of the United States, in the cases of road vehicles and train collisions, the likelihood of the road vehicle driver's death is more than 20 times higher than in other road incidents [7].

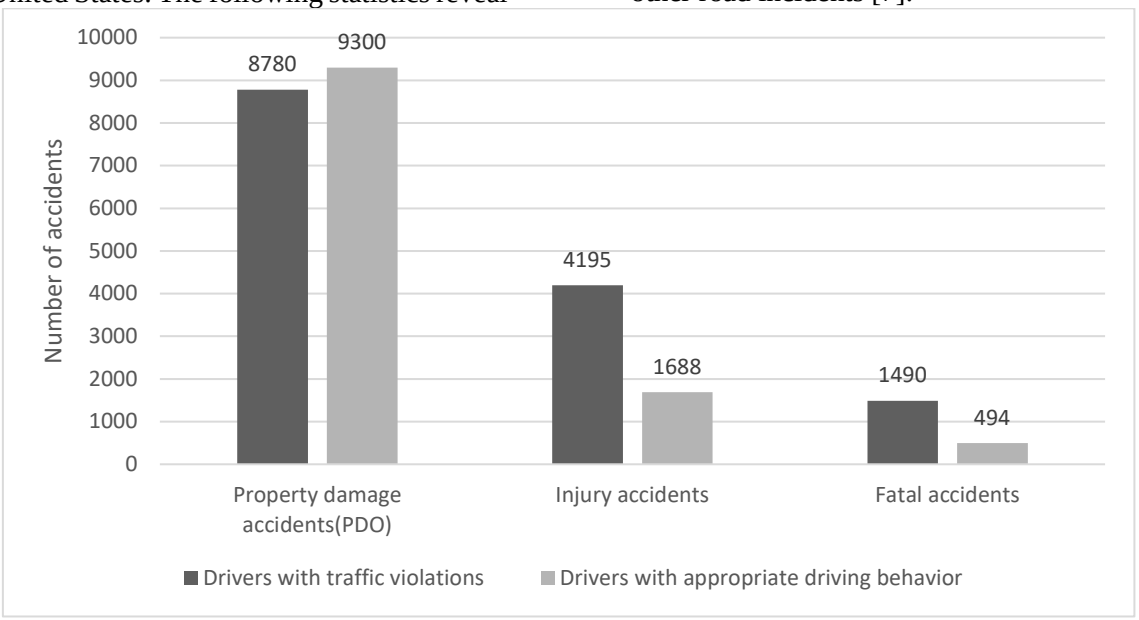

Fig. 1: The role of road vehicle driver behavior in the level of severity of the incident at the grade crossings in the United States [4]

The FRA report for the year 2023 shows that 225 incidents occurred at grade crossing in the United States. The following statistics reveal that while there are many researches conducted on the safety of these points, the incidence of incidents at these crossings is high. However, in addition to the frequency of incidents at these points, one should focus on the severity of the incidents. In this regard, many characteristics and parameters influence the level of risk at the grade crossing. However, it is necessary to identify which are more critical and have the highest impact on the level of risk at the grade crossing. According to the FRA and the National Highway Traffic Safety Administration of the United States, in the cases of road vehicles and train collisions, the likelihood of the road vehicle driver's death is more than 20 times higher than in other road incidents [7]. Thus, this study aims at identifying factors that contribute to the severity of incidents of road vehicle drivers from a different perspective by looking at the information on past incidents. Sufficient comprehension of the factors that affect the degree of incidents at grade crossings will enable the identification of factors that have a higher

impact, thus enhancing the safety level at grade crossings. This study aims to answer two important questions: 1) What are the factors influencing the severity of incidents for road vehicle drivers at grade crossings? 2) Which factors have a more significant impact on the severity of incidents for road vehicle drivers at grade crossings?

In their 2012 research, based on the literature review from 2010 onwards, Novin Eluru et al. used a latent class modeling technique to analyze factors influencing the severity of road vehicle drivers' injuries at highway-railway grade crossings. This method applies probabilistic models to determine hidden classes or clusters of data that are hidden in the dataset but are different from each other. The paper employs an ordinal algorithm with reference to the latent segmentation. This algorithm can compare the influence of different factors on the extent of the injury and define the influence of each factor in different groups [8]. The same year in 2016, Ghomi et al. conducted a study on factors influencing injury severity of incidents at grade crossings involving vulnerable transport drivers. Some of the techniques used in this article include the association rules model and classification decision tree algorithms on data of grade crossing incidents from the year 2007 to 2013. The findings of the research thus reveal that train speed has the highest correlation with the severity of the injuries to the drivers of public transport in incidents. Also, appropriate lighting helps in decreasing the impact of the incident. The probability of the severity of the injury also rises with the occurrence of snow or rain, especially where the train is moving at a faster pace. Public transport drivers are relatively older and women are likely to be more severely injured particularly at night and in adverse weather conditions [9]. The same authors, in a subsequent article the following year, examined factors affecting the severity of injuries to road vehicle drivers at highway-railway crossings using data mining algorithms. Their article aims to identify factors leading to driver injury severity in highway-railway grade crossing incidents, focusing on discovering interactions and differences between these factors. Their paper utilized classification decision tree algorithms and association rules on the Federal Railroad Administration's highway-railway grade crossing incident database for the period 2006 to 2013 to identify factors affecting the severity of injuries to road vehicle drivers in highway-railway grade crossing incidents. Both algorithms were effective in providing meaningful insights into incident factors and their interactions. The results of the two algorithms were never contradictory. Results indicate that train speed, type of road vehicle, driver's age and gender, position of the road vehicle before the collision, type of incident, and type of road surface are among the key factors affecting driver injury severity [10]. On a similar topic using data from grade crossing incidents in the United States, Fan et al. conducted an analysis of vehicle incident severity at highway-rail grade crossings in 2016. The objective of this study was to develop an ordered logit response model to examine the impact of various variables on three different levels of severity in vehicle incidents at highway-rail grade crossings in the United States. The model results show that vehicle and train incidents at high speeds, in high temperatures, with male drivers and older drivers, increase the likelihood of fatal incidents and serious injuries. Additionally, incidents involving trucks with trailers, incidents under weather conditions such as rain and snow, in commercial, residential, industrial, and institutional areas, with asphalt and flange road surfaces, and with daily traffic exceeding 20,000, reduce the likelihood of fatal incidents and serious injuries [11]. In the same year, Zheng et al. carried out a study with the help of machine learning techniques. The purpose of this research was to analyze and assess the decision tree model for estimating the incident rates at highway-rail grade crossings and determining the key factors. The data applied in this research concerned grade crossing incidents in one of the United States during the period of 1996-2014. The analysis has revealed that factors such as the amount of road and rail traffic, as well as the speed of trains, are some of the key drivers of the probability of an incident. Furthermore, warning systems and train detection aids also minimize the chances of an incident occurring [12]. In 2018, Chang Shima et al. also, in their study, focused on vehicle driver behavior in grade crossing incidents. The aim of this article was to estimate the impact of aggressive driving behaviors on driver injury severity in incidents at highway-rail grade crossings. Chang Shima et al. also utilized ten years of data from the FRA of the United States in their study and implemented a mixed logit modeling approach.

