RESEARCH ARTICLE

REVISED Accident severity prediction modeling for road safety using random forest algorithm: an analysis of Indian highways [version 2; peer review: 1 approved, 2 approved with reservations]

Humera Khanum[1,2], Anshul Garg[1], Mir Iqbal Faheem[3]

[1]School of Civil Engineering, Lovely Professional University, Phagwara, Punjab, 1444411, India
[2]Civil Engineering Department, Symbiosis Institute of Technology, Pune Campus, Symbiosis International (Deemed University), Pune, Maharashtra, 412115, India
[3]Civil Engineering Department, Deccan College of Engineering and Technology, Hyderabad, Telangana, 500001, India

v2 **First published:** 15 May 2023, **12**:494
https://doi.org/10.12688/f1000research.133594.1

Latest published: 20 Oct 2023, **12**:494
https://doi.org/10.12688/f1000research.133594.2

Abstract

Background: Road accidents claim around 1.35 million lives annually, with countries like India facing a significant impact. In 2019, India reported 449,002 road accidents, causing 151,113 deaths and 451,361 injuries. Accident severity modeling helps understand contributing factors and develop preventive strategies. AI models, such as random forest, offer adaptability and higher predictive accuracy compared to traditional statistical models. This study aims to develop a predictive model for traffic accident severity on Indian highways using the random forest algorithm.

Methods: A multi-step methodology was employed, involving data collection and preparation, feature selection, training a random forest model, tuning parameters, and evaluating the model using accuracy and F1 score. Data sources included MoRTH and NHAI.

Results: The classification model had hyperparameters 'max depth': 10, 'max features': 'sqrt', and 'n estimators': 100. The model achieved an overall accuracy of 67% and a weighted average F1-score of 0.64 on the training set, with a macro average F1-score of 0.53. Using grid search, a random forest Classifier was fitted with optimal parameters, resulting in 41.47% accuracy on test data.

Conclusions: The random forest classifier model predicted traffic accident severity with 67% accuracy on the training set and 41.47% on

Open Peer Review

Approval Status ✓ ? ?

	1	2	3
version 2 (revision) 20 Oct 2023	✓ view		? view
version 1 15 May 2023	? view	? view	

1. **Hamsa Zubaidi**, University of Al-Qadisiyah, Diwaniyah, Iraq
2. **Sanjeev Sinha**, National Institute of Technology Patna, Patna, India
3. **Holman Ospina-Mateus**, Universidad Tecnologica de Bolivar, Cartagena, Colombia

Any reports and responses or comments on the article can be found at the end of the article.

the test set, suggesting possible bias or imbalance in the dataset. No clear patterns were found between the day of the week and accident occurrence or severity. Performance can be improved by addressing dataset imbalance and refining model hyperparameters. The model often underestimated accident severity, highlighting the influence of external factors. Adopting a sophisticated data recording system in line with MoRTH and IRC guidelines and integrating machine learning techniques can enhance road safety modeling, decision-making, and accident prevention efforts.

Keywords
Traffic Accidents, Accident Severity, Road Safety, Accident Prediction Modeling, Random Forest

Corresponding authors: Humera Khanum (humera.khanum@sitpune.edu.in), Anshul Garg (anshul.18374@lpu.co.in)

Author roles: Khanum H: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Garg A**: Conceptualization, Supervision, Writing – Review & Editing; **Faheem MI**: Conceptualization, Supervision, Validation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

How to cite this article: Khanum H, Garg A and Faheem MI. **Accident severity prediction modeling for road safety using random forest algorithm: an analysis of Indian highways [version 2; peer review: 1 approved, 2 approved with reservations]** F1000Research 2023, **12**:494 https://doi.org/10.12688/f1000research.133594.2

First published: 15 May 2023, **12**:494 https://doi.org/10.12688/f1000research.133594.1

REVISED **Amendments from Version 1**

We express our gratitude to the reviewers for their constructive comments that greatly enhanced the manuscript's quality. In response to Reviewer-1's insightful observation on the term "accident", we have clarified the context and regional usage of the term in the paper, while also acknowledging the global standards. As per suggestions, we've incorporated the Random Forest algorithm formulation in the methods section and discussed the significance of the Gini impurity test in understanding the importance of explanatory variables.
Addressing Reviewer-2's feedback, we've emphasized the motivation behind our work and its academic contribution in the Introduction section, ensuring a comprehensive understanding of our study's significance. We revisited and corrected specific literature citations, particularly references 7, 14, and 15. The choice of factors, their contribution, and the transferability of the developed model are now prominently highlighted in a dedicated subsection. Further, to address the concerns about the novelty, we've incorporated a detailed paragraph in the conclusion, elucidating the innovative aspects and unique contributions of our predictive model in comparison to existing literature.
We believe these revisions have amplified the clarity, depth, and significance of our research, ensuring alignment with the journal's standards and objectives.

Any further responses from the reviewers can be found at the end of the article

Introduction

Road accidents are a significant public health concern worldwide, with an estimated 1.35 million deaths caused by road traffic accidents each year.[1] Developing countries, such as India, are disproportionately affected, with over 150,000 fatalities reported annually.[1] Road safety is a major concern in India, with a large number of accidents and fatalities reported each year. According to the Ministry of Road Transport and Highways, there were 449,002 road accidents in India in 2019, resulting in 151,113 deaths and 451,361 injuries.[2]

Road accidents pose a threat to health, leading to approximately 1.35 million deaths globally each year. Countries such as India are particularly affected, experiencing over 150,000 fatalities annually. This alarming situation emphasizes the importance of comprehending the factors that contribute to these accidents in order to develop prevention measures. Our research is motivated by the urgency to improve road safety in high-risk areas, like India, where the number of accidents and fatalities remains distressingly high.

Our research delves into the modeling of accident severity, a statistical technique in the field of road safety. Traditional statistical models like logit and probit have been used since the 1990s to predict traffic accident severity, but they have limitations, especially when their underlying assumptions are violated.[3,4] On the other hand, Artificial Intelligence (AI) models offer adaptability and can handle complex nonlinear relationships without being constrained by such assumptions. We focus on the Random Forest (RF) algorithm, an advanced machine learning model with distinct advantages over other algorithms for predicting accident severity.[5] Through tuning parameters and preprocessing techniques, we aim to improve the performance of RF even further, making our academic contribution significant in the pursuit of more accurate and reliable predictions for accident severity.

