Jordan Journal of Civil Engineering, Volume **18**, No. **4**, **2024**
DOI: https://doi.org/10.14525/JJCE.v18i4.05

Jordan Journal of Civil Engineering

Journal homepage: https://jjce.just.edu.jo

Exploring the Utility of Auto-machine Learning in Predicting Traffic Accident Severity in Jordan

R. Al Shafie [1], Kh. Jadaan [2], Sh. El-Badawy [3], S. Shwaly [4], U. Shahdah [5]*

[1] Department of Civil Engineering, Mansoura University, Egypt. E-Mail: ranaalshafie@std.mans.edu.eg
2) Department of Civil Engineering, The University of Jordan, Jordan. E-Mail: kjadaan@gmail.com
3) Department of Civil Engineering, Mansoura University, Egypt. E-Mail: sbadawy@mans.edu.eg
4) Department of Civil Engineering, Mansoura University, Egypt. E-Mail: sayed_shwaly@mans.edu.eg
5) Department of Civil Engineering, Mansoura University, Egypt.
** Corresponding Author. E-Mail: usama.shahdah@mans.edu.eg.*

ARTICLE INFO

Article History:
Received: 29/3/2024
Accepted: 25/8/2024

ABSTRACT

The World Health Organization (WHO) estimates that approximately 1.35 million annual fatalities are caused by road traffic accidents, representing a significant global challenge. This study focuses on creating a machine-learning model to forecast traffic accident severity in Jordan. The objective of this study is to help reduce fatalities and economic losses caused by accidents, which can achieve using various machine-learning classifiers, like Decision Trees, Random Forests, Light Gradient Boosting Machines, Extra Trees, Bagging Classifiers, and Gradient Boosting. These models were assessed before and after down-sampling to deal with imbalance in accuracy, balanced accuracy, recall, precision, and F1-score metrics with and without hyper parameter tuning. The study emphasized the importance of advanced analytics in improving road safety measures and reducing accident severity. The researchers analyzed a dataset of 115,148 accidents. Factors considered included traffic volume, environmental conditions, and road geometry features. The data was segmented into urban and rural categories for customized modeling. Down-sampling improved the models' ability to detect injuries and deaths in both urban and rural areas. Hyper parameter tuning offered an additional improvement in balanced recall and F1-score, particularly after down-sampling. In urban areas, the LGBM and Gradient Boosting classifiers showed the most significant gains in recall for minority classes, while in rural areas the Random Forest, Bagging, and Extra Trees classifiers maintained a better balance between precision and recall. All classifiers achieved high accuracy (above 0.97) in both urban and rural areas, so that these models can accurately forecast the severity of accidents, including property damage, fatalities, and injuries. The authors have also recommended incorporating variables of temporal patterns, driver behavior, demographic data, and geographic location into the models of accident prediction to make them more efficient and stronger.

Keywords: Accident severity prediction, Auto-machine learning, Accident classifiers, Python, Jordan.

INTRODUCTION

Transportation networks are crucial for efficient mobility, but road safety requires addressing issues, such as road design, traffic volume, speed, driver health, and weather conditions (Hareru et al., 2022). The World Health Organization estimates that road traffic accidents cause 1.35 million deaths and from 20 million to 50 million injuries annually, with considerable economic and social repercussions (Hareru et al., 2022). In Jordan, traffic accidents account for 4.5% of deaths and rank as the third biggest cause of mortality (Al-Omari et al., 2013; Miqdady & de Oña, 2020). In 2022, the Public Security Directorate reported 169,404 events, resulting in 11,510 injuries with an estimated cost of 322 million Jordanian dinars, underlining the critical need for comprehensive safety measures.

The primary goals are to help reduce fatalities and alleviate the economic impact of traffic accidents. This study highlights numerous other noteworthy objectives that align with its primary premise, including the improvement of the accuracy of AI classifiers in predicting accident severity to better anticipate serious accidents. This research tackled several other significant objectives that align with its primary emphasis, including:

1. Predicting a severity classifier on the data for Jordan.
2. Identifying characteristics that escalate accident severity.