They examined the determinants of driver injury severity both with and without considering aggressive driving behaviors at highway-rail grade crossings [4]. Continuing the use of data mining methods in grade crossing incidents, Soleimani et al. in 2021 implemented the possibility of using advanced machine learning and text mining techniques to retrieve information from highway-rail grade crossing incident data. They explored a platform for developing an integrated model capable of identifying the most suitable highway-rail intersections for closure. Their results showed that 15 percent of crossings in the eastern section should be closed or undergo safety improvements [13]. With the success of the machine learning approach in this subject, Rana et al. in 2023 examined hotspot points and highway-rail grade crossing incidents. Given the development of social areas around railway tracks, there has been an increase in road-rail intersection points, leading to incidents, fatalities, and injuries. This study focused on identifying high-potential incident points and factors affecting the severity of casualties at road-rail intersection points. Using a machine learning model to analyze road-rail intersection incidents and associated fatality severity between 2001 and 2022 in Canada, derived from two sources including data obtained from the Canadian government and data from the Railway Incident Database System, they were able to manage the complex relationship between variables used for modeling. Based on these results, authorities and officials can evaluate these locations and implement necessary safety measures to reduce casualties, such as installing lighting sources, clearing obstacles from adjacent areas, or installing gates and automatic railway control crossings, among others [14].

Reviewing previous studies, it can be concluded that although numerous efforts have been made to identify and analyze factors influencing the severity of incidents at highway-rail grade crossings, there is still a need for further investigations and improvement of prediction methods. This is because the characteristics present in incident data have complex interactions and high dispersion, which conventional statistical methods or previous machine learning algorithms are unable to account for the specific features of these incidents. For example, Novin Eluru et al. (2012) used latent class modeling to identify influential factors, while Ghomi et al. (2016) examined drivers using data mining models. Additionally, Chang Shima et al. (2018) analyzed aggressive driving behaviors and their impact on injury severity, and Soleimani et al. (2021) employed machine learning techniques to develop models. Fewer studies have focused on the severity of road vehicle driver injuries; these injuries have complex interactions with characteristics such as vehicle speed, number of cars, and visual obstacles. The XGBoost algorithm can identify these complex patterns and non-linear interactions between variables. Accurate identification of influential features on injury severity and their ranking, given the high dispersion of incident data, requires high precision. The Random Forest algorithm, by creating a large number of decision trees and combining their results, provides higher accuracy in feature ranking compared to other algorithms. Therefore, considering the high data dispersion and complexity of interactions, using XGBoost and Random Forest algorithms complementarily is essential for identifying the impact of various features on road vehicle driver injury severity. Moreover, using the XGBoost or Random Forest algorithms alone cannot simultaneously manage data dispersion and complex interactions between data. Most of the previous researches have solely relied on statistical analysis like ordered logit response model. However, the multinomial logistic regression model has less restrictive assumptions concerning the order and continuity of the dependent categories, making it more versatile in the analysis of data. Besides, the multinomial logistic regression model is more appropriate in handling the interactions between the independent and dependent variables, making it more effective in establishing and explaining different factors that may be associated with the severity of an incident. Previous studies have rarely investigated a broad range of significant factors and combined them with state-of-the-art machine learning techniques and statistical procedures. This research employs wider data that covers a longer operational time and integrates various modeling approaches to have a better and more accurate estimation of influential factors, hence, improved prediction and minimized prediction errors. The combination of a statistical model with other

machine learning models is ideal for initial evaluation and the determination of significant variables because of the model's simplicity and ease of interpretation. In other words, this study, using larger and better data and more sophisticated methods, aims to gain a better and more precise picture of factors affecting the severity of incidents at highway-rail grade crossings and to identify sound approaches to improving safety at these locations.