The modelling process involves analyzing data on past accidents and identifying the factors that contributed to their occurrence and severity.[6] These factors can include road conditions, weather, driver behaviour, and vehicle type, among others. The goal of accident severity modelling is to identify the most important factors contributing to accidents and to develop evidence-based strategies to improve road safety and reduce the number and severity of accidents.[7–9]

Statistical models have been widely used for predicting traffic accidents' severity.[3,4] Artificial intelligence models, in contrast, do not make any assumptions and are more adaptable. They can handle intricate nonlinear relationships and generally offer higher predictive accuracy than statistical approaches.[5] RF, in particular, has been successfully applied in various contexts, and its performance can be further improved by tuning key parameters and careful data preprocessing.[4,5,10,11] The performance of the RF algorithm is significantly influenced by the selection of hyperparameters.[12] To optimize its performance, identifying the optimal parameter values is crucial.

The main objective of our study is to develop a predictive model for the severity of traffic accidents on Indian highways using RF models due to their accuracy and interpretability.

The findings of our study will be used to develop a predictive model for accident severity that can inform road safety policies and interventions. This model can be used to identify high-risk areas and prioritize resources for accident prevention and mitigation.

Study Areas

The study areas selected are the National Highways two stretches as mentioned below

1. Pune-Sholapur Section of NH-9 in km.144/400 to Km. 249/000 in the State of Maharashtra (Figure 1).

2. Six-Laning of Barwa-Adda-Panagarh Section of NH-2 from km 398.240 to km 521.120 including Panagarh Bypass in the States of Jharkhand and West Bengal (Figure 2).

The study areas for this research project were selected based on specific criteria. Firstly, the researchers had prior experience of working on one of the stretches, which is the Pune-Sholapur Section of NH-9 in km.144/400 to Km. 249/000 in the State of Maharashtra. This experience could have provided insights and knowledge that could be useful in conducting the study.

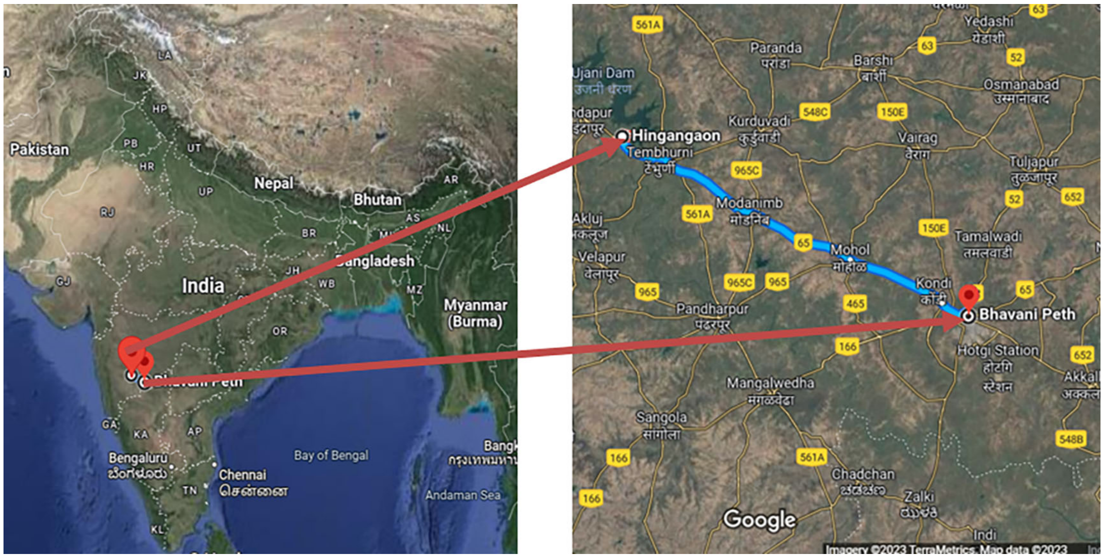

Figure 1. **Pune-Sholapur Section of NH-9 in km.144/400 to Km. 249/000 in the State of Maharashtra.**

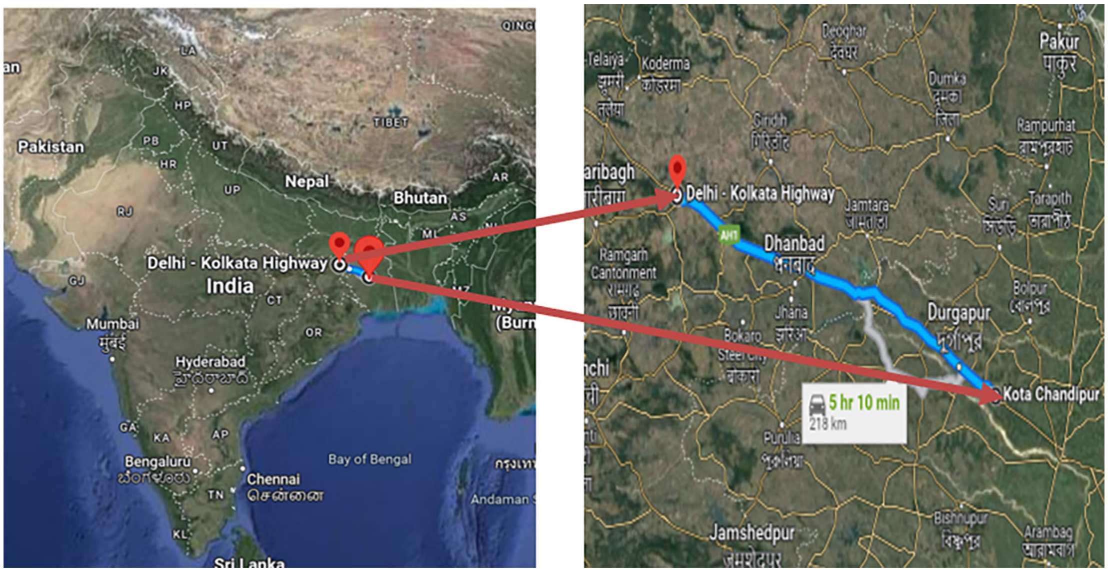

Figure 2. **Six-Laning of Barwa-Adda-Panagarh Section of NH-2 from km 398.240 to km 521.120 including Panagarh Bypass in the States of Jharkhand and West Bengal.**

Additionally, data was also provided by the same concessionaire as of the previous stretch on request for another stretch, which is the Six-Laning of Barwa-Adda-Panagarh Section of NH-2 from km 398.240 to km 521.120 including Panagarh Bypass in the States of and West Bengal. This data could have been relevant to the research objectives and could have assisted in achieving the desired outcomes.