Accident classification models in Jordan, built using data from 2015 to 2022, combine variables, such as Average Annual Daily Traffic (AADT), AM Peak Hour Volume (PHV), geometric features, and accident data, with validation based on 20% of the dataset. Globally, road safety remains a concern due to the huge socio-economic cost of accidents (Al-Masaeid, 2009). In Jordan, the rise in traffic accidents, hurting public well-being, needs urgent actions to address issues linked to driving behavior, infrastructure, and pedestrian interactions (Al-Rousan et al., 2021; Miqdady & de Oña, 2020).

Machine learning (ML) models have become vital tools for predicting accident severity by evaluating historical data and recognizing patterns (Al-Rousan et al., 2021). These models, including Decision Trees (DTs), Random Forests (RFs), Support Vector Machines (SVMs), and Neural Networks, are chosen based on the specific context and available data. AutoML, or Automated ML, facilitates model selection and parameter tuning, boosting accessibility and accuracy (Kunwar, 2019).

Among these ML techniques, Gradient Boosting stands out for its efficiency and excellent performance with huge datasets, albeit it requires careful tweaking to minimize overfitting (Li et al., 2019). The RF classifier, which aggregates predictions from numerous DTs, is resistant against noise and overfitting but may face issues with high-dimensional data (Afshar et al., 2022; Berhanu et al., 2023). The Extra Trees classifier, similar to the RF classifier, averages predictions from several randomized DTs and is successful for feature selection (Ampomah et al., 2020). Lastly, the Bagging classifier prevents overfitting by training classifiers on random sub-sets of data and pooling their predictions (Hussain & Ashraf, 2023).

The Random Forest (RF) classifier is a powerful method for evaluating accident severity and boosting road safety, especially when standard statistical models struggle with parameter volatility. This ensemble technique constructs numerous Decision Trees (DTs) and aggregates their predictions to avoid issues of overfitting and noise, although it may suffer scalability challenges with high-dimensional data (Dogru & Subasi, 2018). Similarly, The Extra Trees classifier balances overfitting risk and computational efficiency by averaging predictions from randomized DTs, whereas the Bagging classifier decreases variance and minimizes overfitting by integrating predictions from several base classifiers. The machine learning approaches are critical for building successful accident prevention and management strategies, since they have shown effectiveness in forecasting accident severity and enhancing safety across varied settings and datasets.

Lee et al. (2019) employed ML approaches, such as RF, ANN, and DT, to model traffic accident severity in Seoul, showing the major influence of weather, especially rainfall, on accident severity and the requirement for weather-related data in models. Aldhari et al. (2022) employed logistic regression, gradient boosting, and RF to forecast highway accident severity in Saudi Arabia, displaying high model accuracy in detecting severity indicators and underscoring the necessity of advanced analytics for enhancing road safety.

Ketha and Imambi (2019) studied road accidents using Logistic Regression, Naïve Bayes, K-Nearest

Neighbors, DT, and RF models, identifying DT as the most accurate model in predicting accident severity. Their findings underlined the need for thorough data and more research to boost prediction accuracy. Sahin (2020) compared ensemble tree approaches, including gradient boosting, and RF, for landslide susceptibility mapping in Turkey. The study indicated RF to be the most effective approach, illustrating the relevance of factor selection and validation in model performance.

Alqudah et al. (2019) discovered that Decision Trees (DTs) were the most successful model among multiple ML models for forecasting road accidents using data from 2012 to 2019, encouraging future research with additional parameters. Baykal et al. (2023) obtained 81.6% accuracy with Auto-ML approaches on 1.7 million US accident records, finding the Random Forest (RF) algorithm to be the most successful approach while emphasizing the importance of refining models and considering aspects, like weather and time.