2. Method

This research aims to investigate factors contributing to the severity of road vehicle driver incidents at highway-rail grade crossings and to make recommendations regarding measures to reduce the severity of such incidents. This section defines the main research methodology and the phases of the research. The first step is data gathering. As noted earlier, the data used in this research were obtained from the FRA of the United States for a period of 12 years, from 2010 to 2022. This dataset contains all the necessary data concerning incidents at highway-rail grade crossings. The variables used in this research are shown in Table 1 below. Further, every variable in the database is given a certain code, which is explained in the explanatory part of Table 1.

The dataset contained 15,716 rows in the initial stage. Data with many missing or wrong values was detected and omitted to enhance the quality and accuracy. If only one feature had missing values, then those missing values were replaced by the mode of the feature. Finally, 15,705 rows were used for the model input. The scales of quantitative variables were standardized to reduce the effect of the difference in scales in the models, while categorical variables were encoded to numerical values for use in the machine learning models. In the subsequent stage, a correlation matrix was computed from the data with the help of SPSS software. This matrix helps determine the independent and dependent variables and also brings out variables that are closely related. Finally, variables with correlation coefficients higher than 0.7 were eliminated from the feature set. Out of the 25 variables mentioned in Table 1, 15 are independent of each other and were chosen as input variables for the model. Such variables include temperature, number of cars, age of the road vehicle driver, train speed, estimated speed of the vehicle, road user, position of the intersection warning, class of the track, visibility, weather conditions, position of the road user, type of equipment involved, presence of visual obstructions, whether the crossing is public or private, and type of track.

The study employs two modeling approaches, supervised learning algorithms and statistical models, due to their specific advantages and disadvantages. Machine learning models like XGBoost and Random Forest handle larger volumes of data and identify complex patterns but are difficult to interpret. Statistical models like multinomial logistic regression are easier to interpret and useful for initial examinations and variable selection. Combining these approaches leverages their strengths for a more accurate assessment of incident severity factors. The statistical approach uses multinomial logistic regression for recognizing and analyzing factors affecting incident severity. This model can handle non-linear data and test the effect of multiple independent variables on a dependent variable, making it suitable for incident data where various factors combine with incident severity [15]. SPSS software is used for the statistical model, while Python is used for machine learning methods. Two machine learning algorithms were chosen: XGBoost and Random Forest. XGBoost captures interactions between features and non-linear relationships, essential for analyzing incident data with high interactions and dispersion [16]. Random Forest models complex feature interactions and has high accuracy in ranking important features [18]. XGBoost handles data dispersion well [17], while Random Forest performs more stably in dispersed data by resampling and creating multiple trees. Combining these algorithms improves prediction and reliability. For better performance, grid search was used to find the best parameters for XGBoost and Random Forest models. After optimization, models were retrained, and their evaluation metrics recalculated. Modeling was conducted as classification to categorize road vehicle driver incident severity into three classes: fatal, injurious, and non-injurious. Classification is fundamental in machine learning and data mining, assigning samples to different classes or labels. The statistical approach uses polynomial regression to check results and make comparisons, suitable for handling non-linear data and testing multiple independent variables' effects on a dependent variable.

Table 1: Data Description

Row	Variable	Description
1	Report Year	2010 to 2022
2	Time	Before noon: 1, After noon: 2
3	Crossing Public/Private Code	Public: 1, Private: 2
4	Road User	1: Passing through safety gates, 2: Stopping then continuing, 3: Passing without stopping, 4: Stopping at intersection, 5: Other, 6: Passing through temporary barriers, 7: Going in front of safety gate and illegally passing through safe zone, 8: Suicide/Suicide attempt
5	Estimated Vehicle Speed	Estimated speed (in miles per hour) at time of collision
6	Vehicle Direction	1: North, 2: South, 3: East, 4: West
7	Road User Position	1: Stopped or trapped at intersection, 2: Stopped at intersection, 3: Crossing intersection, 4: Trapped at intersection due to traffic, 5: Blocked at intersection due to gates
8	Equipment Involved	1: Train (traction units), 2: Train (pushing units), 3: Train (stopped), 4: cars (s) (moving), 5: cars (s) (stopped), 6: Light locomotive (moving), 7: Light locomotive (stopped), 8: Other cases
9	Temperature	In Fahrenheit
10	Visibility	1: Dawn, 2: Day, 3: Dusk, 4: Darkness
11	Weather Conditions	1: Clear, 2: Cloudy, 3: Rainy, 4: Fog, 5: Blizzard, 6: Snow
12	Equipment Type	1: Freight train, 2: Passenger train
13	Track Type	1: Main Track, 2: Maneuvering Track, 3: Side Track, 4: Industrial Track
14	Track Class	1 to 9
15	Number of Locomotive cars	-
16	Number of cars	-
17	Train Speed	In miles per hour
18	Train Direction	1: North, 2: South, 3: East, 4: West
19	Road Conditions	1: Dry, 2: Wet, 3: Snow or mud, 4: Ice, 5: Gravel, dirt, oil, sand, 6: Water (stagnant, moving)
20	Intersection Warning Position	1: Both sides, 2: Near side of vehicle, 3: Opposite side of vehicle
21	Warning Connected to Signal	1: Yes, 2: No
22	User Age	-
23	User Gender	1: Male, 2: Female
24	Visual Obstacles	1: Permanent structure, 2: Railway equipment, 3: Passing train, 4: Topography, 5: Plants, 6: Road vehicles, 7: Other, 8: No visual obstacle
25	Driver Status	Fatal, Injured, Uninjured

Model evaluation and comparison were conducted, and recommendations to minimize the severity of road vehicle driver incidents at level crossings were made based on the results. Python was used for data preprocessing and machine learning models, and SPSS for forming the correlation matrix and statistical modeling.

