Transportation Research Interdisciplinary Perspectives 19 (2023) 100814

Contents lists available at ScienceDirect

Transportation Research Interdisciplinary Perspectives

journal homepage: www.sciencedirect.com/journal/transportation-research-interdisciplinary-perspectives

A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance

Shakil Ahmed [a], Md Akbar Hossain [a,*], Sayan Kumar Ray [b], Md Mafijul Islam Bhuiyan [c], Saifur Rahman Sabuj [d]

[a] School of Business and Digital Technologies, Manukau Institute of Technology, Auckland 2104, New Zealand
[b] School of Computer Science, Taylor's University, Selangor, Malaysia
[c] Department of Computational Physics, University of Alberta, Edmonton, Canada
[d] Department of Electrical and Electronic Engineering, Brac University, Bangladesh

ARTICLE INFO

Keyword:
Road accident
Explainable machine learning
SHAP
Feature
Injury severity

ABSTRACT

Road accidents are increasing worldwide and are causing millions of deaths each year. They impose significant financial and economic expenses on society. Existing research has mostly studied road accident prediction as a classification problem, which aims to predict whether a traffic accident may happen in the future or not without exploring the underneath relationships between the complicated factors contributing to road accidents. A number of research have been done to date to explore the importance of road accident contributing factors in relation to road accidents and their severity, however, only a few of those research have explored a subset of ensemble ML models and the New Zealand (NZ) road accident dataset. Therefore, in this paper, we have evaluated a set of machine learning (ML) models to predict road accident severity based on the most recent NZ road accident dataset. We have also analysed the predicted results and applied an explainable ML (XML) technique to evaluate the importance of road accident contributing factors. To predict road accidents with different injury severity, this work has considered different ensembles of ML models, like Random Forest (RF), Decision Jungle (DJ), Adaptive Boosting (AdaBoost), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LGBM), and Categorical Boosting (CatBoost). New Zealand road accident data from 2016 through 2020 obtained from the New Zealand Ministry of Transport is used to perform this study. The comparison results show that RF is the best classifier with 81.45% accuracy, 81.68% precision, 81.42% recall, and 81.04% of F1-Score. Next, we have employed the Shapley value analysis as an XML technique to interpret the RF model performance at global and local levels. While the global level explanation provides the rank of the features' contribution to severity classification, the local one is for exploring the use of features in the model. Furthermore, the Shapley Additive exPlanation (SHAP) dependence plot is used to investigate the relationship and interaction of the features towards the target variable prediction. Based on the findings, it can be said that the road category and number of vehicles involved in an accident significantly impact injury severity. The identified high-ranked features through SHAP analysis are used to retrain the ML models and measure their performance. The result shows 6%, 5%, and 8%, increase, respectively, in the performances of DJ, AdaBoost, and CatBoost models.

1. Introduction

Every day thousands of people are killed and injured on our roads. Men, women or children walking, biking or riding to school or work, playing in the streets or setting out on long trips can become victims of road accidents leaving behind shattered families and communities. Each year, millions of people spend long weeks in hospital after severe crashes, and many lose the ability to live, work, or play normally as they used to do. Sufferings for victims and their families from road traffic-related injuries are incalculable. According to the World Health Organisation (WHO), each year approximately 1.35 million annual road accidents happen seriously injuring between 20 to 50 million people worldwide. It is the eighth leading cause of global death and may become seventh by 2030 if the current trend continues. Table 1 (Transport - road accidents - oecd data, 2021) shows the last ten years' road fatality statistics for the Organisation for Economic Co-operation

* Corresponding author.
E-mail address: akbar.hossain@manukau.ac.nz (M.A. Hossain).

https://doi.org/10.1016/j.trip.2023.100814

Available online 6 April 2023

Table 1
Road accident fatality statistics for last decade (Transport - road accidents - oecd data, 2021).

Country	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Changes from 2019 (%)
Australia	1350	1277	1300	1187	1151	1204	1292	1222	1134	1186	1095	−8.31
Canada	2238	2023	2075	1951	1841	1889	1899	1856	1922	1762	1745	−0.97
Denmark	255	220	167	191	182	178	211	175	171	199	163	−22.09
France	3992	3963	3653	3268	3384	3461	3477	3448	3248	3239	2780	−16.51
Germany	3648	4009	3600	3339	3377	3459	3206	3180	3275	3046	2719	−12.03
Japan	5828	5535	5261	5165	4838	4885	4698	4431	4166	3920	2839	−38.08
New Zealand	375	284	308	253	293	318	327	378	378	352	320	−10.00
United Kingdom	1905	1960	1802	1770	1854	1804	1860	1856	1839	1752	1516	−15.57
United States	32999	32479	33782	32893	32744	35484	37806	37473	36560	36120	38680	6.62

Table 2
Road Safety vision and targets.

Country	Vision	Targets
Australia	Vision Zero	Targets by 2030: Reduced 50% fatalities by 2030 and 30% serious injuries
Canada	Towards Zero- The safest roads in the world	RSS2025: Seeks to achieve directional downward trends in the rate-based number of fatalities and serious injuries
Denmark	Action plan 2021–2030	Maximum fatalities should be 90 or below and maximum serious injury should be 900 or below
France	Eu road safety targets	Reduction of 50% fatalities and serious injury by 2030
Germany	Eu road safety targets	Reduction of 50% fatalities and serious injury by 2030s
Japan	Safest country for road traffic	Fewer than 2300 road crashes by 2032
New Zealand	Vision Zero	40% reduction in annual deaths and serious injuries by 2030
United Kingdom	Safe system approach	Reduction of 50% fatalities and serious injury by 2030
United States	Highest standards of excellence in road safety	Reduction of 50% fatalities and serious injury by 2030

and Development (OECD) countries. It is good to see that most countries achieve a negative trend of road accident fatality except the United States of America (USA). Unfortunately, road accidents have increased slightly in the USA even though the number of miles travelled by car has decreased by 13% from the previous year.

Besides the fatality, socio-economic and financial impact of road crashes cannot be ignored. Taking the example of New Zealand (NZ), from the economic perspective, the impact of road accidents costs almost 3% of the country's gross domestic product which includes health associated expenses, employers' costs, and household costs (M. of Transport, 2020). The value of a statistical life (VSL) is widely used to estimate the local trade-off rate between fatality risk and money. According to the NZ Ministry of Transport (2020), the average social cost per fatality is NZD 4.46 million and VSL is NZD 4.42 million per fatality as of June 2020 (M. of Transport, 2020). Moreover, average social costs for serious injuries and minor injuries are NZD 467,700 and NZD 25,300 per injury, respectively.

Countries worldwide have taken significant measures, like, updating policies and road safety strategies, and adopting technologies to minimise road accidents and their impact. Table 2 provides a summary of the national road safety strategies of OECD countries and targets to achieve. Reducing fatality and serious injury by 50% are the main targets for most countries. Besides the road accident prevention policy and strategy, it is important to accurately understand and analyse the contributing factors of road accidents and their impacts in order to design safer roads. Road accident is a complex phenomenon influenced by various contributing factors. There is no straightforward or linear relationship between injury severity and multiple factors in an accident, which is deemed one of the significant challenges in developing road safety models. Accident prevention models are usually used to monitor the effectiveness of various road safety policies that have been introduced to minimise accident occurrences. They also give transportation planners and/or engineers an idea to determine new policies and strategies for

Table 3
Road accident contributing factors taxonomy.

Road Characteristics	Vehicle	Human Factors	Environment	Speed Limit	Others
Road type	Vehicle type	License condition	Bright sun	Proposed speed limit	Holiday
Number of lanes	Damage area	License type	Overcast	Temporary speed limit	Weekdays
Special purpose lane	Vehicle usage	License status	Twilight	Advisory speed limit	Weekend
Road feature	CC rating	overseas license	Dark		School zone
Rail crossing	Total Passenger front	Fatigue	Rain		Day and time
Junction type	Total Passenger back	Drug	Mist or Fog		Side road
Road marking	Too fast for condition	Alcohol	Snow		On state highway
Road curvature	type of load	Seat-belt used	Hail or Sleet		Open speed zone
Gradient	Load height	Age	Fine		
Surface type	Load secured	Gender	Wind		
Street lights	Safety rating	Ethnicity	Frost		
Barrier type	Warranty/certificate of fitness				
Barrier location	Mental states				
Road category	Permit type	Medical Condition			

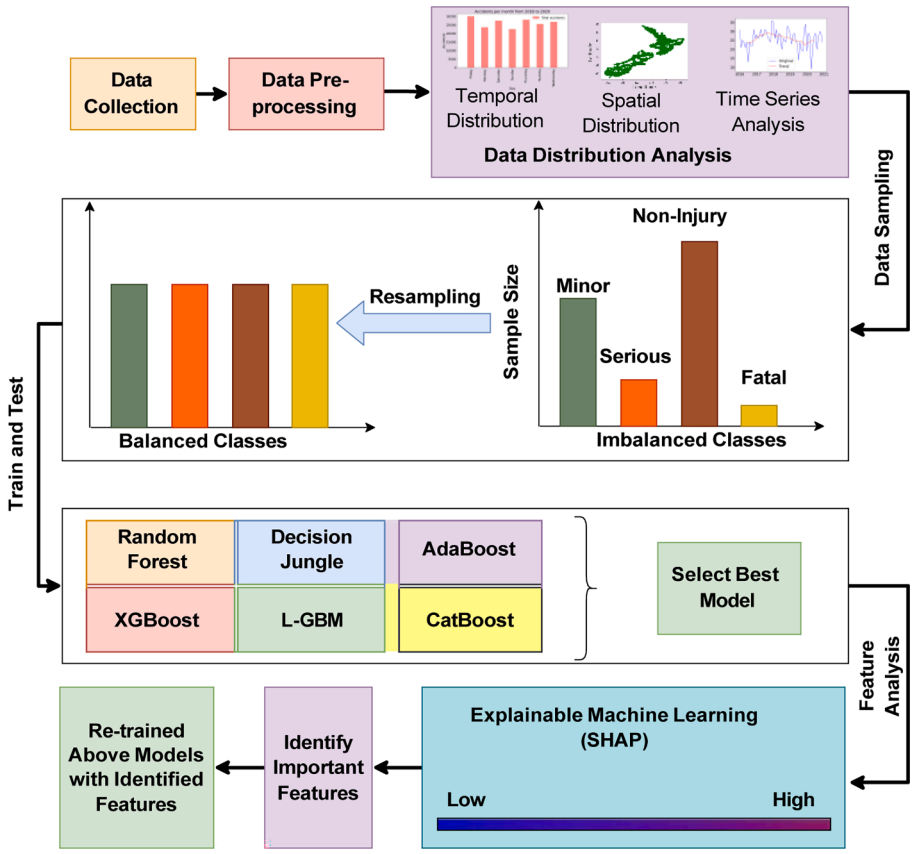


Fig. 1. Feature analysis using machine learning.

road safety. The use of ML approaches to analyse accident data from different circumstances facilitates not only to determine the importance of contributing factors of injury severity but also to select of appropriate input data for developing predictive models. The contributing factors to road accidents can be categorised based on different characteristics such as road, vehicle, human, speed limit, and others. The details taxonomy is presented in Table 3.