Unlike previous studies in Jordan that predominantly used conventional statistical techniques, our research introduces cutting-edge AI classifiers, like Decision Trees, Random Forests, Light Gradient Boosting Machines, Extra Trees, Bagging classifiers, and Gradient Boosting classifiers. We bridged this gap using these sophisticated tools and thoroughly assessed their performance in predicting accident severity. This comprehensive approach provides accurate and reliable forecasts, thus instilling confidence in our findings.

This research gives vital insights into road safety in Jordan by applying ML algorithms, such as RF, Forest classifier, Extra Trees, and Bagging classifier to examine accident injury severity. By constructing predictive models employing these approaches, the project tackles the current gap in their application inside Jordan, seeking to improve evidence-based decision-making and enhance road safety.

METHODOLOGY

Data Collection

A comprehensive dataset was compiled from multiple sources, including incident reports from the Jordan Traffic Institute (JTI), traffic volume and speed data from the Greater Amman Municipality, infrastructure documentation from the Ministry of Public Works and Housing, and geospatial information from site visits and Google Earth. It was a disaggregate data. Road parts were selected to cover Jordan's entire distance, connecting its northern and southern regions and eastern and western regions, around 12 rural road segments and 18 urban road segments. These road parts represented mountainous roads, leveled areas, valleys, sandy conditions, curves, different grades of roads, weather conditions, and different volume levels of traffic, together with the role played by confined spaces, and low visibility conditions influencing accident severity. The detailed screening process provided an opportunity to study the accident patterns along various stretches of roads in Jordan.

Defined variables for the traffic data include speed, AADT, and PHV, involving traffic volumes. The variables of road geometry include the lane width, the median width, and the number of access points in each road section. Accident-based variables include the accident type, surface conditions, severity of accident, vehicle type, road surface type, geometric properties, weather conditions, and lighting conditions. The accident locations were initially recorded using the Universal Transverse Mercator (UTM) projected coordinate system, specifically the World Geodetic System (WGS 1984 Zone 36 North). In accident scenes, traffic officers use GPS devices to obtain UTM coordinates. Data verification was performed in several ways, like comparing information from several sources to find and fix any inconsistencies, performing site inspections, and conducting field surveys to confirm that reports on road conditions and GIS data are accurate. It is important to apply strict pre-processing and data cleaning procedures to eliminate errors or inconsistencies, apply statistical techniques to evaluate data dependability, and ensure that the data satisfies the requirements for correctness and completeness. By implementing these methods, the study guarantees the validity of the research findings and their implications for boosting road safety by ensuring the data reliability and accuracy.

Data Pre-processing

The present study focuses on data pre-processing due to the need to develop a clean and complete dataset for analysis and modeling. The pre-processing steps, outlier detection, residual analysis, feature selection based on statistical chi-square test, data balancing, and data partitioning, are described. Outliers are detected using CURE plots. Due to a high frequency of missing entries, missing values have been eliminated. All the statistical tests are conducted with Python and R-studio, with a threshold P-value set at 0.05 for statistical significance. Data balancing is done to reduce the size of the majority class, since it is big in proportion to ensure correct predictions. The dataset was partitioned using 80% for training and 20% for testing. Pre-processing steps and data manipulation were performed using the Lazy attribute found in Python to close the dataset. The dataset consisted of 114,383 incidents trained to differentiate between urban and rural areas, validated, and tested. This division followed best practices for robust model evaluation, ensuring that the models were trained and tested on diverse sub-sets of data, as illustrated in Figure (1).