Methods

The proposed methodology for this research involves the following steps for implementing a RF model machine learning technique for accident severity prediction.

Data Preparation: The first step in implementing a RF model for accident severity prediction is to collect and prepare data. Raw data on road accidents for the selected stretches of the highway can be obtained from secondary sources such as the Ministry of Road Transport and Highways (MoRTH) and the National Highways Authority of India (NHAI).[2] Data wrangling and mining techniques can be used to clean and preprocess the data.

Feature Selection: Once the data is prepared, selecting appropriate features for the model becomes crucial. Feature selection plays a vital role in reducing the dimensionality of the data and enhancing the model's accuracy. There are several techniques available for feature selection, such as statistical tests, correlation analysis, and principal component analysis (PCA).[13]

Model Training: In the next step, an RF model can be trained on the preprocessed data. The model can be developed using a machine learning-based framework, as described in Breiman's work on RF.[14] The RF algorithm involves bagging and random feature selection techniques to create multiple decision trees that are aggregated to form a stronger learner.[15]

RF Algorithm Formulation: The RF algorithm can be represented as:

$$[\mathrm{RF(X)} = (1/\mathrm{B}) * \Sigma(\mathrm{T}_{\mathrm{b(X)}}) \text{ from } \mathrm{b} = 1 \text{ to } \mathrm{B}]$$

where X are the input features, B is the number of trees, and $\mathrm{T}_{\mathrm{b(X)}}$ is the prediction of the b-th individual decision tree.

Parameter Tuning: To improve the performance of a RF model, it is important to fine-tune its parameters. The three key parameters that significantly impact the tuning performance of the RF model are the total number of trees (n_estimators), the number of features used for each node segmentation (max_feature), and the maximum depth of a tree (max_depth).[16]

In the construction of the RF model for predicting accident severity, Gini impurity is employed as a criterion to evaluate the significance of different explanatory variables. Gini impurity, a measure utilized within the framework of decision trees (the base learners in a RF), is crucial for the optimal selection of features at each node split. It offers a quantitative metric to discern the effectiveness of a variable in segregating the target classes.

Mechanism of Gini Impurity: In the context of binary classification, the Gini impurity for a node is calculated as:

$$I_G(p) = \sum_{k=1}^{n} p^2 k$$

where, pk is the proportion of samples classified to class k at that node, and the summation operates over all classes. A lower Gini impurity score suggests a higher purity of the node, indicating an enhanced classification.

Gini Importance in RF: In the developed RF model, the Gini impurity plays a dual role: **Node Splitting**: It aids in the identification of the most significant variable at each node by evaluating the potential reduction in impurity for each split, and **Feature Importance**: Post model training, the average decrease in impurity caused by each feature across all trees is computed, known as Gini importance. This metric offers insights into the relative significance of different features for the prediction task.

Model Evaluation: After training the RF model and optimizing its parameters, it is important to evaluate the model's performance. Various evaluation metrics can be used, including accuracy, precision, recall, F1 score, and Area Under the Curve - Receiver Operating Characteristics (AUC-ROC) curve.[17]

Model Implementation: Once the model has been trained and evaluated, it can be deployed for accident severity prediction. The methodology can be designed using Python for building the model and forecasting the severity of road traffic accidents on Indian highways.

Source data

Data on road accidents from selected stretches of highways was obtained from the Concessionaires of the National Highways Authority of India (NHAI) for two projects: Pune-Solapur and Bengal (BAEL) Section. For the Pune-Solapur Section of NH-9, which is located between km. 144/400 and km. 249/000 in the state of Maharashtra, accident dates from 2013 to 2018 were used. For the Six-Laning of Barwa-Adda-Panagarh Section of NH-2, which includes Panagarh Bypass and is located in the States of Jharkhand and West Bengal Stretch, accident dates from 2015 to 2019 were used for the stretch between km 398.240 and km 521.120. The raw data was subject to exploratory data analysis, as detailed in the following section.

Data Preparation

In this stage, data gathering and exploration is performed using secondary source data. The dataset consists of 3257 observations out of which the 1855 observations are of Bengal (BAEL) Section and 1402 observations are of Pune-Solapur and 32 variables, including the target variable “accident severity.” The 32 attributes and their corresponding mappings are presented in Table 1.

Table 1. Dataset Attributes and Parameters Mapping.

Sl No	Attributes	Mapping
1	Accident Index	
2	Date	
3	Day of Week	1-Sunday, 2-Monday, 3-Tuesday, 4-Wednesday, 5- Thursday, 6-Friday, 7-Saturday
4	Time of Accident, Accident Location-A	1-Urban, 2-Rural, 3-Unallocated
5	Accident Location-A Chainage-km	
6	Accident Location-A Chainage-km-RoadSide	LHS, RHS
7-9	Nature of Accident-B1, Nature of Accident- B2, Nature of Accident-B3	1-Overturning, 2-Head on collision, 3-Rear End Collision, 4-Collision Brush/Side Wipe, 5-Right Turn Collision, 6- Skidding, 7a-Others-Hit Cyclist, 7b-Others-Hit Pedestrian, 7c-Others-Hit Parked Vehicle, 7d-Others-Hit Fixed Object, 7e-Others-Wrong Side Driving, 7f-Others-Hit Animal, 7g- Others-Hit Two Wheeler, 7h-Others-Unknown, 7i-Others- Fallen down, 8-Overtaking vehicle, 9-Left Turn Collision
10	Accident Severity -C	1-Fatal, 2-Grevious Injury, 3-Minor Injury, 4-No Injury
11-13	Classification of Accident-C1, Classification of Accident-C2, Classification of Accident-C3	1-Fatal, 2-Grevious Injury, 3-Minor Injury, 4-Non - Injury (Damage only)
14-18	Causes-D1, Causes-D2, Causes-D3, Causes- D4, Causes-D5	1-Drunken, 2-Overspeeding, 3-Vehicle out of control, 4a- Fault of driver of motor vehicle, 4b-Driver of other vehicle, 4c-Cyclist, 4d-Pedestrian, 4e-Passenger, 4f-Animal, 5a- Defect in mechanical condition of motor vehicle, 5b-Road condition
19	Road Feature-E	1-Single lane, 2-Two lanes, 3-Three lanes or more without central divider median, 4-Four lanes or more with central divider alongwith carriageway width
20	Road Condition-F	1-Straight Road, 2-Slight Curve, 3-Sharp Curve, 4-Flat Road, 5-Gentle incline, 6-Steep incline 7-Hump, 8-Dip
21	Intersection Type-G	1-T Junction, 2-‘Y Junction, 3-‘Four arm junction, 4- Staggered junction, 5-Roundabout, 6-Uncontrolled junction
22	Weather Conditions-H	1-Fine, 2-Mist/Fog 3-Cloud, 4-Light Rain, 5-Heavy Rain, 6-Hail/sleet, 7- Snow, 8-Strong Wind, 9-Dust Storm 10-Very Hot, 11-Very Cold, 12-Other extraordinary weather condition
23-26	Vehicle Type Involved-J-V1, Vehicle Type Involved-J-V2, Vehicle Type Involved-J-V3, Vehicle Type Involved-J-V4	1-Car/Jeep/Van, 2-SUV, 3-Bus, 4-Mini Bus, 5-Truck, 6- Two Wheeler, 7-Three Wheeler, 8-Cycle, 9-Pedestrian, 10- Tractor, 11-Unknown, 12-Animal, 13-Objects, 14-LCV, 15- MAV