Measuring road safety has always been a critical and complicated component of road safety management systems. Road infrastructure safety management, road authorities, road designers, and road safety practitioners can use this road accident prediction model presented in this paper to analyse the potential safety issues, identify safety improvements, and estimate the potential effect of these improvements in terms of crash reduction. This research aims to gain a deep understanding of the critical variables that significantly contribute to road accidents. The research addresses road accident prediction as a classification issue. Several studies have been conducted to predict road accidents and investigate the severity of road accidents, however, very few of them focus on analysing the relationships between road accidents and the factors contributing to those accidents. This study has evaluated a set of ML models to predict road accident severity. Also, the predicted results are analysed and an explainable ML (XML) technique is applied to evaluate the importance of road accident contributing factors. Furthermore, the study investigates the connection between factors contributing to road accidents and accident severities. Table 3 provides a generic set of factors involved in a road accident. We have categorised the contributing factors into six types: road characteristics, vehicle, human factors, environment, speed, and others which is similar to what has been reported in Rolison et al. (2018). In this study, we have used a set of ML algorithms, namely, RF, DJ, ADABoost, XGBoost, L-GBM, and CatBoost. From our previous study, we have seen that ensemble ML algorithms perform better in high dimensional datasets compared to single model ML algorithms (Ahmed et al., 2021). Later we conducted a feature importance analysis to understand how a specific feature will affect the model in predicting the target variable. Further, a feature importance analysis is useful for model improvement and model interpretability. Model interpretability refers to interpreting and communicating the model behaviour and performance to a wider audience. We have used explainable ML methods to understand what features are important and how the features interact with the outcome. Explainable ML helps us to comprehend the model behaviour and interpret the prediction made by the ML algorithm. There are several explainable ML techniques which can be broadly categorised as Ante-hoc and post hoc models. In ante-hoc, the explainability of a model is considered right from the beginning. Reverse time attention (RETAIN) and Bayesian deep learning (BDL) are good examples of ante-hoc models. On the contrary, post hoc models allow models to be trained as normal, and explainability is based on the results of the model's output. Shapley Additive Explanations (SHAP), Local Interpretable Model-agnostic Explanation (LIME), and layer-wise relevance propagation (LRP) are widely used post hoc techniques. SHAP uses the Shapley value, which is the average marginal contribution of a feature value over all possible coalitions. LIME works based on sparse liner models for each prediction to explain how the model works in the local vicinity. Despite the time it takes to compute the Shapley value, SHAP guarantees accuracy and consistency in explainability. Moreover, SHAP has better visualisation features than LIME, making it more popular when communicating with stakeholders from diverse backgrounds. This research uses the SHAP as a post hoc XML technique to interpret the model output and understand the feature contributions toward the target variable. For this, an analysis of the Shapley value is performed for global and local explanations. We have collected and used the New Zealand road accident dataset to train the ML method and analyse the findings to identify the underlying relationships of the road accident's contributing factors. The flow of this

work is illustrated in Fig. 1.

The rest of the paper is organised as follows: Section 2 highlights previous studies related to this research. Section 3 discusses the data collection, preprocessing and descriptive statistics, temporal, and spatial distribution of the dataset considered in this study. Section 4 briefly describes the different ML methods used in this work. Section 5 presents the results for ML models followed by the use of explainable ML for feature analysis. The implication and limitations of this study are discussed in Section in 6 and finally, Section 7 concludes the paper.

2. Background study and literature review

A detailed review of the work done by the research fraternity on this topic is presented here. To alleviate the adverse consequences of road accidents, existing literature examined major intrinsic and extrinsic factors associated with the increased risk of fatal road crashes. Rolison et al. (2018) assessed the human factors and identified characteristics of drivers (e.g., age, gender, adopted safety measures, risk-taking behaviour) that influence the severity of crash outcomes (Rolison et al., 2018). The study utilises law enforcement views, opinions of drivers and road accident records for the assessment and comparison. The results indicate that young drivers are higher risk-takers than older drivers. Moreover, lack of adequate driving experience and driver distraction are major concerns for young drivers. For middle-aged drivers with adequate driving experience, involvement in drugs or alcohol impairment frequently leads to road accidents. On the contrary, visual and cognitive impairment are the foremost causes of crashes involving vehicles driven by older drivers (Rolison et al., 2018; Hammad et al., 2019).

Driving volatility at the time of collision has a significant contribution to crash injury severity (Wali et al., 2020). Driving volatility captures the variation in driving, such as fast acceleration, hard braking, jerky manoeuvres, and risky behaviour towards other road users. Multinomial logit models involving fixed and random parameters were used to estimate the crash severity and its relationship with driving volatility. The study found that injury severity is highly correlated to the driving volatility before 30 s of crash(Wali et al., 2020). A similar logit-based study explored the relationship between alcohol, driver's age, and the influence of passengers on accident severity (Keall et al., 2004). Furthermore, the logit model proposed in Alogaili et al. (2020) shows the relationship of the driver's nationality, cultural background, and education with accident injury severity. Besides the driving volatility, several locality factors such as residential/school zones, intersections with/without signals, number of through lanes, and rain or mist positively correlate with accident severity.

Other studies examined the role of naturalistic environmental factors (e.g., road types, light conditions, wind direction, weather conditions) in amplifying the risk of exposure to fatal road injuries (Pervez et al., 2022; Bergel-Hayat et al., 2013; Ahmad et al., 2020; Hammad et al., 2019). Pervez et al. found that day, time, age, fatigue and speeding were positively associated with severe crashes on mountainous freeway tunnel groups (Pervez et al., 2022). In Bergel-Hayat et al. (2013) authors performs a time series analysis on the aggregated monthly dataset of road crashes for France, the Netherlands, and the Athens region. The findings show that rainfall and temperature are positively correlated with the number of injuries, but frost is negatively correlated. According to (Hermans et al., 2006), the rate of accidents per month changes by 5% depending on weather conditions. While good weather conditions positively contribute to accident severity, it is negative in bad weather due to cautious driving. The study also shows that light condition has a significant contribution to road accident severity, for example, the impact of an accident that happens at night without road lights is likely to be more severe than others (Ullah et al., 2021). In Hammad et al. (2019), the authors studied 50 road accident cases that occurred in the year 2011 to 2016 in District Vehari Punjab, Pakistan. It is found that the number of road accidents is proportional to rainfall patterns as also claimed in other studies (C. Caliendo et al., 2007; Keay et al., 2006). Though the finding matches the previous studies on the impact of rainfall on road accidents, it is impossible to generalise the findings with such small datasets. The study doesn't provide enough statistical inference to justify the results.

Estimating road accident severity is critical and complex particularly due to the involvement of multiple contributing factors, which is why it is crucial to uncover the underlying relationships between accident contributing factors and injury severity. As mentioned above, most research on this topic has used the traditional statistical and regression approach, which is fundamentally based on the assumption of linear and non-linear relationships between input and output parameters. However, in recent days, ML-based feature analysis has been extensively studied. Different Artificial Neural Networks (ANN) are explored in predicting road accident injury severity based on a set of contributing factors (Ogwueleka et al., 2014; Amiri et al., 2020; Shiran et al., 2021). Ogwueleka et al. (2014) discusses an ANN model that utilises past experience to predict the outcome without prior knowledge of the relationships amongst exploratory variables. Reported results show that the ANN model has outperformed other statistical methods. Amiri et al. (2020) used ANN and a hybrid intelligent genetic algorithm to investigate the effects of psychological conditions and physical features of elderly drivers (i.e. aged 65 years or more) when predicting the severity of run-off-road crashes. ANN models predict minor and non-injury severity more accurately than severe injury cases. Moreover, the performance of ANN models significantly degrades when dealing with multiclass classifications, which is why Decision Tree (DT)-based accident severity prediction techniques have become popular in such research (Zheng et al., 2016; Chen et al., 2020; Shiran et al., 2021). DT identifies important features and their relationship with injury severity (Chen et al., 2020). Also, the accuracy of accident prediction in DT-based techniques is significantly higher than that in ANN-based models (Shiran et al., 2021).

Due to the involvement of a high number of contributing factors in road accidents, several studies have investigated the use of support vector machines (SVM) and Random Forest (RF) methods to predict the injury severity (Sharma et al., 2016; Mokhtarimousavi et al., 2019). Sharma et al. (2016) has modelled the traffic accident injury severity as a classification problem using SVM with Gaussian kernel. The SVM model exhibited better prediction accuracy than the multilayer perceptron network such as ANN. To achieve the superior performance of SVM, there is a requirement to have equal data size for each class as prediction accuracy varies based on the size of the dataset (Li et al., 2012). Random forest (RF) is another popular non-parametric ML method extensively studied in the literature to analyse large datasets using numerous independent variables, which are a good fit for road accident research (Mondal et al., 2020; Chen et al., 2020; Yan et al., 2022). Mondal et al. (2020) applied the RF model on the Connecticut crash data for four years (2015–2018) on five injury classes and concluded how the crash took place and the weather conditions that time were the two primary contributing factors. To further improve the performance of the RF method, work done in Yan et al. (2022) integrated RF with Bayesian optimisation (RF-BO) and applied the proposed RF-BO method on the USA road accident dataset from February 2016 to March 2019. The results showed better prediction accuracy compared to traditional ML models. Apart from these, boosting algorithms, such as AdaBoost, L-GBM, XGBoost, and CatBoost, are explored to predict road accident injury and improve prediction accuracy along with reduced computational cost (Parsa et al., 2020; Pradhan et al., 2020; Qu et al., 2019; Ma et al., 2021). For example, Parsa et al. (2020) detected the occurrence of road accidents using XGBoost and tested it on a set of real-time data consisting of road infrastructure data, traffic data, and weather data. The result showed excellent performance of 99% detection accuracy with only 0.16% false alarm. CatBoost-based ML model on analysing the important features (e.g. presented in 3) and their importance on road accident prediction accuracy is studied in Ma et al. (2021).

Most of the above-mentioned models discussed the prediction or

classification accuracy without a deeper understating of the achieved accuracy. Hence model interpretation is essential when dealing with big datasets consisting of many features to explore how features (e.g. road accident contributing factors) interact with the severity of accidents. In context to the existing literature review, below is a list of our contributions in this paper:

- Investigate a set of ensemble ML models to predict the road accident injury severity.
- Evaluate and compare the ML models' performance in prediction accuracy and sensitivity.
- Use Shapley value as an explainable ML technique to rank the road accident contributing factors.
- Re-train the same ML ensemble model using high-rank features and compare the performance.
- Use the most recent New Zealand road accident data up to 2021, including the accident data during the Covid-19 pandemic.