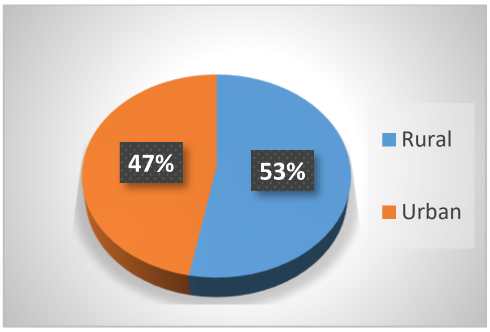

Figure (1): Classifying the data into two distinct segments

Model Selection

The study assessed 20% of the dataset for testing accident severity prediction, together with an analysis of categorical features and their relationship to the prediction. It further assesses the accuracy, recall, precision, and F1-score as predictive capability measures in AutoML through systematic experiments across various algorithms. Equation (1) deals with accuracy, the performance measured by several metrics, such as true positives, false negatives, and false positives, enabling a detailed assessment of recall (Equation 2), precision (Equation 3), and F1-score (Equation 4).

$$\text{Accuracy} = \frac{\text{TP+TN}}{TP+TN+FP+FN} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$Precision = \frac{TP}{TP\ FP} \quad (3)$$

$$F1 - sorce = 2 * \frac{(Recall*Precision)}{(Recall+Precision)} \quad (4)$$

where (TP) is the number of sites correctly classified as hazardous, (TN) is the number of sites correctly classified as non-hazardous, (FP) is the number of false positives, and (FN) is the number of false negatives.

RESULTS AND DISCUSSION

Urban Road Section Data

Accidents are categorized by severity: Property Damage Only (PDO), injury, and fatality. Factors affecting injury severity include geometric features, environmental conditions, accident variables, and traffic flow characteristics. Data for urban and rural areas is presented as counts (%) (Table 1 and Table 6). Chi-square tests were used if cell counts exceeded 5, otherwise Fisher's exact test was applied. A p-value < 0.05 indicated significance. Significant features were used in the AI model, while non-significant ones were excluded. Python 3.11.4 was used for analysis, except for Fisher's test, conducted in R 4.3.1.

Fatalities were the highest in residential areas (0.43%), while injuries were the highest in sub-urban areas (3.217%). Curved roads had the highest injury rates. Roads with 3.5m lane width had higher death rates (0.51%) than roads with 3.1m lane width (0.33%). More U-turns were correlated with higher injury rates, though fewer accidents occurred in these sections. Two-way divided roads had higher fatality rates with (0.39%). The significant differences in injury rates between one-way, two-way divided, and two-way undivided roads by Fisher's exact test had a p-value =0.0004998. This trend might be due to higher speeds on the separated traffic flows. On the other hand, not-level roads had the highest injury rates (5.09%) More access points increased injury and fatality rates. Roads with four lanes had the highest injury rates (0.06%). Roads with wider shoulders (>2.5 meters) had the highest injury rates (0.053%). Wider medians (>4 meters) had the lowest injury and death rates, but more accidents. Roads with lower grades had high injury and fatality rates. Sections that are >3 km long had the lowest injury and fatality rates.

Injuries were the highest on not dry roads, but not with a significant relationship. Unpaved roads had the highest injury rates, but no significant relationship was found due to low unpaved road counts. Night lighting conditions significantly affected injury rates, with higher percentages of inadequate lighting (5.08%). Adverse weather slightly increased injury rates, but the relationship wasn't statistically significant. Run-over accidents had high death rates with (20%). No significant relationship was found between vehicle type and injury or death rates.

Higher speeds correlated with higher injury and death rates. The highest death rate was in the >60 km/h group (0.013%). Peak hour volume and AADT had significant relationships with injury rates. Higher peak hour volumes had the highest death rates, while lower volumes had the highest injury rates. Lower AADT had the highest death (0.006%) and injury rates (0.039%). Higher speeds were significantly correlated with high injury and death rates.

The percentages of PDO, injury and death accidents among the different significant features are compared in Table 1.