Table 1. *Continued*

Sl No	Attributes	Mapping
27	Number of Vehicles	
28	Number of Casualties-Fatal	
29	Number of Casualties-Grievous Injury	
30	Number of Casualties-Minor Injury	
31	Number of Casualties-Non Injured	
32	Number of Casualties	

Data Modelling

The RF classification algorithm has been employed in this study to forecast the severity of road traffic accidents in India. This section details the procedure for implementing the model, performance evaluation, and discuss the results obtained. The RF algorithm is written using python programming language.

The target variable for the random RF is selected as the'Accident Severity' which has classes as Fatal, Grevious Injury, Minor Injury and No Injury and indexed as 1-Fatal, 2-Grevious Injury, 3-Minor Injury, 4-No Injury.

The dataset is partitioned into training and testing sets with a ratio of 80% and 20%, respectively. The hyperparameters'n_estimators' and'max_depth' are specified, and a grid search is conducted with cross-validation (cv=5) to identify the optimal hyperparameters. The best parameters and scores are obtained. The best estimator is fit on the training data. Predictions are made on the test data and the accuracy of the model is obtained.

The algorithm and programme for Accident Severity Modelling using RF are written in the Python programming language, and the code is made available to the public for further development. The source code can be accessed via the software availability statement.

Accuracy analysis on test data: Three metrics were employed to evaluate the effectiveness of the algorithms: accuracy, precision, and recall. These metrics are defined as follows:

Accuracy: The formula for a metric that measures the proportion of correctly predicted observations to the total number of observations is represented as:

$$(TP+TN)/(TP+TN+FP+FN)$$

Precision is a metric that indicates the ratio of correctly predicted positive observations to the total number of predicted positive observations, and is calculated using the formula:

$$TP/(TP+FP)$$

Recall is a metric that reflects the ratio of correctly predicted positive observations to the total number of actual positive observations, and is determined using the formula:

$$TP/(TP+FN)$$

Result and Discussion

Model Performance

The classification model used three hyperparameters -'max_depth': 10,'max_features':'sqrt', and'n_estimators': 100, and the results generated a confusion matrix for the training set. The matrix indicated the number of correctly and incorrectly classified instances for each class. The classification report provided precision, recall, and f1-score for each class, along with support. The model showed high precision and recall for class 1 but low precision and recall for classes 2, 3, and 4, with an overall accuracy of 67% and a weighted average f1-score of 0.64 on the training set. The macro average f1-score, which assigns equal weight to each class, was 0.53.

The optimal parameters for a RF classifier model were determined through a grid search, with a max depth of 2, n estimators of 5000, and a random state of 0. The model was then applied to the test data, and the predictions were saved in an Excel file called "predicted output3.xls" for further analysis. The accuracy of the model on the test data was determined to be 0.4147, or approximately 41.47%, indicating that it accurately predicted the severity of traffic accidents in about 41.47% of test cases.

Predicted outputs

Comparative analysis of observed and predicted accident severity index against dates

The actual accident severity indices are represented by the observed values, while the predicted values are generated by the RF model using the input features.

The following is a summary (Figure 3) of the comparison between the observed and predicted values:

On dates such as 25-02-2017, 17-04-2017, and 22-04-2017, the RF model accurately predicts the accident severity index.

In a number of instances, the model predicts a lower accident severity index value than the observed value. 18-02-2017, 23-02-2017, and 27-03-2017, for example.

Occasionally, the model overestimates the accident severity index by predicting a higher value than the observed value, as on 24-05-2017 and 20-10-2017.

In general, the model frequently predicts a severity index of 2 for accidents, even when the observed values are distinct. This may indicate a bias in the model, possibly as a result of an imbalance in the training dataset, in which severity index 2 occurs more frequently than other categories.

Comparative analysis of observed and predicted accident severity index against time

Figure 4 displays the date, day of the week, and time of the accident, as well as the observed and predicted accident severity indices. The plotted for the 165 rows of predicted data doesn't fit in A4 sheet hence the data is published and the link is provided in the Tableau graphs visuals availbility [i].