3. Data processing and statistics

It is very important to clean the raw data to get precise and accurate information. The analysis result depends on the quality of the data set, which includes data consistency, accuracy, and non-missing and valuable information. Hence, we first discuss the data processing procedure followed by descriptive statistics of the cleaned dataset.

3.1. Data processing

This study utilises New Zealand's five years (2016–2020) accident data records collected from the Crash Analysis System (CAS)[1] of Te Manatū Waka Ministry of Transport. This data is also available in the open data portal[2].

Two data sets were collected from the CAS system corresponding to information about the person concerned, vehicles, and accident information. Two datasets, namely the 'person' dataset and the 'accident' dataset, were merged into one master dataset containing the accident-causing factors. The merged dataset initially had 378820 rows and 101 columns; however, several out-of-context columns (out of the 101 columns) were removed from the study as the contents in those columns were irrelevant to the factors causing accidents. For example, a column containing information about the nearby police stations was not considered for this study. We thereby selected 36 features related to crashes, such as crash type, location of crash characteristics, environmental factors, vehicle type(s), vehicle factors, and personal (or user) factors that influence the severity of an accident. In this work, we consider the accident types based on severity. According to our data, there are four types of accident severity, and their definition is as follows:

- Fatal Crash: A road crash that results in death
- Serious injury crash: A road crash where any of the parties required medical attention and was taken to hospital.
- Minor injury crash: A road crash where no one needed medical attention but sustained some bruising and superficial cuts.
- Non-injury crash: A road crash where no one sustained any injuries. The police might not always attend to these crashes.

Moreover, the original dataset had multiple duplicates and repeating factors per accident. For example, if one accident involved five persons (two drivers and three passengers), the dataset included all personal information in the factors contributing to the accident, which is unnecessary for our work. Hence, when selecting factors contributing to an accident, we chose only one row per accident by selecting the road user type as "Driver" and the driver's contribution as "primary". In the absence of any "primary" contribution in an accident, we selected "secondary" and no contribution. Thus, after removing all the duplicated data of accident factors, we ended up with 184317 rows of data to work. Details of that data can be found in Section 3.2. These 184317 data rows were further reduced to a dataset of 67971 after cleaning and removing all null and unknown values.

Table 4
Descriptive statistics of dependent variables.

Dependent Variables			
Crash Severity	**Variables**	**Count**	**Percentage %**
	Fatal	1543	0.84
	Serious	10582	5.74
	Minor	42888	23.27
	Non Injury	129304	70.15

3.2. Road accident data analysis

Primarily, the crash data used in this research was the NZ road accident data for the last five years (2016–2020), downloaded from the According to Federal Emergency Management Agency (FEMA) (Goss, 1996), an emergency is defined as an event or occurrence of natural calamities like a storm, tsunami, tornado, hurricane, flood, volcanic eruption, earthquake, landslide, snowstorm, sandstorm, forest fire, nuclear accident, or human-made mishap. Prompt action requires sending a warning of these catastrophes or needing a quick response to save lives. Information is one of the critical elements to taking necessary immediate actions by gathering and sharing data and resources, decisions, and activities during those catastrophe events. A secure framework is mandatory to interact with the infrastructure and get reliable and valuable data from external sources. Data then needs to be analysed to provide information so that the first responder has more accurate and precise information to accomplish the complex task. Device-to-Device (D2D) and the Internet of Things (IoT) under 5G systems are emerging technologies that can play a significant role in enhancing the performance of the response system by providing delay tolerance early warning and post-disaster communication to the affected peoplePraPT database. In most cases, NZ road accident data are derived from the respective crash reports that are completed by NZ police officers present at the crash sites.

Even though we do not have in-depth knowledge of the road accident data collection policy of the NZ police, however, we assume a subset of the input features might be influenced by self-selection bias. These input features are school zone, road curvature, gradient, state highway, number of vehicles involved, and day of the week. The basis of our assumption is manifold such as i) police might be more alert near school areas, curved roads, steep slopes (i.e., high gradient), and state highways; hence accidents occurring in those areas are reported to the police easily, ii) most of the accidents involved with more than two vehicles are generally reported to the police iii) police might increase their activities on Friday night and hence, most of the accidents on that night might be recorded by the police.

The CAS system plays a significant role in supporting NZ road safety strategy "Road to Zero", which states that NZ roads should be safe for travellers and envisions that no people should be killed or seriously injured in road accidents while travelling. This section provides the descriptive statistics of the data and its spatial and temporal distributions.

3.2.1. Descriptive statistics of road accident data

Table 4 provides the descriptive statistics of the variables in terms of counts and percentages for dependent variables. According to the dataset, the dependent variables are divided into four categories based

[1] https://cas.nzta.govt.nz/.
[2] https://opendata-nztaopendata.arcgis.com/.

Table 5
Descriptive statistics for independent variables (features).

Independent Variables	Values	Count	Fatal (%)	Serious (%)	Minor (%)	Non-Injury (%)
Year	2016	37212	0.76	5.67	20.84	72.73
	2017	39295	0.87	6.10	21.59	71.45
	2018	38454	0.86	5.49	23.96	69.69
	2019	36893	0.81	5.77	25.21	68.21
	2020	32463	0.89	5.66	25.06	68.39
Day	Sat	27458	0.92	6.65	22.57	69.86
	Sun	22564	1.16	6.84	24.30	67.70
	Mon	23702	0.81	5.63	23.56	69.99
	Tue	25561	0.70	5.20	23.38	70.72
	Wed	26789	0.73	5.34	23.41	70.52
	Thu	28075	0.75	5.25	23.08	70.92
	Fri	30168	0.82	5.46	22.86	70.87
Weekend	Yes	59441	1.05	6.61	23.03	69.31
	No	124876	0.74	5.33	23.38	70.55
School Zone	Yes	154875	0.59	4.47	19.24	59.73
	No	29432	0.25	1.27	4.03	10.42
	Unknown	10				
Intersection	Yes	62840	0.16	1.71	7.81	45.74
	no	121476	0.68	4.04	15.46	24.42
	Unknown	1				
urban or open	Open	64319	0.60	2.78	9.05	22.47
	Urban	119997	0.24	2.96	14.22	47.68
	Unknown	1				
Junction type	Cross Road	20389	0.06	0.53	2.69	7.79
	Driveway	12570	0.04	0.41	1.70	4.68
	End of road	169	0.00	0.00	0.02	0.07
	Multileg	838	0.00	0.02	0.08	0.36
	Roundabout	9533	0.01	0.14	0.89	4.14
	T junction	35059	0.10	1.08	4.46	13.39
	Y Junction	566	0.00	0.02	0.04	0.25
	Nill	105193	0.63	3.56	13.41	39.48
Road Curvature	Straight	130404	0.43	3.63	16.22	50.48
	Curved	53628	0.41	2.11	7.05	19.52
	Null	285				
Gradient	Flat	143541	0.62	4.33	18.14	54.79
	Hill Road	36059	0.21	1.38	5.02	12.95
	Null	4717				
Gender	Female	32939	0.18	1.64	8.34	7.71
	Male	61878	0.64	3.86	13.83	15.24
	Unknown	89500	0.01	0.24	1.10	47.20
Primary surface condition	Dry	134604	0.65	4.52	17.47	50.39
	Ice/Snow	1578	0.00	0.04	0.23	0.58
	Wet	44009	0.19	1.17	5.55	16.97
	Null	4126				
Natural Light	Bright Sun	65310	0.28	2.26	8.81	24.09
	Dark	51126	0.30	1.68	6.06	19.69
	Overcast	55094	0.22	1.53	7.23	20.92
	Twilight	8705	0.04	0.27	1.16	3.26
	unkown	4082				
Primary weather condition	Fine	140227	0.68	4.72	18.56	52.12
	Hail	65	0.02	0.10	0.35	0.97
	Heavy rain	6537	0.03	0.17	0.81	2.53
	Light rain	27273	0.11	0.69	3.39	10.61

(*continued on next page*)

Table 5 (*continued*)

Independent Variables	Values	Count	Fatal (%)	Serious (%)	Minor (%)	Non-Injury (%)
	Mist or Fog	2656	0.02	0.10	0.35	0.97
	Snow	192	0.00	0.00	0.02	0.08
	Null	7367				
Drug suspected	Not Suspected	111007	0.31	3.54	17.19	39.19
	Suspected	3293	0.19	0.30	0.58	0.72
	Unknown	70017	0.34	1.90	5.50	30.25
Alcohol	Not suspected	27350	0.18	1.95	8.68	4.03
	Suspected	60912	0.62	3.32	12.78	16.33
	Unknown	96055	0.04	0.47	1.81	49.80
Road usage type	50 Max	5	0.00	0.00	0.00	0.00
	Bus	1861	0.01	0.06	0.19	0.75
	Car/wagon	131018	0.45	3.45	16.33	50.85
	Left scene	2221	0.00	0.03	0.09	1.08
	Moped	662	0.00	0.06	0.21	0.09
	Motorcycle	4615	0.11	0.79	1.11	0.50
	SUV	12864	0.08	0.35	1.62	4.94
	Train	3	0.00	0.00	0.00	0.00
	Truck	8410	0.06	0.23	0.82	3.45
	Truck HPMV	34	0.00	0.00	0.00	0.01
	UTE	5772	0.04	0.19	0.76	2.15
	Van	15546	0.09	0.54	2.02	5.78
	Uncoupled towed vehicle	6	0.00	0.00	0.00	0.00
	Unknown	863	0.00	0.02	0.05	0.40

on injury severity: fatal, serious, minor, and non-injury crashes. Among 184321 road crashes, 70.15% have not caused any injury (non-injury crashes), while 23.27% and 5.74%, respectively, have caused minor to serious injuries, and 0.84% were fatal crashes. The statistics for independent variables are given in Table 5. These independent variables have been selected based on different categories mentioned in Table 3 (Rolison et al., 2018) such as road characteristics, human factors, environment, day of the week, and others. The highest fatal crashes (340) occurred in 2017, but considering the overall number, the percentage was more (0.89% fatal crashes) in 2020. Overall, there were fewer road crashes in 2020 (17.61%) due to the COVID-19 lockdown. Moreover, data indicate that Friday (16.37%) has been the deadliest day of the week. Crash proportion during the week has been significantly higher (nearly 35%) than during the weekend, but the occurrence of fatal crashes was more over the weekend (0.30% higher). One of the most alarming findings from the studied data was that school zones are not safe as they should be. 84.02% crashes occurred in the school zones; however, luckily, the majority (59.73%) of them were non-injury crashes. From the perspective of road infrastructure, although 65.10% of crashes happened on urban roads, crashes that took place on open roads proved to be deadlier in comparison (0.46% higher). Similarly, more accidents (70.75%) were recorded on straight roads compared to curved ones. Lastly, from the dataset studied, we can also conclude that dry, sunny weather has proved to be more dangerous in measuring the severity of road crash injuries.