Table 1. Comparing the percentages of accident severity (PDO, Injury and death accidents) among a sample of significant features (geometric, environmental, and traffic flows) in urban areas

Features	Categories	Total	PDO	Injuries	Deaths	P-val.
Geometric Features						
Road Section Type	One Way	5510 (6.76%)	5407 (98.1%)	92 (1.67%)	11 (0.2%)	<0.001
	Others	2292 (2.81%)	2276 (99.3%)	15 (0.654%)	1 (0.044%)	
	Two Way Divided	59720 (73.29%)	57673 (96.6%)	1812 (3.034%)	235 (0.394%)	
	Two Way not divided	13968 (17.14%)	13552 (97.0%)	385 (2.756%)	31 (0.222%)	
Geometrics of Road Section Properties	Level	77994 (95.71%)	75622 (97.0%)	2126 (2.726%)	246 (0.315%)	<0.001
	Not Level	3496 (4.29%)	3286 (94.0%)	178 (5.092%)	32 (0.915%)	
Section Length (km)	<1km	5919 (7.26%)	5729 (96.8%)	171 (2.889%)	19 (0.321%)	<0.001
	1-3km	28704 (35.22%)	27678 (96.4%)	912 (3.177%)	114 (0.397%)	
	>3km	46867 (57.51%)	45501 (97.1%)	1221 (2.605%)	145 (0.309%)	
Environmental features						
Lighting Conditions	Adequate night lighting	19432 (23.85%)	18502 (95.2%)	799 (4.112%)	131 (0.674%)	<0.001
	Daylight	57632 (70.72%)	56233 (97.6%)	1280 (2.221%)	119 (0.206%)	
	Inadequate night lighting	4426 (5.43%)	4173 (94.3%)	225 (5.084%)	28 (0.633%)	
Traffic Flow Features						
AM Peak Hour Volume (veh/hr)	< 1138	19735 (24.22%)	19075 (96.7%)	581 (2.944%)	79 (0.4%)	0.001
	1138-1724	21701 (26.63%)	20972 (96.6%)	656 (3.023%)	73 (0.336%)	
	1724-2393	19576 (24.02%)	19051 (97.3%)	478 (2.442%)	47 (0.24%)	
	>2393	20478 (25.13%)	19810 (96.7%)	589 (2.876%)	79 (0.386%)	
Average AADT (veh/day)	< 4356	19256 (23.63%)	18371 (95.4%)	761 (3.952%)	124 (0.644%)	<0.001
	4356-6855	19247 (23.62%)	18756 (97.4%)	454 (2.359%)	37 (0.192%)	
	6855-10786	21475 (26.35%)	20891 (97.3%)	525 (2.445%)	59 (0.275%)	
	>10786	21512 (26.4%)	20890 (97.1%)	564 (2.622%)	58 (0.27%)	
Speed (km/hr)	< 40	29737 (36.49%)	29068 (97.7%)	628 (2.112%)	41 (0.138%)	<0.001
	40-50	18183 (22.31%)	17711 (97.4%)	430 (2.365%)	42 (0.231%)	
	50-60	27355 (33.57%)	26329 (96.2%)	913 (3.338%)	113 (0.413%)	
	>60	6215 (7.63%)	5800 (93.3%)	333 (5.358%)	82 (1.319%)	