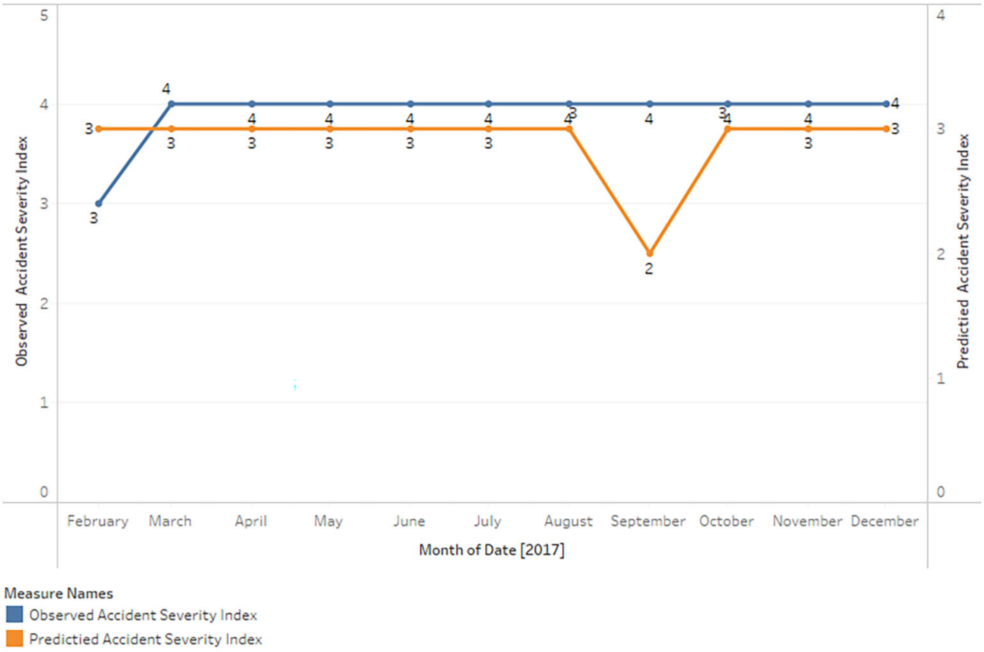

Figure 3. Comparison between Observed and Predicted Accident Severity Index.

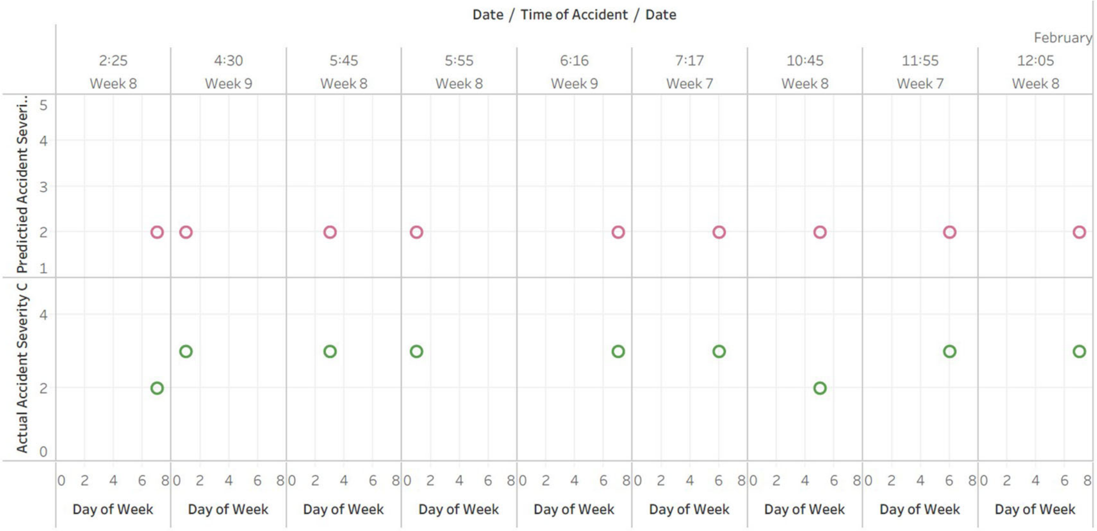

Figure 4. Comparative analysis of observed and predicted accident severity index against time.

The dataset contains accident data from February 18, 2017 to December 31, 2017, as determined by Tableau analysis of the plot generated from the provided Excel table.

The observed accident severity index ranges from 1 to 4, where 1 corresponds to the least severe accident and 4 to the most severe accident.

The observed severity index for the vast majority of accidents in the dataset is 3, followed by 4. 2 indicates a less severe accident, while 4 indicates a more severe accident.

The majority of accidents within the dataset have a predicted severity index of 2, followed by an index of 1.

The analysis of the scatter plot reveals that the predicted severity index is typically lower than the observed severity index. This suggests that the model used to predict the severity of accidents is not always accurate and could be improved.

Comparative analysis of observed and predicted accident severity index against Location and Chainages- RHS

The Tableau plot (Figure 5) presents a detailed visual analysis of accident data on the right-hand side of the road. The data is organized by date and day of the week, displaying the accident location, observed accident severity index, and predicted accident severity index for each incident. The plot effectively illustrates the spatial distribution of accidents and their severity over time, enabling the identification of patterns and trends. The Tableau plot doesn’t fit in A4 sheet hence the data is published and the link is provided in the Tableau graphs visuals availability [ii].

It is evident from the analysis that the majority of accidents have an observed severity index of 2 or 3, indicating a moderate severity. However, the predicted accident severity index largely remains at 2, indicating that the predictions may be somewhat conservative and do not fully capture the observed severity range.

In addition, there appears to be no correlation between the day of the week and the frequency or severity of accidents across the different days of the week. This may suggest that external factors, such as traffic patterns or weather conditions, have a greater impact on the occurrence and severity of accidents than the day of the week.

Comparative analysis of observed and predicted accident severity index against Location and Chainages- LHS

The graph displays (Figure 6) the date, day of the week, and accident location on Left Hand Side (LHS) of the road, as well as the observed and predicted accident severity indices. The plotted of predicted data doesn’t fit in A4 sheet hence the data is published and the link is provided in the Tableau graphs visuals availability [iii].

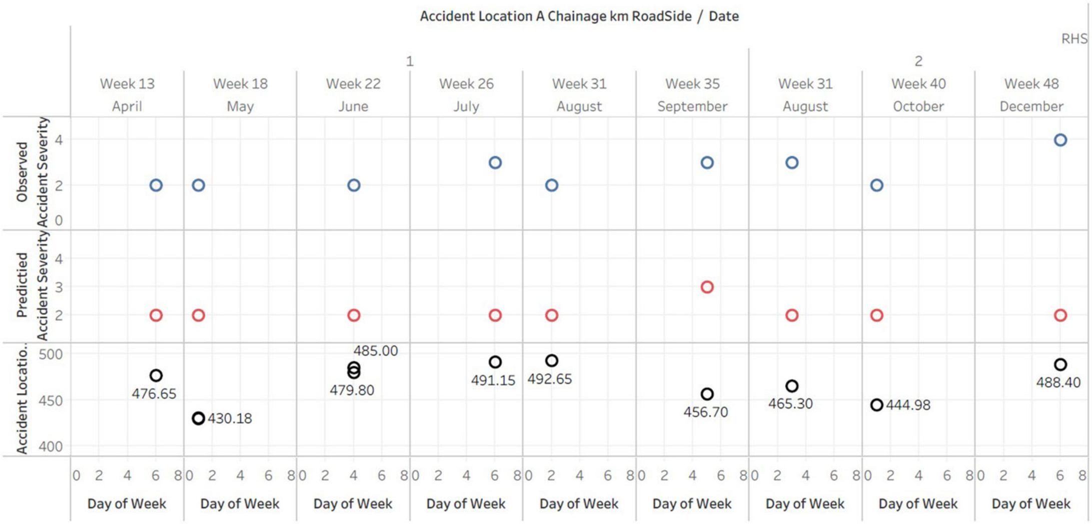

Figure 5. Comparative analysis of observed and predicted accident severity index against Location and Chainages-RHS.