3.2.2. Temporal distribution

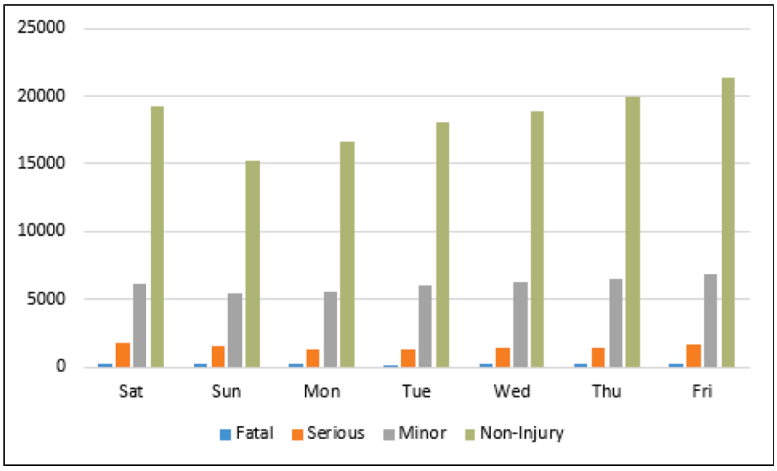
Here we discuss the temporal distribution of road accidents with respect to four injury severities. As illustrated in Fig. 3(a), Friday proves to be the most vulnerable day of the week in any year. Speaking of months, as shown in Fig. 3(b), the highest number of accidents occurred in June and December, while the lowest was observed in April. Also, Fig. 3(c) shows that between 2016 to 2020, the maximum number of accidents happened in 2017, and it is also true for all injury severity classes except minor injuries. We performed a temporal pattern analysis of the time series data to understand the accident trend. The temporal pattern for the fatal crash is shown in Fig. 4 (a) presents the decreasing trend of the accident. A sharp fall can be observed in March/April of 2020 due to a nationwide lockdown in NZ. The seasonal analysis in the same figure shows the variation of crashes throughout the year. It also shows the cycle in the data trend for different years. Fatal, serious, and minor (Fig. 4(a), (b), (c)) crashes occur primarily during winter, whereas non-injury crashes (Fig. 4(d)) peak around the fall.

3.2.3. Spatial distribution

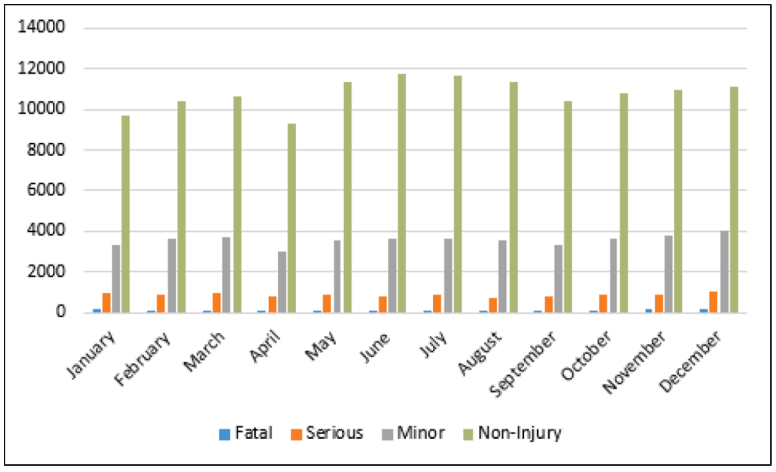
We compare the spatial distribution of road accident data with injury severity, as shown in Fig. 2. As we consider the road accidents for the entire New Zealand, spatial distribution in the figure almost resembles the map of the country. New Zealand consists of two major islands, North Island and South Island, which can be seen from the spatial distribution of accident data in the map. This is because most accidents occurred in major cities like Auckland, Wellington, Tauranga, Christchurch, and Queenstown, located on one of the two major islands. In the figure, also can be seen few other white spaces that refer to remote areas, such as small islands having limited access to vehicles.

4. Methodology

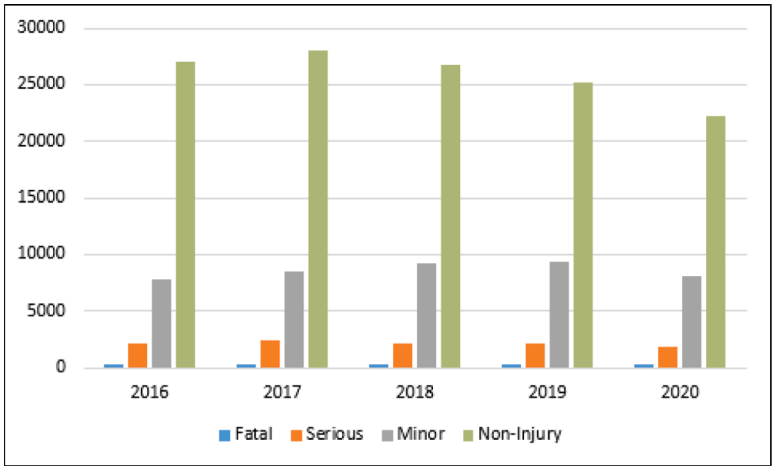
In this research, we investigate road accident prediction as a classification problem that can classify an accident's severity into four categories: fatal, serious, minor, and non-injury. Since more than two classes (or multiple classes) are involved in this road accident severity prediction research problem, it becomes a multiclass classification problem. This research has considered ensemble ML algorithms to analyse road accident datasets. An ensemble ML algorithm combines several base models to produce optimal performance and reduce the dispersion of prediction. To improve the acceptability of road accident prediction, it is essential to understand these factors and their influence on a model,



(a) Day

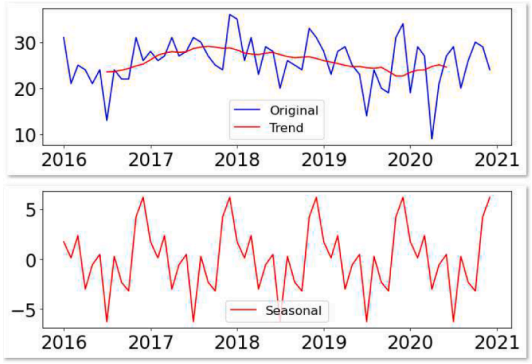


(b) Month

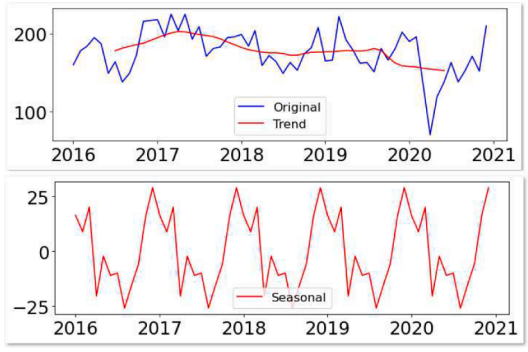


(c) Year

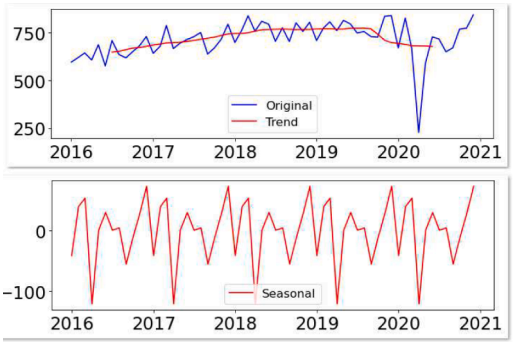
Fig. 3. Crashes by (a) day (b) month (c) year in New Zealand.



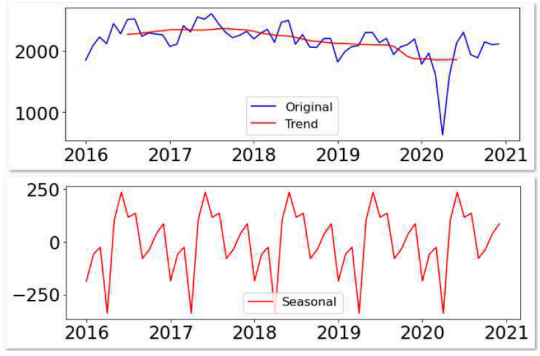
(a) Fatal



(b) Serious



(c) Minor



(d) Non-Injury

Fig. 4. Time series analysis of (a) fatal (b) serious (c) minor (d) non-injury crashes in New Zealand.

which is the primary focus of this research. Thus the Shapley value is analysed to understand the feature contributions toward the target variable. The findings were used to determine the underlying relationships of the contributing factors to the road accident. Fig.1 depicts the flow of this work. The performances of the different ensemble ML models were analysed in this research in terms of accurately predicting road accident severity, understanding the precision of prediction, and calculating the F1-score and recall. The following subsections briefly describe all the ML methods explored in this paper.

4.1. *Random forest (RF)*

RF is a widely used decision tree-based ensemble ML supervised learning algorithm. It is based on the ensemble of decision trees to reduce the risk of overfitting. The performance of RF depends on three main hyperparameters: the number of trees, node size, and the number of features sampled. Feature importance analysis can be evaluated by estimating the road accident contributing factors and their influence on injury severity.

4.2. *Decision jungle (DJ)*

DJ is an extension of decision trees and forests to limit the exponential growth of the tree, which is memory hungry and exhaustive. Unlike RF, in DJ, directed acyclic graphs (DAGs) allow multiple paths from the root to each leaf instead of only one path to every node, as can be found in conventional decision trees (Shotton et al., 2013).

4.3. *Adaptive boosting (AdaBoost)*

AdaBoost is the simplest boosting algorithm based on an ensemble decision tree. It uses an iterative adaptive approach in which weights are adjusted at each iteration by assigning higher weights to incorrectly classified instances. Unlike other decision tree-based algorithms, the decision tree in AdaBoost only starts with one node and two leaves instead of the full-depth decision tree.

4.4. *Extreme gradient boosting (XGBoost)*

XGBoost is another ensemble decision tree-based ML algorithm, an advanced version of the gradient boosting algorithm. It includes system optimisation and algorithmic enhancement. Unlike gradient boosting, XGBoost utilises Taylor expansion to calculate the value of the loss function for base learners and accordingly builds the tree in a greedy manner (Chen et al., 2016).