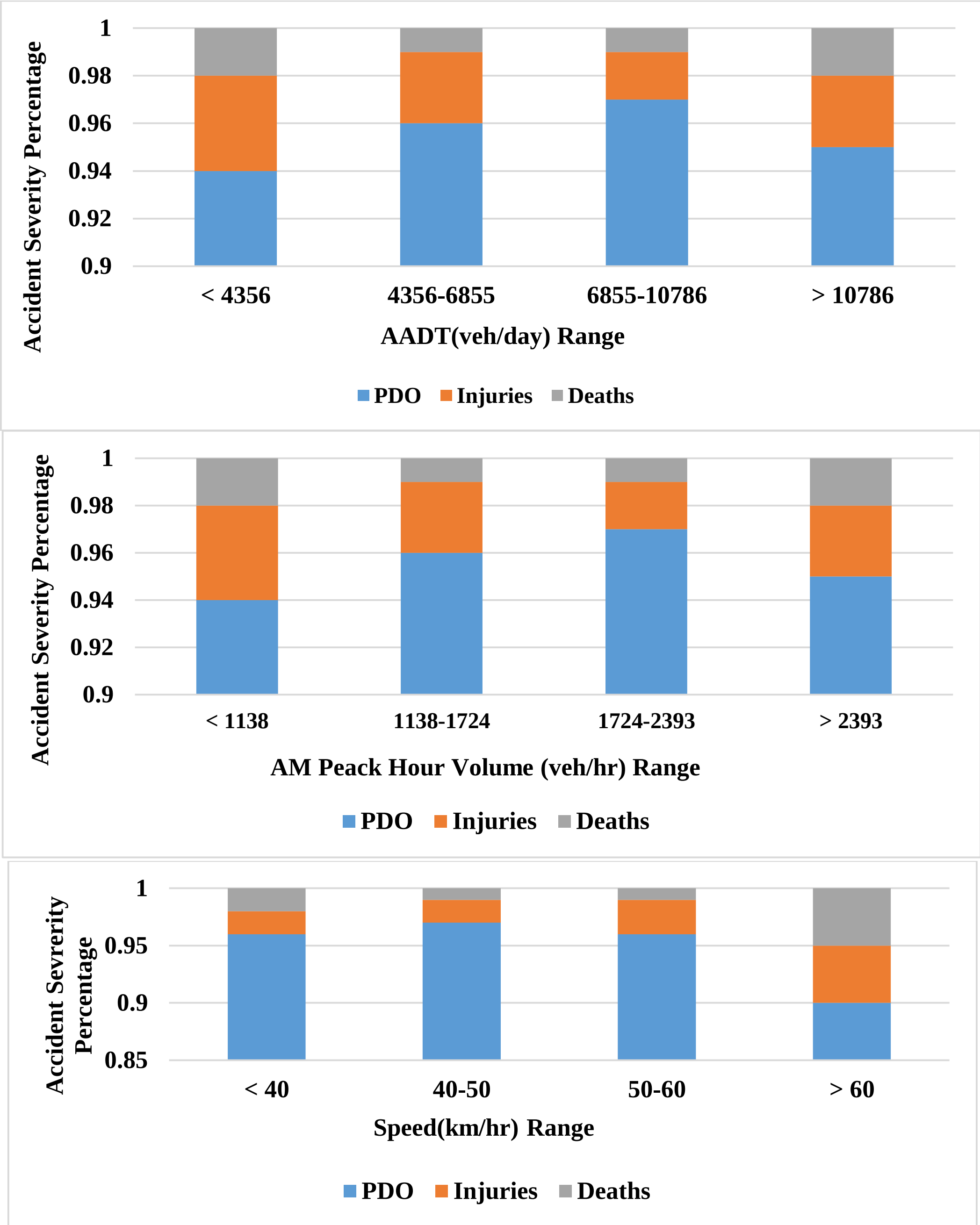

Figure (2): Bar plots showing significant traffic flow features and the percentages of different severity levels in urban data

Down sampling played a pivotal role in feature selection and ML post-data segmentation. The primary objective was to enrich insights from the data, diminish data dispersion, and enhance algorithm efficiency by reducing the number of data points in the majority class. A significant data imbalance existed, with 53,396 records for PDO, 1,316 records for injuries, and 145 fatalities. However, successful down sampling led to a more balanced dataset comprising 159 records for PDO, 159 for injuries, and 145 for fatalities. A comparison of the performance metrics of the machine learning classifiers on accident data in urban areas is carried out before and after down sampling, with and without hyper-parameter tuning, by assessing their performance measures of accuracy, balanced accuracy, recall, precision, and F1-score.

Tables 2 and 3 how the performance measurements of the different classifiers on the urban data before down sampling, both with and without hyper-parameter tuning.