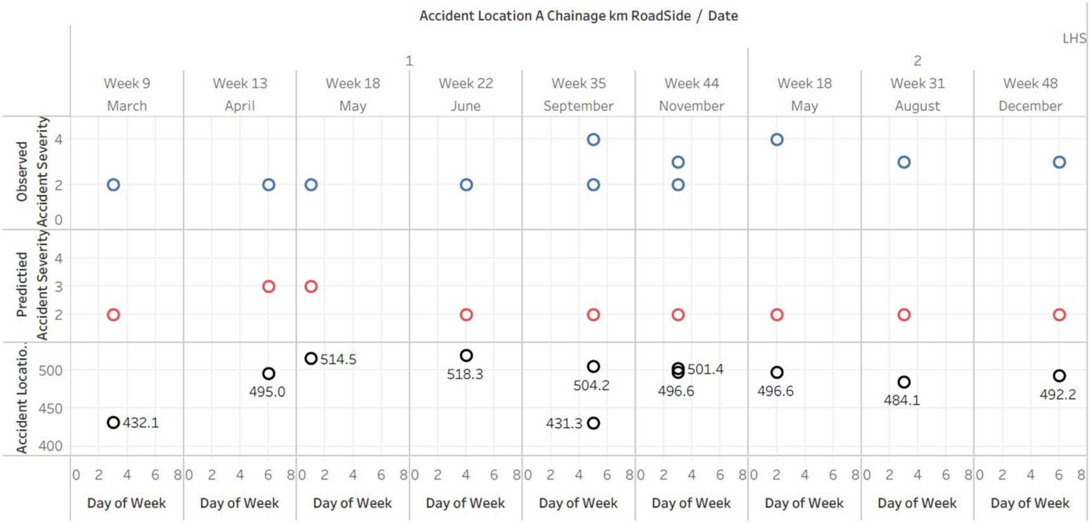

Figure 6. Comparative analysis of observed and predicted accident severity index against Location and Chainages-LHS.

The scatterplot reveals that the majority of accidents on the left side of the road had a severity index of 2 or 3, with only a few instances of severity index 1 and 4. This indicates that the majority of collisions on the left side of the road were of moderate severity.

In the majority of cases, the predicted accident severity index was 2, with only a few instances of values 3 and 4. This suggests that the predictive model may be biased towards predicting less severe accidents.

There was no discernible pattern or trend between the day of the week and the occurrence of accidents. Accidents appeared to occur every day, indicating that the day of the week may not be a significant predictor of accident severity on the left side of the road.

The accident locations, as measured by Accident Location A Chainage km, were scattered along the roadway at various distances. This suggests that there may not be a particular accident hotspot or concentration on the left-hand side of the road.

Data Recording and availability

The recording of road accident data in India must comply with the MoRTH & IRC guidelines, utilizing the Road Accident Recording and Reporting Formats. Despite this, there exists a need for a more advanced data recording system to effectively model road safety. The digital monitoring of road accidents can increase the frequency of data collection and minimize the absence of crucial information. Often, the lack of a system or individual to document the accident leads to the absence of important road accident data. This missing data can be regained through the use of machine learning, thus enhancing the accuracy of road safety modeling.

Conclusion

The RF classifier model predicted the severity of traffic accidents with an overall accuracy of 67% on the training set and approximately 41.47% on the test set. Indicating possible bias or imbalance in the training dataset, the model tended to predict a lower severity index than the observed values. There were no discernible relationships between the day of the week and the occurrence or severity of accidents. The performance of the model can be enhanced by correcting the dataset imbalance and refining the model's hyperparameters.

The observed and predicted accident severity indices were compared against a number of variables, including dates, times, and locations on both sides of the road. In some instances, the model accurately predicted the accident severity index, but it frequently underestimated accident severity. No discernible patterns or trends were observed in terms of accident location, indicating that external factors may have a greater influence on the occurrence and severity of accidents.

To improve road safety modelling, it is essential to adopt a more sophisticated data recording system consistent with MoRTH and IRC recommendations. Digital monitoring of road accidents can increase the frequency of data collection and reduce the loss of vital information. Integrating machine learning techniques can contribute to more effective interventions and decision-making in the field of traffic accident prevention and mitigation.

In the field of accident severity modeling our research stands out for its contributions. We focus on leveraging Artificial Intelligence (AI) models, which excel at capturing the relationships that traditional statistical methods often overlook. Our, in depth study of the Random Forest (RF) algorithm, combined with careful parameter adjustment and data preprocessing highlights its potential in this area. This research addresses not concerns but also specifically tackles India's road safety challenges providing insights applicable worldwide as well as tailored solutions for the region. A key aspect of our approach is our unwavering dedication to improving accuracy positioning our work as a standard for precise and reliable accident severity predictions. Overall this study makes a contribution to literature, in this field.

Future Scope

The study presented provides a good starting point for future research in the field of road safety modeling and accident prevention for Indian highways. However, with the limitations of the present study there opens potential areas for future research as mentioned below which will be taken up in continuation.

Dataset improvement: The study identified the possibility of dataset bias and imbalance affecting model performance. Future research will focus on improving the quality and quantity of data, reducing bias and improving model performance. This will involve exploring alternative data sources, enhancing data collection methods, and addressing data quality issues.

Model improvement: The study used the RF algorithm to develop a predictive model for traffic accident severity. In future research, other machine learning algorithms or ensemble models to improve model performance will be explored. Additionally, refining hyperparameters and addressing dataset imbalance will be done to improve model accuracy.

External factors analysis: The study highlighted the influence of external factors on accident severity prediction. Future research can focus on exploring the impact of external factors such as weather conditions, road infrastructure, and driver behavior on accident severity. This can enhance the accuracy of predictive models and inform decision-making in accident prevention efforts.

Real-time monitoring: The study highlighted the need for a sophisticated data recording system in line with MoRTH and IRC guidelines. Future research can focus on developing a real-time monitoring system that can capture road safety data in real-time and provide insights for accident prevention efforts.