3. Results

To analyze and examine factors affecting the severity of road vehicle driver incidents at level crossings of roads and rails, features are divided into two groups: uncontrollable and controllable variables. Uncontrollable features include those that cannot be managed and consist of temperature, age of the vehicle driver, visibility, weather conditions, type of track, class of track, and whether the crossing is private or public. Controllable features are those that can be influenced by managerial actions, such as train speed, number of cars, estimated vehicle speed, road user status, intersection warning position, road user position, involved equipment, and visual obstacles. This classification aims to enhance the understanding and management of factors that impact safety and incidents at road-rail level crossings [19].

As depicted in the table 2, the XGBoost model, with higher accuracy and performance compared to the Random Forest model, is the most appropriate model to use in predicting the severity of road vehicle driver incidents in this study. In Table 2, it is possible to observe the accuracy and other comparable characteristics. Table 2 shows the comparison of two algorithms, namely "XGBoost" and "Random Forest," based on the severity of road vehicle driver incidents at level crossings of roads and rails. The XGBoost algorithm has a slightly higher accuracy of 72 and F1 score of 69. However, it is better in terms of precision and recall rate, so it can be considered that the proposed approach is effective. On the other hand, the Random Forest algorithm has a slightly lower accuracy of 69 and a recall rate of 71. As a result of this comparison, it can be deduced that while the overall performance of both algorithms is almost the same, there are differences in certain aspects.

Table 2: Comparing the performances of various algorithms in forecasting the severity of road vehicle driver incidents at level crossings of roads and rails.

Algorithm	Accuracy	Precision	Recall	F1-Score
Extreme Gradient Boosting	0.72	0.64	0.72	0.69
Random Forest	0.71	0.69	0.71	0.68

The results obtained from modeling with the two aforementioned algorithms, which represent the variables influencing the severity of road vehicle driver incidents at level crossings of roads and rails, are shown in Figures 2 and 3.

First, the Random Forest model results are shown; this algorithm, thanks to its simple and easily understandable structure, can be used as an introduction to the results analysis. The random tree, which is a hierarchical structure, can directly point out which variables have the largest effect on the incident's severity. The Random Forest model for the severity of road vehicle driver incidents at level crossings of roads and rails is presented in Figure 2. These results are depicted in the form of feature importance chart and show the significance of each feature in the model performance. Based on the obtained results, the number of cars was found to be the most important controllable factor with an importance of 0.142, which means that the number of cars has a direct effect on the severity of road vehicle driver incidents at level crossings of roads and rails. This could be because of the increased length and weight of the trains, which results in increased kinetic energy. After that, train speed with a weight of 0.133 and estimated vehicle speed with a weight of 0.077 have been defined as the most critical factors in the model based on the importance level. Higher train speed, because of increased kinetic energy, can cause more serious incidents, and the high speed of road vehicles can aggravate the severity of road vehicle driver incidents at level crossings of roads and rails because of the short time for reaction and increased kinetic energy in impacts. As for the uncontrollable factors, the most important one is temperature, with an importance of 0.145, followed by the age of the road vehicle driver, with an importance of 0.141.

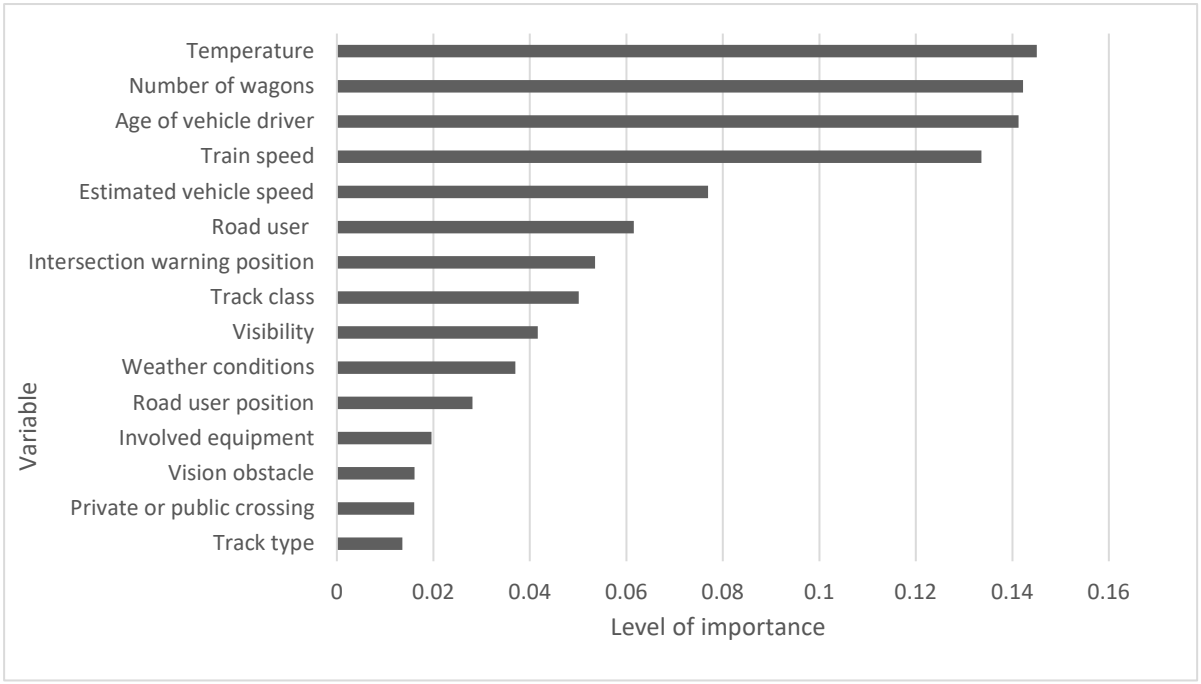

Fig. 2: The importance of variables on the severity of road vehicle driver incidents at level crossings of roads and rails based on the Random Forest algorithm