Table 2. Urban data before down sampling (without hyper-parameter tuning)

	Accuracy	Balanced	Recall			Precision			FI-Score		
			PDO	Injuries	Deaths	PDO	Injuries	Deaths	PDO	Injuries	Deaths
Decision Tree	0.972	0.334	0.999	0.002	0.002	0.97	0.01	0.2	0.986	0.011	0.005
LGBM	0.973	0.333	1	0.003	0.001	0.975	0.01	0.01	0.987	0.002	0.001
Random Forest	0.973	0.333	1	0.002	0.001	0.974	0.02	0.01	0.988	0.003	0.002
Bagging	0.973	0.334	0.999	0.003	0.002	0.974	0.01	0.125	0.986	0.002	0.005
Extra Trees	0.973	0.333	1	0.002	0.001	0.974	0.01	0.01	0.986	0.001	0.001

Table 3. Urban data before down sampling (with hyper-parameter tuning)

	Accuracy	Balanced	Recall			Precision			FI-Score		
			PDO	Injuries	Deaths	PDO	Injuries	Deaths	PDO	Injuries	Deaths
Decision Tree	0.973	0.334	0.999	0.002	0.003	0.974	0.012	0.25	0.986	0.012	0.005
LGBM	0.973	0.334	0.997	0.004	0.002	0.974	0.02	0.09	0.985	0.003	0.002
Random Forest	0.973	0.334	1	0.003	0.002	0.974	0.023	0.03	0.986	0.004	0.003
Bagging	0.973	0.334	0.999	0.0031	0.003	0.974	0.012	0.127	0.986	0.003	0.005
Extra Trees	0.973	0.333	1	0.0032	0.002	0.974	0.013	0.012	0.986	0.002	0.002

Hyper-parameter tuning resulted in subtle improvements to the accuracy and other performance metrics. To make the point clearer, hyper-parameter tuning done on the decision tree classifier guaranteed that accuracy remained stable at 0.973 while offering some minor improvements in balanced recall and precision for PDO; that is, the injuries resulted in a somewhat more balanced performance. Minor improvements for the LGBM classifier were also shown, especially in balanced recall and precision for injuries and deaths, though the overall accuracy remained stable. The Random Forest classifier has similarly maintained consistent class accuracy while its balanced recall and precision are tuned towards particular classes. Bagging and Extra Trees classifiers marginally improved in terms of balanced recall and precision, indicating better balancing in class prediction. That is, most of the classifiers are relatively stable with respect to the accuracy metric, remaining at about 0.973 in value, while what got mainly improved by hyper-parameter tuning were the balanced recall and precision scores, more so for PDO, enhancing classifier ability to predict less frequent outcomes.

Tables 4 and 5 present the performance measures of different classifiers following down sampling, both with and without hyper parameter tuning.

Table 4. Urban data after down-sampling (without hyper-parameter-tuning)

	Accuracy	Balanced	Recall			Precision			FI-score		
			PDO	Injuries	Deaths	PDO	Injuries	Deaths	PDO	Injuries	Deaths
Decision Tree	0.973	0.886	1	0.761	0.916	0.941	0.941	0.83	0.969	0.842	0.871
LGBM	0.973	0.886	1	0.761	0.92	0.941	0.941	0.826	0.96	0.842	0.86
Random Forest	0.973	0.886	1	0.761	0.941	0.914	0.914	0.826	0.969	0.831	0.86
Bagging	0.973	0.88	1	0.785	0.941	0.942	0.942	0.846	0.969	0.857	0.88
Extra Trees	0.973	0.893	1	0.761	0.941	0.941	0.941	0.83	0.969	0.842	0.871

Table 5. Urban data after down sampling (with hyper-parameter tuning)