4.5. *Light gradient boosting (L-GBM)*

L-GBM is another decision tree-based algorithm in which a tree grows vertically (called leaf-wise tree growth), where the leaf that can minimise the information loss will split, and the rest of the leaves remain at the same level. Faster training speed with high accuracy and lower memory usage are the advantages of L-GBM (Bentéjac et al., 2021).

4.6. *Categorical boosting (CatBoost)*

CatBoost is known as categorical boosting based on gradient boosting of decision trees. It uses one-hot encoding to handle categorical data. In Bentéjac et al. (2021), authors argue that by using the minimal variance sampling technique for node split, the number of instances needed for each boosting iteration can be reduced, which in turn can significantly improve the quality of the model.

5. Experimental results

Performance of the above-mentioned ML algorithms was measured

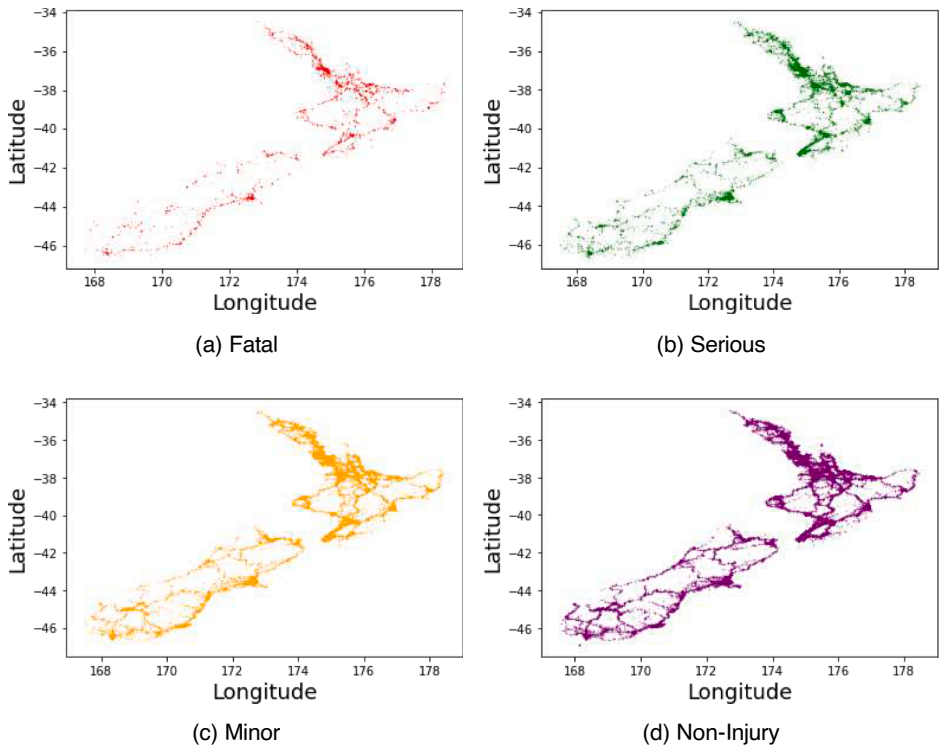


Fig. 2. Spatial distribution of different crash severities.

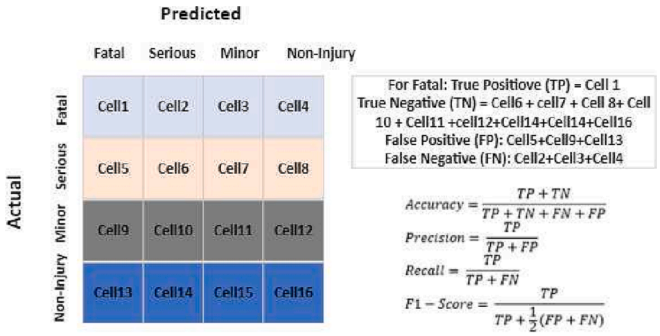


Fig. 5. Performance metrics for classification.

based on the performance metrics for classification as shown in Fig. 5: In context to accident severity categories, the dataset's severity classes were distributed as 1.25%, 9.27%, 44.31%, and 45.18% for fatal, serious, minor, and non-injury, respectively. The distribution indicates the imbalance in data for multi-classes severity. To handle this imbalance, we have used the synthetic minority oversampling technique (SMOTE)(Chawla et al., 2002) and have generated synthetic positive instances to balance all of the classes (Parsa et al., 2020) in the training dataset. However, we have used imbalanced multi-class data to infer the trained model. We have also converted alphabetic values to numeric values to fit the CatBoost model data and performed one (1) hot encoding to convert data to numeric values. To avoid overfit and underfit-related issues, we have implemented stratified k-fold validation in which k represents the number of groups for training data split. In these experiments, we have used $k = 10$. Furthermore, for training and testing of the model, the dataset has been split in the ratio of 80% for training and 20% for testing.

The result for prediction is presented in Table 6. The result shows that RF is superior to the other five models considered. More specifically, the model accuracy for RF is 81.45%, whereas XGBoost generates the second highest accuracy of 78.52%, followed by 76.94%, 74.12%, 65.61%, and 69.68%, respectively, for L-GBM, DJ, AdaBoost, and CatBoost. This is similar to the results reported in Wahab et al. (2019). In this context, it should be noted that prediction accuracy can be misleading for a multiclass classification problem as there are possibilities that the model might neglect some of the classes. Hence, we have also computed the precision, recall, and F1-score. Precision measures the exactness of the model means What proportion of identifications was correct, whereas recall measures the probability of relevant items being detected. The results in Table 6 indicate that RF provides the maximum precision (81.68%) and outperforms XGBoost (79.87%), L-GBM

Table 6
Performance of models for multiclass injury severity prediction.

	Imbalance Data				Balance Data (After SMOTE)			
Model	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
Random Forest	52.36	42.14	35.96	37.12	81.45	81.68	81.42	81.04
Decision Jungle	51.2	38.04	35.58	35.86	74.12	73.65	74.17	73.87
ADABoost	55.35	59.19	38.14	39.03	65.61	66.65	65.52	65.06
XGBoost	55.54	48.44	38.36	39.16	78.52	79.87	78.68	78.16
L-GBM	55.73	52.21	38.51	39.54	76.94	78.54	77.12	76.42
CatBoost	55.91	53.44	38.97	40.06	69.68	70.83	69.76	69.68

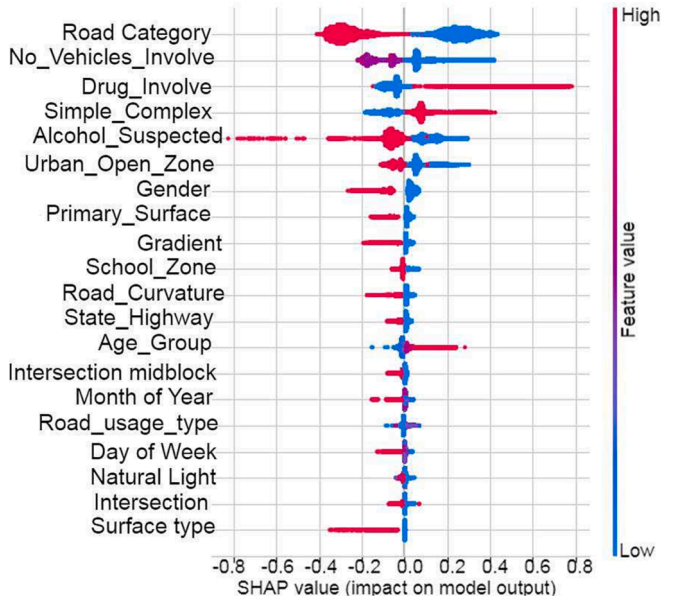


Fig. 6. Evaluating feature influence of RF model performance using SHAP.

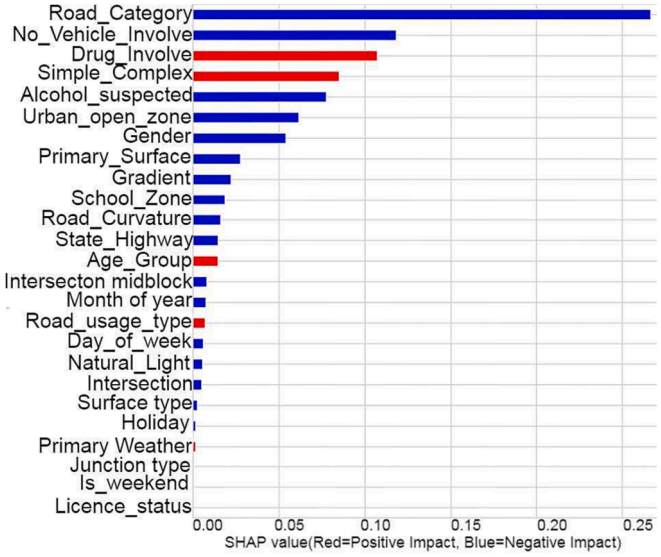


Fig. 7. Feature importance using SHAP for RF.

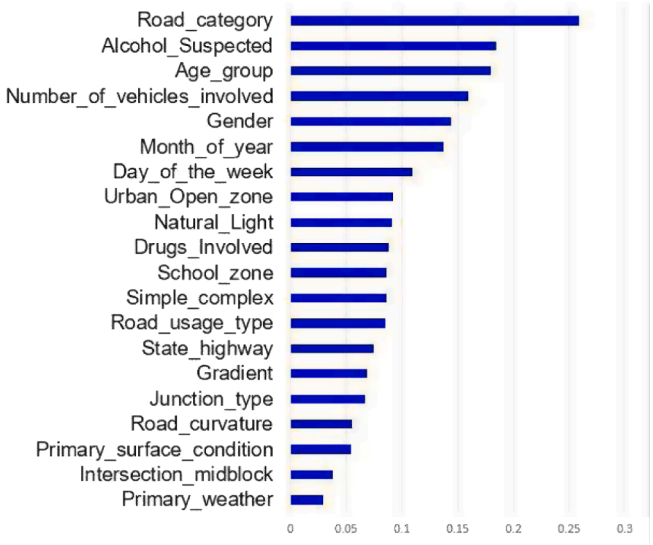


Fig. 8. Feature importance using Permutation for RF.

(78.54%), DJ (73.65%), AdaBoost (66.65%), and CatBoost (70.83%). For RF, a similar trend can be observed for recall and F1-score, where RF's recall score (81.42%) and F1-score (81.04%) are found to be reasonably better than that of XGBoost (78.68% for recall score and 78.16% for F1-score). This is consistent with the results reported in Wahab et al. (2019).