Figure 3 also shows the outcome of the XGBoost model for the severity of road vehicle driver incidents at level crossings of roads and rails. The XGBoost algorithm is more complex than the Random Forest algorithm and hence can capture more complex and non-linear interactions between variables. This algorithm is based on gradient boosting techniques to enhance the base models and can give better results in prediction. For this reason, presenting the results of this algorithm after the Random Forest, which has a relatively simpler and more linear structure, enables the analysis of features affecting the severity of road vehicle driver incidents to be carried out with basic analyses and then with more complex analyses. According to the results obtained from the XGBoost model, visual obstacles, which according to Table 1, with an importance of 0.259 include eight states (1: Permanent structure (fixed and immobile obstacles such as buildings, walls, bridge pillars, etc. around the crossing), 2: Railway equipment, 3: Passing train, 4: Topography, 5: Plants, 6: Road vehicles, 7: Other obstacles, 8: No visual obstacle), have been identified as the most important controllable feature in the model. This means that the existence of visual barriers can influence the extent of the incident rate of road vehicle drivers at level crossings of roads and rails. Following the visual obstacles, the next most significant controllable feature in the XGBoost model is the road user position with an importance of 8. Table 1 shows that the road user position, which has an importance of 0.08, has five states: 1: Stopped or trapped at the intersection, 2: Stopped at the intersection, 3: Crossing the intersection, 4: Trapped at the intersection due to traffic, and 5: Blocked at the intersection, has been identified as the next most important controllable feature in the XGBoost model. An improper position on the road may be attributed to low visibility and a high probability of encountering the train, which raises the risk of road vehicle driver injuries at level crossings of roads and rails. In this model, two attributes, namely track type, which is the main track, the maneuvering track, the side track, or the industrial track, with a weight of 0.166, and the private/public crossing, with a weight of 0.115, have influenced the severity of road vehicle driver incidents at the level crossings of roads and rails as the uncontrollable factors.

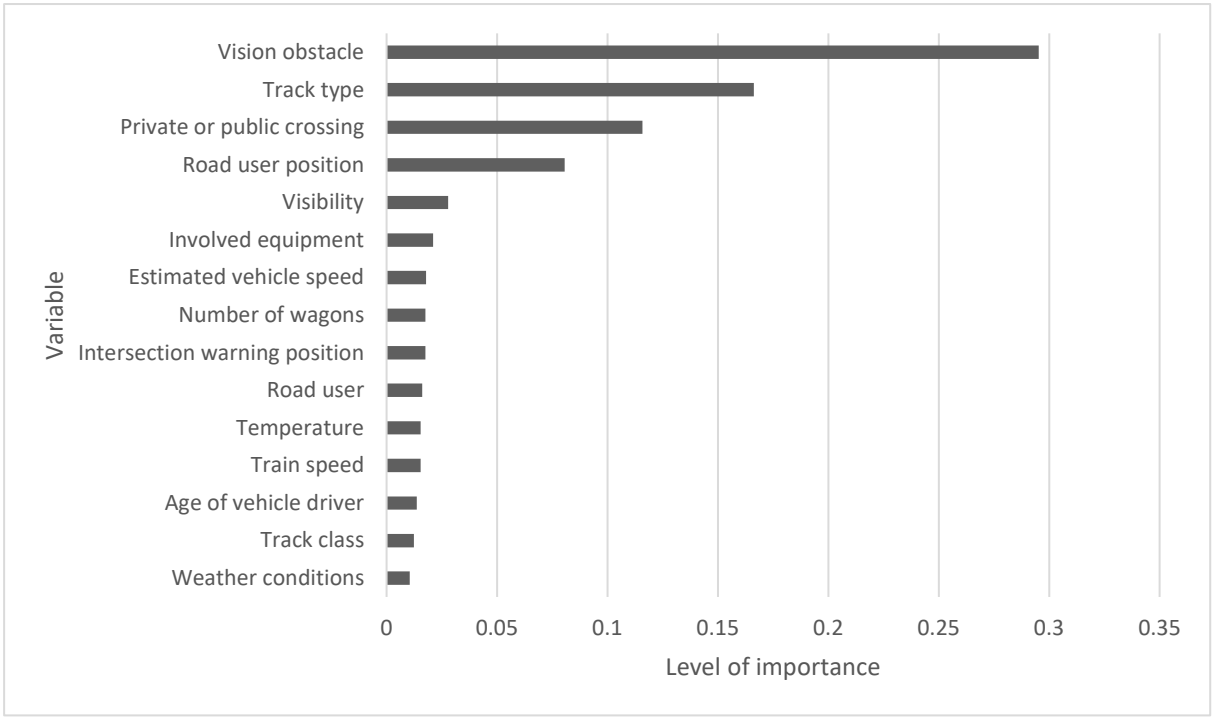

Fig. 3: The importance of variables on the severity of road vehicle driver incidents at level crossings of roads and rails based on the XGBoost algorithm

In the statistical modeling section, the multinomial logistic regression model was used after the use of the Random Forest and XGBoost algorithms. As mentioned, this model was chosen due to its ability to model the relationship between a multi-state dependent variable (car driver status at three levels: fatal, injurious, and non-injurious) and multiple independent variables. The reason for presenting the multinomial logistic regression model after the machine learning algorithms is the ease of interpreting results, validating and confirmation of more complex results, as well as the ability to compare and contrast the results with other sophisticated models .Table 3 shows the results of statistical modeling of the impact of various parameters on the severity of road vehicle driver incidents at level crossings of roads and rails. In this table, the beta value represents the impact coefficient of each parameter, and the significance level indicates its statistical importance. If the significance level is less than 0.05, the parameter is considered significant. A positive beta means an increase in incident severity with an increase in the variable, and a negative beta means a decrease in incident severity with an increase in the variable.