Data availability

Zenodo. Data for Accident Severity Prediction Modelling for Indian Highways Case Study, https://doi.org/10.5281/zenodo.7773156.[18]

This project contains the following underlying data:

- Accdataset_hk_PS_BAEL_Combined.csv (The dataset consists of 3257 observations out of which the 1855 observations are of Bengal (BAEL) Section and 1402 observations are of Pune- Solapur.)

- predicted_output_1.xlsx (This is level-2 processed data derived from raw accident data using prediction modeling. The data has been indexed from 1 to 4 for further analysis, and there are a total of 165 rows in the predicted output observations.

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Software availability

- Github: https://github.com/humera-k/RF_Accident_Severity

- https://zenodo.org/badge/latestdoi/616376786

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Tableau graphs visuals availability

1. Accidental_Analysis_1 | Tableau Public (Comparative analysis of observed and predicted accident severity index against time)

2. Accidental_Analysis_1 (Comparative analysis of observed and predicted accident severity index against Location and Chainages-Right hand Side (RHS))

3. Accidental_Analysis_1 (Comparative analysis of observed and predicted accident severity index against Location and Chainages-Left Hand Side (LHS))

Acknowledgement

We are grateful to National Highways Authority of India and IL&FS Engineering and Construction Company for making the raw accident data available.

References

1. World Health Organization: **Global Status Report on Road Safety 2018.**
Reference Source

2. Ministry of Road Transport & Highways: **Road Accidents in India 2019.**
Reference Source

3. Mohamed AA, Mannering FL: **Probabilistic models of traffic accident occurrence and severity.** *Handbook of traffic psychology.* Elsevier; 2010; (pp. 129–143).
Publisher Full Text

4. Sze NN, Wong SC: **Diagnostic analysis of the logistic model for pedestrian injury severity in traffic crashes.** *Accid. Anal. Prev.* 2007; **39**(6): 1267–1278.
PubMed Abstract | Publisher Full Text

5. Hoang Long V, Ahmed K, Ma W: **A Random Forest Approach to Predicting Traffic Accident Severity.** *IEEE Access.* 2021; **9**: 1219–1232.
Publisher Full Text

6. Mannering F, Bhat C: *Statistical Methods for Analyzing Highway Accident Data.* John Wiley & Sons; 2014.

7. Barbosa P, Andrade M, Ferreira S: **Machine learning applied to road safety modeling: A systematic literature review.** *Journal of Traffic and Transportation Engineering (English Edition).* 2020; **7**(6): 775–790.
Publisher Full Text

8. Al-Mistarehi BW, Alomari AH, Imam R, *et al.*: **Using Machine Learning Models to Forecast Severity Level of Traffic Crashes by R Studio and ArcGIS.** *Front. Built Environ.* 2022; **8**(April): 1–14.
Publisher Full Text

9. Moghaddam FR, Afandizadeh S, Ziyadi M: **Prediction of accident severity using artificial neural networks.** *Int. J. Civ. Eng.* 2011; **9**(1): 41–48.

10. Abdel-Aty M: **Analysis of driver injury severity levels at multiple locations using ordered probit models.** *J. Saf. Res.* 2003; **34**: 597–603.
PubMed Abstract | Publisher Full Text

11. Shiran G, Imaninasab R, Khayamim R: **Crash Severity Analysis of Highways Based on Multinomial Logistic Regression Model, Decision Tree Techniques, and Artificial Neural Network: A Modeling Comparison.** *Sustainability.* 2021; **13**(10): 5670.
Publisher Full Text

12. Yan L, Liao W: **Evolutionary hyperparameter optimization for random forest.** *J. Ambient. Intell. Humaniz. Comput.* 2019; **10**(7): 2801–2810.
Publisher Full Text

13. Adele Cutler DRC, Stevens JR: **Random Forests.** 2012.
Publisher Full Text

14. Breiman L: **Random forests.** *Mach. Learn.* 2001; **45**(1): 5–32.
Publisher Full Text

15. Liaw A, Wiener M: **Classification and regression by randomForest.** *R News.* 2002; **2**(3): 18–22.

16. El-Basyouny K, Sayed T, Abdel-Aty M: **Predicting accident occurrence and severity on arterials using random parameter and random effect models.** *Accid. Anal. Prev.* 2010; **42**(3): 718–727.

17. Sokolova M, Lapalme G: **A systematic analysis of performance measures for classification tasks.** *Inf. Process. Manag.* 2009; **45**(4): 427–437.
Publisher Full Text

18. Khanum H, Garg A, Faheem MI: Data for Accident Severity Prediction Modelling for Indian Highways Case Study (Accidentdata_V1). [Data set]. *Zenodo.* 2023.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✓ ? ?

Version 2

Reviewer Report 13 January 2024

https://doi.org/10.5256/f1000research.157426.r228819

?

Holman Ospina-Mateus
Universidad Tecnologica de Bolivar, Cartagena, Bolívar, Colombia

The article proposes an analysis of road safety in the context of India, using a random forest methodology, as a proposal to predict conditions on highways. Below are some aspects that should be improved within the article:
• In the abstract, it is noted: "AI models", random forest is a traditional machine learning method.
• The results in the abstract: They can be written better.
• Road accident datasets are traditionally unbalanced and may be subject to considering sampling techniques that allow them to be balanced or leveled.
• Before explaining the method, it would be good to see an overview in descriptive statistics of the country's official data on road safety, where the participation by road actors, spatiotemporal conditions, and sociodemographic conditions can be seen.
• Figure four has errors in structure, labels, and legends.
• In the analysis and relationships that the variables may have, there is no significant contribution.
• It is recommended within the model to reduce the categories of the variables, seeking better relational results.
• The ROC curve is not observed.
• No technique was considered to balance accident sampling in relation to predictions.
• The model suffers from a high hyper parameterization; the authors did not validate a pre-analysis that linked or filtered the variables that could be part of the model.
There was a lack of clear discussion at the end of the document, contrasting the results with similar investigations.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Road safety.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 01 November 2023

https://doi.org/10.5256/f1000research.157426.r217364

✓ **Hamsa Zubaidi**
Roads and Transport Engineering Department, University of Al-Qadisiyah, Diwaniyah, Al-Qādisiyyah Governorate, Iraq

The authors addressed my comments, thank you.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Transportation Safety

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 05 October 2023

https://doi.org/10.5256/f1000research.146596.r173654

?