5.1. *Feature analysis using explainable ML*

To understand the impact of road accident contribution factors towards an accident's severity, we performed feature analysis using XML. Feature analysis is an effective way to uncover the relationship between the features and output, i.e. target variable. For this, we performed feature importance analysis using an explainable ML technique. Explainable ML helps users to understand and trust the results predicted/classified by an ML algorithm. In this research, we utilise the Shapley value as one of the explainable ML techniques due to its model-agnostic characteristics. The Shapley value is a cooperative game-theoretic concept introduced by Lloyd Shapley in 1951 to evaluate each player's contribution towards the overall gain of the game (Shapley, 1951). We have conducted both the global and local levels of Shapley value analysis in this work. While the global level provides insight into the overall uses of features in the model, the local level analysis helps to understand a decision for an individual data point. Here, the SHAP value analysis is performed on the RF model using balanced data as it shows superior performance than other ML models in this study.

5.1.1. *Global SHAP value analysis*

The SHAP quantifies each feature's contribution to the target variables produced by the ML model (e.g. here, we will use fandom forest). Each point on the summary plot represents a Shapley value for the respective feature and the corresponding data point. The features in the SHAP plot are arranged in descending order, i.e., feature importance is ordered from high (red color) to low (blue color) in the *Y* axis, and its influence on model output is shown in the *X* axis. The points on the SHAP plot represent the training data. The value on the left of 0 in the *X* axis represents the observation that shifts (i.e. move to a negative or positive direction) the target value in the negative direction, and the value on the right shifts in the positive direction. As shown in Fig. 6, the road category has the greatest impact on the model performance. The high road category values (Vehicle track, Motorway) on the left side negatively influence the model, as most accidents happen in rural and urban areas. These findings are similar to the results reported in Sun et al. (2021). Fig. 6 also shows drug consumption results in severe accidents. The Shapley value for the drug-involved feature increases with the drug consumption, i.e., the higher the drug consumption, the more the probabilities of accident and injury severity. This finding is consistent with the previous study (Zhang et al., 2021), where drug consumption was ranked as the 6th important factor in crash severity.

Fig. 7 is the simplified version of Fig. 6 and shows the feature importance ranking for RF using a balanced dataset. As the name suggests, feature importance ranking is a technique to evaluate the importance of a feature to predict a target variable.Fig. 7 illustrates the feature importance considering the SHAP value and its impact on global model output. Feature with large absolute values is considered an important feature in SHAP feature importance analysis. The global feature importance can be calculated using $I_j = \frac{1}{N}\sum_{i=1}^{N}|\phi_j^{(i)}|$ where N is the size of the dataset and ϕ_j is the SHAP value of feature j. A permutation is one of the other ways to generate feature importance, as shown

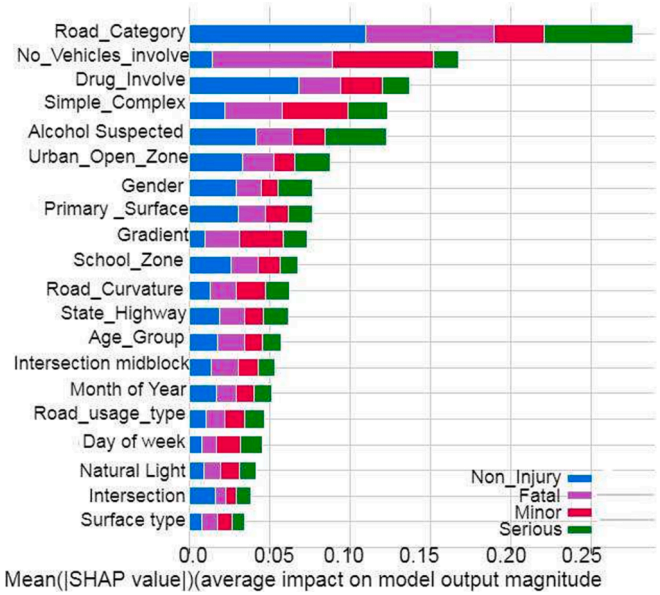
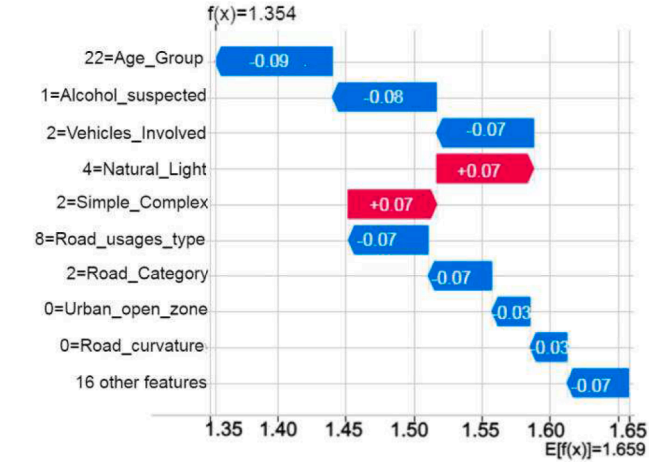
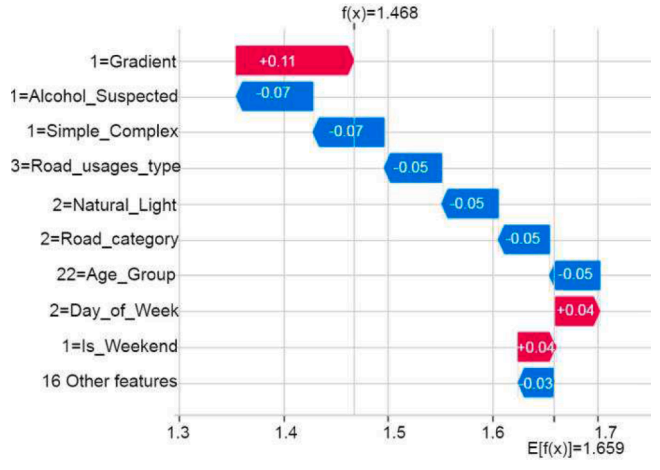


Fig. 9. Multiclass Feature importance using SHAP for RF.



(a) Data point 1



(b) Data point 1000

Fig. 10. Individual SHAP waterfall Plot for Observations (a) point 1 (b) point 1000.

in Fig. 8. Permutation feature importance is based on the model's prediction error, where a feature is important if shuffling its values increases the model error; otherwise, it is unimportant. According to Fig. 7 and 8, the road category is the most significant feature compared to any other features. Though a little variation can be observed in feature importance ranking between SHAP and permutation-based approach, a common set of features is identified by both feature importance rankings. Further, we generate feature importance in Fig. 9 for the multiclass scenario based on SHAP value. We noticed that road category and number of vehicles involved in an accident are the highest rank features for fatal severity. For serious severity, road category and alcohol are the dominant features. An interesting observation, in this case, could be that the number of vehicles is the highest rank feature for a minor injury accident, followed by multiparty involvement in a crash. The multiparty involvement is termed the "Simple_complex" feature in our dataset and refers to the number of parties involved in an accident.

5.1.2. *Local SHAP value analysis*

The local Shapley value analysis aims to explore how the model uses the features to predict the target variable for each data point, i.e., in a single row in the dataset. We obtain a Shapley value for every feature and sample in the training set. Therefore, there is no Shapley value of a feature, but rather a feature has as many Shapley values as there are samples. For a sample data point x, our model outputs the value $f(x)$. For feature j, the Shapley value $\phi(x,j)$ is the contribution that features j has on the output $f(x)$. These contributions are given with respect to a base value, namely, the average value of $f(x)$ for all training samples x. This can be written as follows:

$$f(x) = E(f(x)) + \sum_j \phi(x,j) \tag{1}$$

Fig. 10 shows the SHAP waterfall plot for data points 1 and 1000 (i.e. row numbers 1 and 1000 in the dataset). The SHAP waterfall plot explains the feature's contribution to the model prediction compared to the mean prediction. For data point 1, the model predicted the output value of 1.354 (at the top of the plot), whereas the mean prediction is 1.659, as shown at the bottom. Here features with red color imply that they push the prediction to the right, i.e., higher, and those with blue color push the prediction to the left, i.e., lower. Moreover, the feature values on a particular data point are on the left of the feature name, while SHAP values for the feature are on the arrows (e.g. blue and red arrows). For example, in Fig. 10(a), the age group reduces the prediction of an accident by 0.09 compared to the average prediction. On the contrary, natural light and multiparty involvement (simple_complex) increase the accident prediction by 0.07 compared to the mean prediction value. Moving on to Fig. 10(b), multiparty involvement (simple_complex) reduces the accident prediction by 0.07, whereas road gradient increases it by 0.11 in comparison to the mean prediction.

5.1.3. *Feature dependency analysis*

Feature dependency analysis in this study is computed using the SHAP dependence plot, which shows the marginal effect that features have on the target variable. It also helps us to uncover the variation and distribution of feature SHAP values. Each point on the dependence plot represents a prediction of the dataset's target value of an individual row. In this plot, we consider the four features, which are road category, drug, age, and gender, and the system automatically selects another feature from the dataset with the strongest interaction, as shown in Fig. 11. For example, Fig. 11(a) shows the effect of road category (1:Rural, 2:Urban, 3:Foot track, 4:Motorway, 5: vehicle track) and the number of vehicles. Many accidents occur in rural areas as the number of vehicles involved is high compared to other road categories. Fig. 11(b) illustrates the interaction of drug and road speed zone (i.e. urban_Open_Zone). In the graph, for urban_Open_Zone, the blue colored plots represent urban and red colored plots indicate open zones. As can be observed in the graph, higher consumption of drugs results in more serious injuries when