According to Table 3, for example, the estimated vehicle speed parameter, as a controllable feature, with a beta of 0.462, a significance level of 0.005, and an impact coefficient of 1.891, indicates that as vehicle speed increases, incident severity significantly increases. In another example, the road user position parameter, also a controllable feature, has a negative beta value, indicating that as the value of this variable increases, incident severity decreases. This shows that at a significance level of 0.03, this parameter is statistically significant. Furthermore, the value of 0.726 for this parameter shows that the shift of position of the road users lessens the likelihood of an incident by 72.6 percent. All in all, the probability estimates of the multinomial logistic regression model show that two variables, estimated vehicle speed and road user position, which are the controllable features, have a significant influence on the severity of road vehicle driver incidents at level crossings of roads and rails.

When the speed of the road vehicle increases, the level of the incident rises as well. As with the case of the Random Forest algorithm, this could be because of short response time by drivers and higher kinetic energy on impact, which are factors that amplify the severity of road vehicle driver incidents at level crossings of roads and rails. Additionally, similar to the finding from the XGBoost algorithm, a wrong position of road user may also be attributed to poor visibility and a higher

probability of coming into contact with the train; this also contributes to higher risks of road vehicle driver incidents at level crossings of roads and rails. The variable of private or public crossing, as an uncontrollable factor, has a significant effect on the severity of road vehicle driver incidents at level crossings of roads and rails.

Table 3: Statistical analysis of the effects of the various factors on the level of incidents in road and rail crossing

Parameter	β	P-value	Effect Size (Exp(β))	sig	Importance Level
public/Private Passcode	0.637	0.002	1.891	00.000	Very High
Estimated Vehicle Speed	0.462	0.005	1.587	00.007	Very High
Road User	-0.321	0.03	0.726	00.000	High
Involved Equipment	0.287	0.04	1.332	00.664	Low
Temperature	-0.187	0.07	0.829	00.077	Medium
Visibility Condition	-0.153	0.08	0.858	00.000	Low
Weather Conditions	0.112	0.1	1.118	00.000	Medium
Track Type	0/089	0.12	1.093	00.000	Low
Track Class	-0.075	0.15	0.927	00.002	Low
Number of Carriages	0.061	0.17	1.063	00.633	Low
Train Speed	0.047	0.2	1.048	00.000	Low
Location of Intersection Warning	0.033	0.23	1.034	00.649	Low
Age of Road Vehicle Driver	-0.019	0.26	0.981	00.000	Low
Visibility Obstructions	0.005	0.3	1.005	00.771	Low

In summary, the multinomial logistic regression model identifies the estimated speed of the vehicle and road user position as influential variables. The XGBoost model highlights visibility obstacles, road user position, and the number of cars. In contrast, the Random Forest model emphasizes train speed and the estimated speed of the vehicle as significant contributors. All these factors impact the severity of road vehicle driver incidents at level crossings of roads and rails.

4. Discussion

In this study, similar to the study by Ghomi et al. in 2016 [9] and another study by the same authors in 2017 [10], train speed has been identified as one of the influential factors in incident severity at level crossings of roads and rails. Zhang et al. 2016 reached a similar conclusion using U.S. incident data [12]. Seven years prior to them, Richard et al. also concluded with similar data, that train speed is one of the influential factors in incident severity at level crossings of roads and rails [20]. Fan et al., 2016, in addition to train speed, also considered the passing vehicle's speed as influential, similar to the results of the current study [11]. The results of these analyses are compared with a similar study conducted by Ghomi et al. (2016) [8] in Table 4. The current study identified the following controllable influencing factors: train speed, estimated vehicle speed, road user position, number of cars, and visual obstacles. Unlike Ghomi et al., the authors did not differentiate between controllable and non-controllable influencing factors; therefore, in order to compare the results, the list of non-controllable influencing factors that affect the current study is also provided, such as the age of the road vehicle driver, temperature, and type of crossing, whether private or public. The following conclusions can be made When comparing the current study's findings with those of Ghomi et al. (2016).

According to Table 4, Both studies highlight factors including train speed, vehicle type, driver age, and prevailing conditions. These consistencies suggest that these factors are always found to be significant predictors of the severity of road vehicle driver incidents at level crossings of roads and rails in different analyses. In the current study, machine learning algorithms, including the XGBoost and Random Forest, have enabled researchers to determine other intricate and accurate interaction effects between various factors. Other conventional techniques like the ordered probit model and classification and regression tree have also been able to identify important factors well, but in terms of accuracy and identification of more intricate interaction, the performance is not as good.