Sanjeev Sinha
Department of Civil Engineering, National Institute of Technology Patna, Patna, Bihar, India

The work is clearly and accurately presented however, the motivation behind the work and its academic contribution needs to be highlighted.

Though relevant literature has been cited, it has not been described for its application in the present study. Further, there is some literature which seems not to be correctly referred and cited. As example the references, 7, 14 and 15 needs to be re-checked.

The title of the work is development of prediction modeling for accident severity, the choice of factors, their contribution and transferability of the model developed needs to be highlighted.

The literature review shows several works related to the topic thus novelty of the work also need to be highlighted.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Traffic Engineering, Transportation Planning, Highway Materials, Logistics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 12 Oct 2023

Humera Khanum

Comment 1: The work is clearly and accurately presented however, the motivation behind the work and its academic contribution needs to be highlighted.
Author's Response: Thank you for your constructive feedback. We agree that highlighting the motivation and academic contribution of this work is crucial for a well-rounded understanding of the study. In the revised manuscript, we have included a section in the Introduction where we elaborate on the motivation behind this work, particularly focusing on the rising number of road accidents and the necessity to develop predictive models to understand and mitigate accident severity. Additionally, we have outlined the academic contribution of our study in the Discussion section, emphasizing the novelty and practical implications of our findings.

Comment 2: Though relevant literature has been cited, it has not been described for its application in the present study. Further, there is some literature which seems not to be correctly referred and cited. As example the references, 7, 14 and 15 needs to be re-checked.
Author's Response: We appreciate your attention to detail. We have revisited the cited literature, especially references 7, 14, and 15, and corrected the citations as suggested. Moreover, we have expanded the section to better explain the relevance and application of the cited works in the context of our study.

Comment 3: The title of the work is Development of prediction modeling for accident severity, the choice of factors, their contribution and transferability of the model developed needs to be highlighted.
Author's Response: Thank you for your suggestion. In the revised manuscript, we have dedicated a subsection within the Methodology section to discuss the choice of factors,

their contribution, and the transferability of the developed model.

Comment 4: The literature review shows several works related to the topic thus novelty of the work also need to be highlighted.
Author's Response: We value your input on highlighting the novelty of our work. In the revised manuscript, we have included a paragraph at the end in the conclusion section where we delineate the novelty of our study in comparison to existing literature. We have emphasized the innovative aspects of our predictive model and how our work contributes uniquely to the understanding and prediction of accident severity.

Competing Interests: NO COMPETING INTERESTS

Reviewer Report 18 September 2023

https://doi.org/10.5256/f1000research.146596.r202786

?

Hamsa Zubaidi
Roads and Transport Engineering Department, University of Al-Qadisiyah, Diwaniyah, Al-Qādisiyyah Governorate, Iraq

1. The author mentioned "accident" in the paper, "Accident" is a broader term that implies an unintentional or unforeseen incident. It can encompass various events, including crashes, collisions, or other incidents resulting from human error, mechanical failure, or environmental factors. "Crash" is a more specific term that typically refers to a collision or impact between vehicles, objects, or individuals. It often implies a sudden and forceful event.

   Ultimately, the choice between "accident" and "crash" should be based on the specific details and nature of the event being described in the paper, and relying on that "accident" should be replaced by "crash".

2. The author should add a random forest formula in the methodology.

3. It might be better if there is a Gini impurity test in the paper to understand the importance of explanatory variables.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Tansportation Safety

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 29 Sep 2023

Humera Khanum

Comment:1 The author mentioned "accident" in the paper, “Accident" is a broader term that implies an unintentional or unforeseen incident. It can encompass various events, including crashes, collisions, or other incidents resulting from human error, mechanical failure, or environmental factors. "Crash" is a more specific term that typically refers to a collision or impact between vehicles, objects, or individuals. It often implies a sudden and forceful event.
Ultimately, the choice between "accident" and "crash" should be based on the specific details and nature of the event being described in the paper, and relying on that "accident" should be replaced by "crash".

RESPONSE: Thank you for your insightful comments and for bringing up the concern regarding the usage of the term "accident" in our paper. Your observation about the broad implication of "accident" and the specificity of "crash" is indeed valid.

I would like to clarify that in the context of our paper, the mention of "accidents" is based on Indian data, which is commonly referred to as road traffic accident data in official and legal documents in India. The term "accident" here encompasses various events, including crashes, collisions, or other incidents resulting from human error, mechanical failure, or environmental factors, as per the definition used in the Indian context.

All the parameters mentioned in the paper have been taken into consideration, and the terminology used is consistent with the data source and regional context. We appreciate the suggestion to use the term "crash" for more specificity and will certainly consider this in our future research and writings, ensuring that the terminology aligns with the global standards and is clear to all readers.

Comment 2: The author should add a random forest formula in the methodology.

RESPONSE: The random forest formula is added to the methods section as below:
Random Forest Algorithm Formulation: The Random Forest algorithm can be represented as:
[RF(X) = (1/B) * Σ(T_b(X)) from b=1 to B]
where X are the input features, B is the number of trees, and T_b(X) is the prediction of the b-th individual decision tree.

Comment 3: It might be better if there is a Gini impurity test in the paper to understand the importance of explanatory variables.

RESPONSE: The random forest formula is added to the methods section as below:
In the construction of the Random Forest model for predicting accident severity, Gini impurity is employed as a criterion to evaluate the significance of different explanatory variables. Gini impurity, a measure utilized within the framework of decision trees (the base learners in a Random Forest), is crucial for the optimal selection of features at each node split. It offers a quantitative metric to discern the effectiveness of a variable in segregating the target classes.
Mechanism of Gini Impurity: In the context of binary classification, the Gini impurity for a node is calculated as:

$I I_{G}(p)=k=1np2k (p) = \sum_{k=1}^{n} p^{2}\mathrm{k}$

where, *pk* is the proportion of samples classified to class *k* at that node, and the summation operates over all classes. A lower Gini impurity score suggests a higher purity of the node, indicating an enhanced classification.

Gini Importance in Random Forest: In the developed Random Forest model, the Gini impurity plays a dual role: **Node Splitting**: It aids in the identification of the most significant variable at each node by evaluating the potential reduction in impurity for each split, and **Feature Importance**: Post model training, the average decrease in impurity caused by each feature across all trees is computed, known as Gini importance. This metric offers insights into the relative significance of different features for the prediction task.

Competing Interests: NO Competing Interests