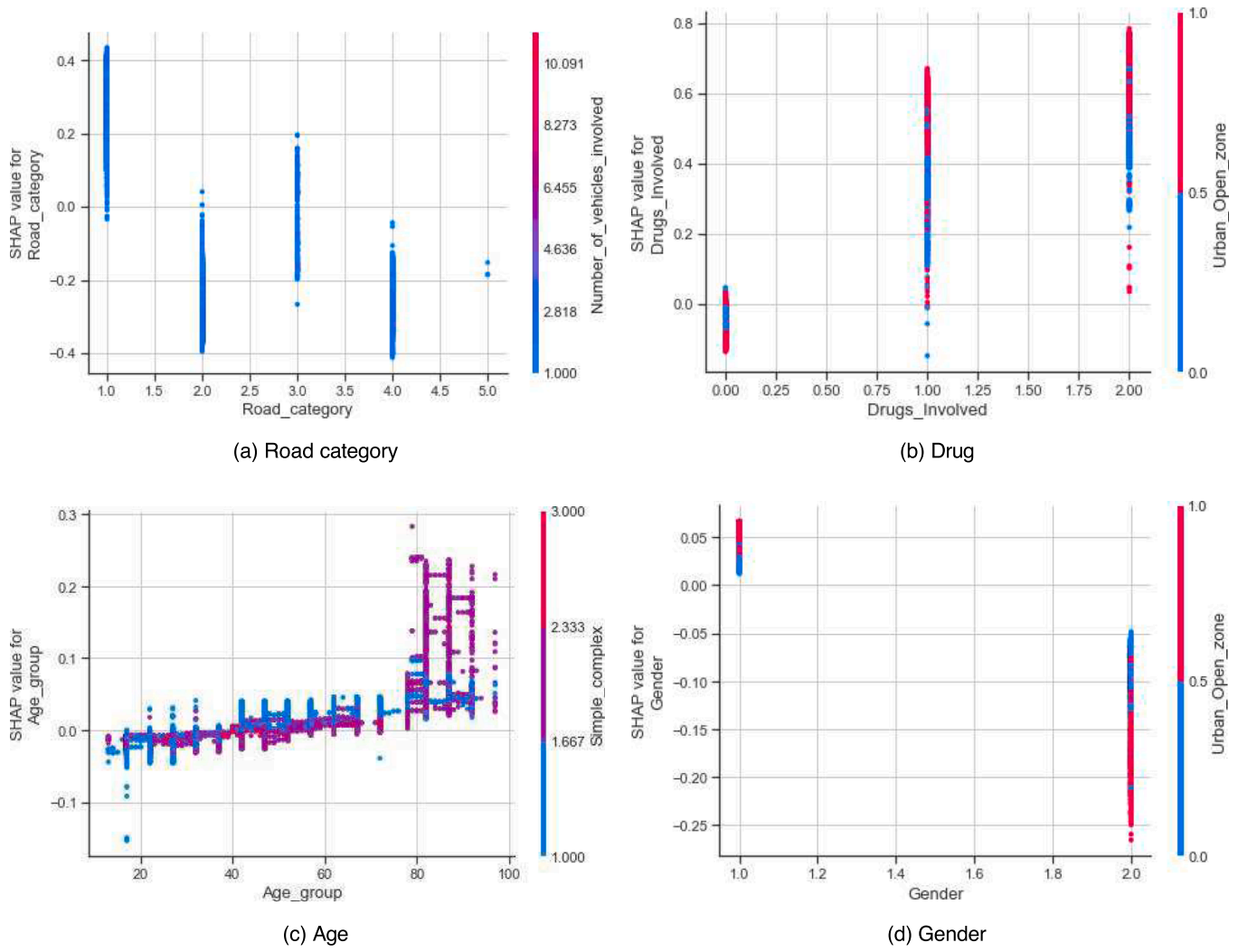


Fig. 11. Feature dependency analysis using SHAP (a) road category (b) drug (c) age and (d) gender.

accidents happen in an open zone where a vehicle may drive at a higher speed compared to that in an urban zone where the speed of the vehicle will be comparatively less. In New Zealand, most open zones have no middle barriers, which increases the probability of head-on collision. The involvement of drugs makes it more deadly as it has a negative impact on the reaction time and concentration of drivers while driving. Fig. 11(c) explains the relationship between age and multiparty involvement in an accident. For multiparty category shown on the right has three types, namely, the single party, and multiple parties simple and complex, which are shown in blue, purple, and red color on the vertical line. Most elderly drivers are involved in simple multiparty accidents compared to other age groups. Middle-aged drivers (say between 38 to 55) are involved in multiparty complex road crash (Zubaidi et al., 2021). Finally, 11(d) shows the interaction between gender and road with different speed limits, with 1 representing male and 2 representing female. From the figure, it can be seen that male drivers are involved in more severe crashes than female drivers despite the fact that female drivers are more prone to accidents in open zones.

5.1.4. *Performance improvement using high ranked features*

The feature rank obtained in subSection 5.1.1 is used to retrain the models, and in this research, we have used the first 15 high-ranked features as shown in Fig. 7. The performance of the ML models after retraining is shown in Table 7. A slight improvement, measuring approximately between 1% to 1.5%, can be observed for RF, XGBoost, and L-GBM models. An 8% performance improvement is achieved for CatBoost, which is quite significant. Similarly, to be noted are 5% enhancement for AdaBoost and 6% for DJ. These findings confirm that the features identified through global and local Shapley value analysis provide valuable insight into road safety design.

Table 7
Performance of ML models with 15 high-ranked features.

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	81.27	82.31	82.97	81.76
Decision Jungle	80.35	79.59	80.19	79.59
ADABoost	70.14	70.76	70.06	68.85
XGBoost	78.84	78.50	78.73	77.58
L-GBM	77.37	76.6	77.26	75.8
CatBoost	77.04	76.63	76.94	75.63

6. Implication and limitations

The findings from this research can be used as evidence-based for road safety professionals and practitioners to design and evaluate road safety policies and to understand progress towards the long-term vision of the Safer Journeys' strategy. Since drug involvement and alcohol

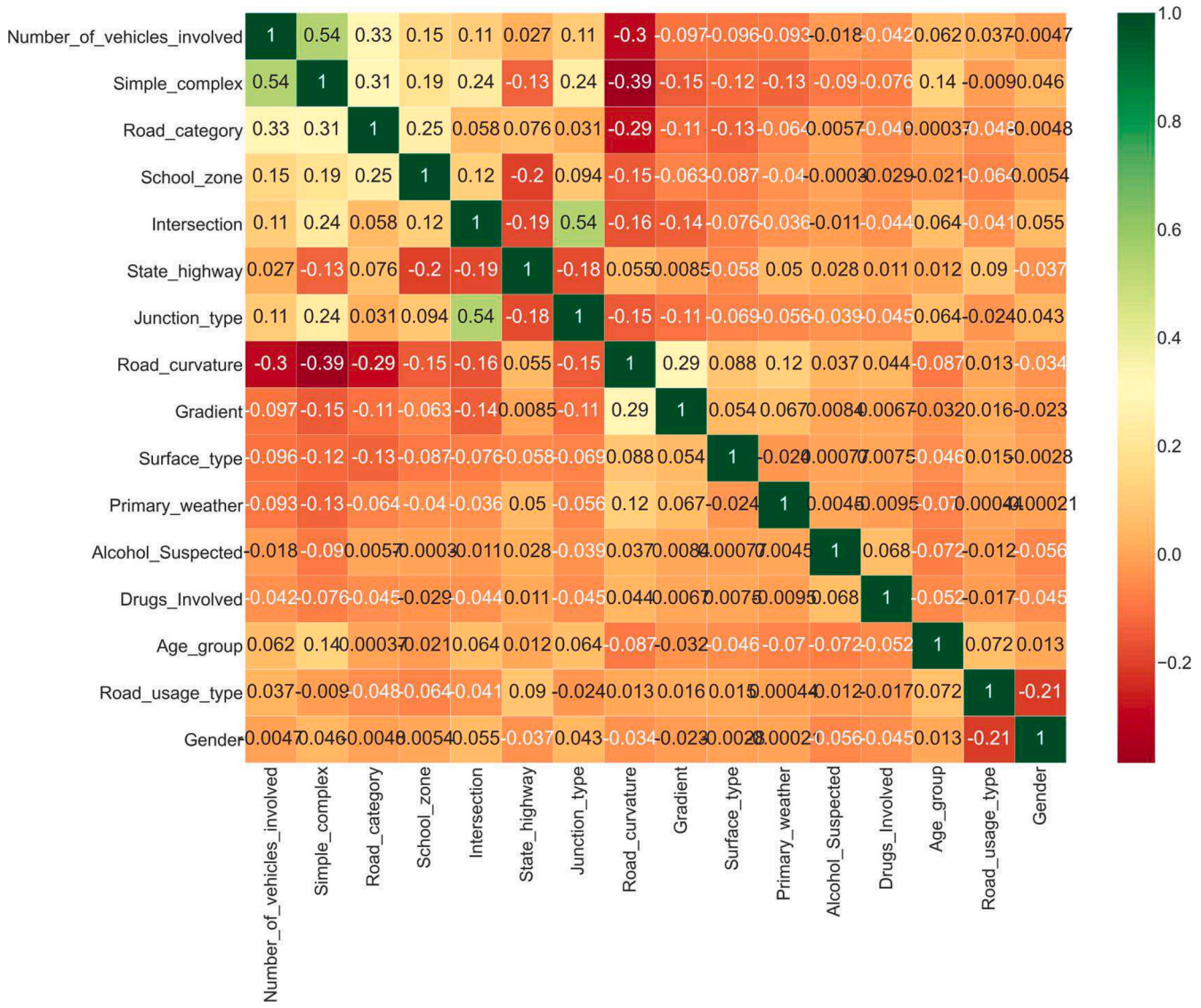


Fig. A.12. Cross-Correlation matrix of independent variables.

consumption are responsible for high crash severity, more preventive and corrective litigation can be considered to minimise them. Our study also shows that road geometry and road categories with different speed zones need to be closely monitored as preventive measures. Moreover, we have shown that elderly drivers are more prone to accidents compared to other age groups and countermeasures like putting up more elderly-friendly road signs, providing them training as needed and making them more aware of the changes can help them to be safe and sound when driving. Furthermore, the insight of this research can serve as an effective tool to minimise the social cost associated with road accidents.

This work does have some limitations as highlighted here. We have considered only the New Zealand road accident dataset, in this study, and hence the findings regarding road accident contributing factors cannot be generalised to any other dataset. However, the relationships between those contributing factors can be well utilized to make strategic decisions for road safety design in other similar countries across the globe. Another limitation of this study is using SMOTE as an over-sampling technique to deal with imbalanced data. We have a comparatively low number of fatal and serious crash records in the dataset. However, it is found that SMOTE doesn't attenuate the bias for high-dimensional data and also suffers from over-generalisation as the majority class is ignored by the method (Nekooeimehr et al., 2016). We plan to explore different combinations of SMOTE and undersampling techniques as part of our future work. Moreover, with the recent advancement of artificial intelligence (AI), deep learning-based models can be investigated to achieve better performance and imbalanced data can be trained with sophisticated large-scale deep neural networks.

7. Conclusion

This research presented extensive insight into road accident injury severity and its contributing factors. We have used the New Zealand road accident dataset for the study. For the evaluation purpose, we have employed ML-based classification algorithms, namely, DJ, RF, AdaBoost, L-GBM, CatBoost, and XGBoost. Referring to the findings presented in Section 5, the accuracy of the RF is 7.33%, 15.84%, 2.93%, 4.51%, and 11.77% higher than Decision Jungle, ADABoost, XGBoost, L-GBM, and catBoost respectively. The precision of the RF model also shows better performance than other models by 3% to 15% as there is no risk of overfitting. The Shapley value is also used in this study as an explainable ML technique to interpret the model performance. By performing global and local SHAP analysis, we are able to establish the relationship between the contributing features on overall model

Table B.8
Additional results on the performance of ML models.