Both studies indicate that in addition to vehicle and driver characteristics, environmental conditions such as temperature, weather, lighting, and obstacles also significantly impact the severity of road vehicle driver incidents. While factors such as time of incident, type of road vehicle, position of road vehicle, type of incident, driver gender, and lighting were identified as influential characteristics in the Ghomi et al. study, they did not prove to be significant in the current study. In the present study, these features were either not included in the model due to their high correlation with other input variables that provide similar information or were indirectly covered through other variables. For instance, visibility status and weather condition variables indirectly accounted for incident time and lighting. Similarly, the model did not include vehicle type and position due to their correlation with estimated speed and road user variables. The type of incident was also excluded from the model due to its high correlation with vehicle type and number of cars. Additionally, the uniform distribution of gender impact in the data excluded driver gender from the model. Given the higher accuracy of the XGBoost algorithm in this study, it can be concluded that the use of advanced machine learning algorithms can contribute to improved prediction and analysis of factors affecting the severity of road vehicle driver incidents, thereby aiding in the formulation of more effective policies and strategies for reducing incidents and injuries.

Table 4: Comparison of the current study with the study by Ghomi et al. [10]

		Current Study	Previous Study (Similar Study)
Data Source		Federal Railroad Administration	Federal Railroad Administration
Data Timeframe		2010 to 2022	2006 to 2013
Main Study Topic		Identifying influential factors in the severity of road vehicle driver incidents on grade crossings road and rail	Identifying factors affecting the severity of driver injuries in road and rail grade crossing incidents
Study Environment		Grade crossings road and rail	Grade crossings road and rail
Modeling Approach		Machine learning algorithms (Random Forest and Extreme Gradient Boosting) and multinomial logistic regression model	Ordered probit model, Classification and Regression Trees (CART), Association Rules
Key Identified Variables	Train speed	✓	✓
	Estimated speed of passing vehicle	✓	-
	Number of cars	✓	-
	Age of road vehicle driver	✓	✓
	Temperature and weather conditions	✓	✓
	Visibility obstacle	✓	✓
	Type of Track	✓	-
	Public or private nature of crossing	✓	-
	Road condition	✓	-
	Time of incident	-	✓
	Type of road vehicle	-	✓
	Position of road vehicle	-	✓
	Type of incident	-	✓
	Driver's gender	-	✓
	Lighting	-	✓

To address the severity of road vehicle driver incidents at level crossings of roads and rails, several recommendations can be implemented. One effective strategy is to lower train speed by installing speed restriction signs, issuing speed reduction instructions to train drivers, and imposing heavy fines for noncompliance. Additionally, creating monitoring systems such as speed cameras and radars [20], along with conducting awareness campaigns and educational workshops, can help control passing vehicle speed. Better management of cars numbers, such as using shorter trains and more precise scheduling, can also reduce incidents.

Removing physical obstacles and creating open visibility areas, as well as improving lighting with smart systems, can enhance visibility. Installing advanced warning systems and optimizing traffic route design can improve road user position, while regular barrier monitoring and testing ensure proper barrier function. Implementing these measures can significantly reduce the severity of road vehicle driver incidents at these crossings.

5. Conclusions

In this study, the multinomial logistic regression model examined the effects of the various factors on incidents and established that two factors, the estimated speed of public vehicles and road user position, which are controllable factors, significantly influence the severity of road vehicle driver incidents at level crossings of roads and rails. In particular, the rise in the speed of road vehicles causes the level of incidents to rise. In the second modeling approach involving more complex machine learning algorithms, the Random Forest algorithm was used to demonstrate that different variables have different levels of significance in determining the severity of an incident. This model's most important controllable variables are the number of cars, the train's speed, and the vehicle's estimated speed. The XGBoost algorithm shows that the controllable variables, such as visual obstacles and the position of the road user, are more influential for the model. Based on these findings, it can be concluded that the XGBoost algorithm assigns the highest weights to the infrastructural and environmental factors, which indicate their high influence on incident severity. Thus, evaluating these three models, it can be stated that all of them accent different variables and are characterized by different effects on the severity of incidents at intersections of roads and rails. The multinomial logistic regression analysis results reveal that the nature of drivers' factors or behavior, which is directly manageable determines the level of road vehicle driver incidents at level crossings of roads and rails. On the other hand, the Random Forest algorithm is more focused on variables of concern with respect to vehicles and the training of the importance of these factors on the extent of road vehicle driver incidents at level crossings of roads and rails. Conversely, the XGBoost algorithm demonstrates that infrastructural and environmental factors significantly influence the severity of road vehicle driver incidents at level crossings of roads and rails and should be paid more attention to in the safety plan. In conclusion, this comparison shows that each model can be applied individually or in conjunction with other models for assessing and forecasting the severity of road vehicle driver incidents at the level crossings of roads and rails. The multinomial logistic regression model alone is used for simple and accurate interpretation of the variables; the Random Forest algorithm and XGBoost algorithm are used individually for accurate and complex data prediction. However, for a comprehensive and complete analysis, statistical models and machine learning models should be used complementarily, as statistical models provide accurate and easily interpretable results. At the same time, machine learning models are more accurate in identifying intricate patterns and also handle the variability of incident data. Thus, selecting an appropriate model depends on the data type, the analysis goals, and the decision-making requirements. Other models including XGBoost and Random Forest can be used for more complex predictions with higher accuracy. In contrast, the multinomial logistic regression model can be useful for more precise and simpler explanations and interpretations. However, based on the findings of this study, the XGBoost algorithm is more accurate and has a higher recall rate than the other algorithms. On the other hand, the Random Forest algorithm was less accurate than the XGBoost algorithm. Thus, comparing these two algorithms proves that their performance is quite similar but different in some aspects.