		Precision	Recall	F1-Score
Random Forest	Fatal	35	24	29
	Serious	20	17	18
	Minor	53	51	52
	Non-Injury	59	63	61
	Macro Avg	42	39	40
	Weighted Avg	52	53	53
Decision Jungle	Fatal	17	24	20
	Serious	15	20	17
	Minor	48	49	48
	Non-Injury	54	50	52
	Macro Avg	34	35	34
	Weighted Avg	47	46	47
ADABoost	Fatal	34	29	31
	Serious	19	35	24
	Minor	57	34	43
	Non-Injury	58	71	64
	Macro Avg	51	38	39
	Weighted Avg	54	51	51
XG Boost	Fatal	48	24	32
	Serious	3335	119	1614
	Minor	58	56	57
	Non-Injury	61	73	66
	Macro Avg	50	40	42
	Weighted Avg	57	59	57
LGBM	Fatal	39	29	33
	Serious	33	11	16
	Minor	58	53	55
	Non-Injury	60	74	66
	Macro Avg	48	42	43
	Weighted Avg	56	58	57
CATBoost	Fatal	51	24	33
	Serious	35	7	12
	Minor	58	53	56
	Non-Injury	60	75	67
	Macro Avg	51	40	42
	Weighted Avg	57	59	56

performance and individual data points. Analysis of results also highlights factors like road category and the number of vehicles involved in an accident that results in severe injuries. Moreover, results highlighted that elderly drivers are more prone to multiparty accidents leading to severe injuries. The highest-ranked features obtained from global SHAP analysis are used to retrain all the previously considered models and performance improvements as a result of that are noted. This can be helpful for future studies as it provides an option to compare the performance of an arbitrarily chosen ML model.

Our future works include further investigation on how to minimize the impact of the self-selection bias in the feature selection process. we also like to investigate the performance of this classification problem using different deep neural architectures together with customized loss functions with proper regularization methods to train large-scale imbalanced datasets. Moreover, in our future research, we are planning to include more road accident contributing factors such as ethnicity, frost, snow, and the medical condition of the driver.

CRediT authorship contribution statement

Shakil Ahmed: Conceptualization, Formal analysis, Software, Data curation. **Md Akbar Hossain:** Conceptualization, Methodology, Formal analysis, Writing - original draft. **Sayan Kumar Ray:** Investigation, Writing - review & editing, Supervision. **Md Mafijul Islam Bhuiyan:** Methodology, Software, Writing - original draft. **Saifur Rahman Sabuj:** Investigation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

The authors express their gratitude to the Ministry of Transport, New Zealand and Manukau Institute of Technology, New Zealand for their support in doing this research.

Appendix A. Cross-correlation matrix

Fig. A.12

Appendix B. Performance of considered ML models

Table B.8

References

Transport - road accidents - oecd data.https://data.oecd.org/transport/road-accidents.htm.

M. of Transport, Te marutau - ngā tatauranga ā-tau: Safety - annual statistics,http://transport.govt.nz/statistics -and-insights/safety- annual-statistics/summary/.

Rolison, J.J., Regev, S., Moutari, S., Feeney, A., 2018. What are the factors that contribute to road accidents? an assessment of law enforcement views, ordinary drivers' opinions, and road accident records. Acc. Anal. Prevent. 115, 11–24. https://doi.org/10.1016/j.aap.2018.02.025 https://www.sciencedirect.com/science/article/pii/S0001457518300873.

Ahmed, S., Hossain, M.A., Bhuiyan, M.M.I., Ray, S.K., 2021. A comparative study of machine learning algorithms to predict road accident severity. In: 2021 20th International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS). IEEE, pp. 390–397.

Rolison, J.J., Regev, S., Moutari, S., Feeney, A., 2018. What are the factors that contribute to road accidents? an assessment of law enforcement views, ordinary drivers' opinions, and road accident records. Acc. Anal. Prevent. 115, 11–24.

Hammad, H.M., Ashraf, M., Abbas, F., Bakhat, H.F., Qaisrani, S.A., Mubeen, M., Fahad, S., Awais, M., 2019. Environmental factors affecting the frequency of road traffic accidents: a case study of sub-urban area of pakistan. Environ. Sci. Pollut. Res. 26 (12), 11674–11685.

Wali, B., Khattak, A.J., Karnowski, T., 2020. The relationship between driving volatility in time to collision and crash-injury severity in a naturalistic driving environment. Analytic Methods Acc. Res. 28, 100136.

Keall, M.D., Frith, W.J., Patterson, T.L., 2004. The influence of alcohol, age and number of passengers on the night-time risk of driver fatal injury in new zealand. Acc. Anal. Prevent. 36 (1), 49–61.

Alogaili, A., Mannering, F., 2020. Unobserved heterogeneity and the effects of driver nationality on crash injury severities in saudi arabia. Acc. Anal. Prevent. 144, 105618.

Pervez, A., Huang, H., Lee, J., Han, C., Li, Y., Zhai, X., 2022. Factors affecting injury severity of crashes in freeway tunnel groups: A random parameter approach. J. Transp. Eng. Part A: Syst. 148 (4), 04022006.

Bergel-Hayat, R., Debbarh, M., Antoniou, C., Yannis, G., 2013. Explaining the road accident risk: Weather effects. Acc. Anal. Prevent. 60, 456–465.

N. Ahmad, A. Ahmed, B. Wali, T.U. Saeed, Exploring factors associated with crash severity on motorways in pakistan, in: Proceedings of the Institution of Civil Engineers-Transport, Thomas Telford Ltd, 2020, pp. 1–10.

Hermans, E., Wets, G., Van den Bossche, F., 2006. Frequency and severity of belgian road traffic accidents studied by state-space methods. J. Transp. Statist. 9 (1), 63.

Ullah, H., Farooq, A., Shah, A.A., 2021. An empirical assessment of factors influencing injury severities of motor vehicle crashes on national highways of Pakistan. J. Adv. Transp.

Caliendo, C., Guida, M., Parisi, A., 2007. A crash-prediction model for multilane roads. Acc. Anal. Prevent. 39 (4), 657–670.

Keay, K., Simmonds, I., 2006. Road accidents and rainfall in a large australian city. Acc. Anal. Prevent. 38 (3), 445–454.

Ogwueleka, F.N., Misra, S., Ogwueleka, T.C., Fernandez-Sanz, L., 2014. An artificial neural network model for road accident prediction: a case study of a developing country. Acta Polytechnica Hungarica 11 (5), 177–197.
Amiri, A.M., Sadri, A., Nadimi, N., Shams, M., 2020. A comparison between artificial neural network and hybrid intelligent genetic algorithm in predicting the severity of fixed object crashes among elderly drivers. Acc. Anal. Prevent. 138, 105468.
Shiran, G., Imaninasab, R., Khayamim, R., 2021. Crash severity analysis of highways based on multinomial logistic regression model, decision tree techniques, and artificial neural network: A modeling comparison. Sustainability 13 (10), 5670.
Zheng, Z., Lu, P., Tolliver, D., 2016. Decision tree approach to accident prediction for highway–rail grade crossings: Empirical analysis. Transp. Res. Rec. 2545 (1), 115–122.
Chen, M.-M., Chen, M.-C., 2020. Modeling road accident severity with comparisons of logistic regression, decision tree and random forest. Information 11 (5), 270.
B. Sharma, V.K. Katiyar, K. Kumar, Traffic accident prediction model using support vector machines with gaussian kernel, in: Proceedings of fifth international conference on soft computing for problem solving, Springer, 2016, pp. 1–10.
Mokhtarimousavi, S., Anderson, J.C., Azizinamini, A., Hadi, M., 2019. Improved support vector machine models for work zone crash injury severity prediction and analysis. Transp. Res. Record 2673 (11), 680–692.
Li, Z., Liu, P., Wang, W., Xu, C., 2012. Using support vector machine models for crash injury severity analysis. Acc. Anal. Prevent. 45, 478–486.
Mondal, A.R., Bhuiyan, M.A.E., Yang, F., 2020. Advancement of weather-related crash prediction model using nonparametric machine learning algorithms. SN Appl. Sci. 2 (8), 1–11.
Yan, M., Shen, Y., 2022. Traffic accident severity prediction based on random forest. Sustainability 14 (3), 1729.
Parsa, A.B., Movahedi, A., Taghipour, H., Derrible, S., Mohammadian, A.K., 2020. Toward safer highways, application of xgboost and shap for real-time accident detection and feature analysis. Acc. Anal. Prevent. 136, 105405.
Pradhan, B., Ibrahim Sameen, M., 2020. Predicting injury severity of road traffic accidents using a hybrid extreme gradient boosting and deep neural network approach. In: Laser scanning systems in highway and safety assessment. Springer, pp. 119–127.
Qu, Y., Lin, Z., Li, H., Zhang, X., 2019. Feature recognition of urban road traffic accidents based on ga-xgboost in the context of big data. IEEE Access 7, 170106–170115.
Ma, Z., Mei, G., Cuomo, S., 2021. An analytic framework using deep learning for prediction of traffic accident injury severity based on contributing factors. Acc. Anal. Prevent. 160, 106322.
J. Shotton, T. Sharp, P. Kohli, S. Nowozin, J. Winn, A. Criminisi, Decision jungles: Compact and rich models for classification, Adv. Neural Inform. Process. Syst. 26.
T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.
Bentéjac, C., Csörgő, A., Martínez-Muñoz, G., 2021. A comparative analysis of gradient boosting algorithms. Artif. Intell. Rev. 54 (3), 1937–1967.
Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.
Wahab, L., Jiang, H., 2019. A comparative study on machine learning based algorithms for prediction of motorcycle crash severity. PLoS one 14 (4), e0214966.
L.S. Shapley, Notes on the n-person game–ii: The value of an n-person game. (1951).
Sun, T.-J., Liu, S.-J., Xie, F.-K., Huang, X.-F., Tao, J.-X., Lu, Y.-L., Zhang, T.-X., Yu, A.-Y., 2021. Influence of road types on road traffic accidents in northern guizhou province, china. Chin. J. Traumatol. 24 (01), 34–38.
Zhang, Y., Liu, S., Wu, J., Lu, Q., Yang, Q., Song, L., Yu, X., Zhang, H., Yang, Z., 2021. Drug-related crash severity analysis using the highway safety information system data. IEEE International Intelligent Transportation Systems Conference (ITSC), IEEE 2021, 3659–3664.
Zubaidi, H.A., Obaid, I.A., Alnedawi, A., Das, S., 2021. Motor vehicle driver injury severity analysis utilizing a random parameter binary probit model considering different types of driving licenses in 4-legs roundabouts in south australia. Saf. Sci. 134, 105083.
Nekooeimehr, I., Lai-Yuen, S.K., 2016. Adaptive semi-unsupervised weighted oversampling (a-suwo) for imbalanced datasets. Expert Syst. Appl. 46, 405–416.