The method of presenting results was such that first, a general and initial evaluation of the effect of the variables was done using the Random Forest technique. Next, the second approach, which is more detailed and accurate, was carried out using the XGBoost algorithm. Finally, the quantitative analysis of the significance of the variables was presented with the help of the findings of the multinomial logistic regression model. This order also demonstrated how different models can be used side by side to analyze and predict the level of incidents. Based on the findings of the study and recognizing the fact that it is possible to determine the controllable parameters that have a higher

impact on the level crossing of roads and rails, various suggestions for reducing the intensity of road vehicle driver incidents were made at level crossing of roads and rails. Some of the mitigation measures include; reducing the speed of the train, reducing the speed of the passing through vehicle, restricting the number of coaches, reducing the number of barriers to vision, and positioning of the road user.

Acknowledgements

The author acknowledges the foundation support from the members of the Transportation Systems and logistics (TSL) lab which is located in the School of Railway Engineering at IUST.

References

[1] M. U. T. Arshad, "4 people dead after train crashes into pickup at Idaho railroad crossing, police say," USA Today, Apr. 14, 2024. [Online]. Available: https://www.usatoday.com/story/news/nation/2024/04/14/idaho-train-crash/73319284007/.

[2] National Safety Council, "Railroad Deaths and Injuries," NSC Injury Facts. [Online]. Available: https://injuryfacts.nsc.org/home-and-community/safety-topics/railroad-deaths-and-injuries/.

[3] U.S. Department of Transportation, "Accident Data, Reporting, and Investigations," Railroads.dot.gov. [Online]. Available: https://railroads.dot.gov/railroad-safety/accident-data-reporting-and-investigations.

[4] C. Ma, W. Hao, W. Xiang, and W. Yan, "The impact of aggressive driving behavior on driver-injury severity at highway-rail grade crossings accidents," J. Adv. Transp., vol. 2018, pp. 1–10, Oct. 2018, doi: 10.1155/2018/9841498.

[5] National Safety Council, "Every four hours someone is hit by a train," National Rail Safety Week.

[6] Federal Railroad Administration (FRA), "Highway-rail grade crossing accident/incident form F6180.57."

[7] Transportation Safety Board of Canada, "Rail transportation occurrences in 2020," BST-TSB.gc.ca. [Online]. Available: https://www.bst-tsb.gc.ca/eng/stats/rail/2020/sser-ssro-2020.html.

[8] N. Eluru, M. Bagheri, L. F. Miranda-Moreno, and L. Fu, "A latent class modeling approach for identifying vehicle driver injury severity factors at highway-railway crossings," Accid. Anal. Prev., vol. 47, pp. 119–127, Jul. 2012, doi: 10.1016/j.aap.2012.01.027.

[9] H. Ghomi, M. Bagheri, L. Fu, and L. F. Miranda-Moreno, "Analyzing injury severity factors at highway-railway grade crossing accidents involving vulnerable road users: A comparative study," Traffic Inj. Prev., vol. 17, no. 8, pp. 833–841, Nov. 2016, doi: 10.1080/15389588.2016.1151011.

[10] H. Ghomi, L. Fu, M. Bagheri, and L. F. Miranda-Moreno, "Identifying vehicle driver injury severity factors at highway-railway grade crossings using data mining algorithms," in Proc. 2017 4th Int. Conf. Transp. Inf. Safety (ICTIS), IEEE, Aug. 2017, pp. 1054–1059, doi: 10.1109/ICTIS.2017.8047900.

[11] E. Fan, W. Gong, L. Haile, "Severity analysis of vehicle crashes on highway-rail grade crossings: ordered response logit modeling," Adv. Transp. Stud., 2016.

[12] Z. Zheng, P. Lu, and D. Tolliver, "Decision tree approach to accident prediction for highway-rail grade crossings: empirical analysis," Transp. Res. Rec. J. Transp. Res. Board, vol. 2545, no. 1, pp. 115–122, Jan. 2016, doi: 10.3141/2545-12.

[13] S. Soleimani, M. Leitner, and J. Codjoe, "Applying machine learning, text mining, and spatial analysis techniques to develop a highway-railroad grade crossing consolidation model," Accid. Anal. Prev., vol. 152, p. 105985, Mar. 2021, doi: 10.1016/j.aap.2021.105985.

[14] P. Rana, F. Sattari, L. Lefsrud, and M. Hendry, "Machine learning approach to enhance highway-railroad grade crossing safety by analyzing crash data and identifying hotspot crash locations," Transp. Res. Rec. J. Transp. Res. Board, Dec. 2023, doi: 10.1177/03611981231212162.

[15] "Multinomial Logistic Regression | R Data Analysis Examples," UCLA Advanced Research Computing Statistical Methods and Data Analytics.

[16] ArXiv, “Feature interactions in XGBoost,” [Online]. Available: https://arxiv.org/abs/2007.05758.

[17] A. Hschcham, “XGBoost: Everything you need to know,” [Online]. Available: https://neptune.ai/blog/xgboost-everything-you-need-to-know.

[18] A. Hapfelmeier, T. Hothorn, K. Ulm, and C. Strobl, “A new variable importance measure for random forests with missing data,” Stat. Comput., vol. 24, no. 1, pp. 21–34, Jan. 2014, doi: 10.1007/s11222-012-9349-1.

[19] A. S. Hakkert and V. Gitelman, “Development of evaluation tools for road-rail crossing consideration for grade separation,” Transp. Res. Rec. J. Transp. Res. Board, vol. 1605, no. 1, pp. 96–105, Jan. 1997, doi: 10.3141/1605-12.

[20] R. A. Raub, “Examination of highway-rail grade crossing collisions nationally from 1998 to 2007,” Transp. Res. Rec. J. Transp. Res. Board, vol. 2122, no. 1, pp. 63–71, Jan. 2009, doi: 10.3141/2122-08.Fracture of Engineering Materials & Structures, Vol.1, No.3, (2002), pp.899-909.