	Accuracy	Balanced	Recall			Precision			FI-score		
			PDO	Injuries	Deaths	PDO	Injuries	Deaths	PDO	Injuries	Deaths
Decision Tree	0.973	0.892	1	0.762	0.896	0.941	0.914	0.827	0.97	0.831	0.86
LGBM	0.973	0.9	1	0.762	0.938	0.941	0.97	0.833	0.97	0.853	0.882
Random Forest	0.973	0.907	1	0.762	0.958	0.941	1	0.836	0.97	0.865	0.893
Bagging	0.973	0.88	1	0.785	0.916	0.941	0.942	0.846	0.969	0.857	0.88
Extra Trees	0.973	0.907	1	0.761	0.958	.0941	1	0.836	0.97	0.864	0.893

All classifiers returned to very high accuracy levels after down sampling and fine-tuning of hyper-parameters, where Bagging classifier showed the highest recall rate and precision rate in case of accidents. LGBM had the best precision and recall in case of injuries. Random Forest classifier returned the lowest F1-score for injuries. However, Bagging and Extra Trees algorithms showed similar performances. Decision Trees classifier had the lowest recall and precision scores.

The results showed improvements in the accuracy for classifiers, like Gradient Boosting, LGBM, Random Forest, Bagging, and Extra Trees classifiers, by down sampling. Balanced recall increased, especially for the minority class of injuries and death cases. The F1-scores improved as well. Gradient Boosting classifier showed the biggest improvement. Furthermore, compared to the default setting, the Decision Tree classifier performed well in balanced recall for injuries and deaths with down sampling.

Finally, the performance of the classifiers was compared before and after down-sampling. Before down-sampling, all classifiers performed highly accurate; however, Bagging classifier resulted in the worst recall and precision rates. The LGBM classifier resulted in the best accuracy, but had a poor retrieval rate. After down-sampling and hyper-parameter tuning, all classifiers displayed very good accuracy, rates with the Bagging classifier attaining the highest recall and precision. The Decision Tree classifier was eventually identified to be the best choice, resulting from its high recall compared to precision.

Rural Road Section Data

Injury severity was significantly related to geometric properties. More injuries and deaths occurred on leveled roads [1661 (0.06196%), 487 (0.01817%)] than on unleveled ones [179 (0.04871%), 54 (0.01469%)]. The study shows that leveled roads claim more injuries and loss of life compared to unleveled ones. This, herefore, calls for all-encompassing measures for road safety. These include higher speeds, complacency in drivers, and lack of vital safety elements, such as (guardrails and barriers, adequate signage, pavement markings, and emergency shoulders), underlining the need for constant driver education and infrastructure building. Injury severity was significantly different for road surface conditions, with more injuries and deaths on dry roads [1744 (0.06108%), 515 (0.01804%)] than on not dry ones [96 (0.04972%), 26 (0.01346%)].

On the other hand, injury severity significantly differed among all flow features. Higher peak hour volumes increased injuries and deaths, with the highest group showing [601 (0.0721%), 183 (0.02195%)] compared to the lowest group [192 (0.04339%), 45 (0.01017%)]. AADT showed similar injury and death rates for the lowest [378 (0.0562%), 132 (0.01963%)] and the highest [458 (0.05986%), 146 (0.01908%)] volumes. The lowest injuries and deaths were in the 1100-1700 AADT group [367 (0.04395%), 70 (0.00838%)], while the highest was in the 1700-2300 group [637 (0.08213%), 193 (0.02488%)]. Injury rates decreased with increasing speed, with the highest speed group having the lowest number of deaths [68 (0.0151%)], while 40-60 km/hr had the highest deaths [243 (0.02016%)]. The results from the lowest AADT group can be compared to those from the highest AADT group, wherein bad driving behaviors and inadequate infrastructure in low-volume areas are combined with better infrastructures and reduced accident severities in high-volume areas. The intermediate AADT group has the best traffic flow, with the lowest injury and fatality rates due to its effective flow and efficient measures for safety.

The percentages of PDO, injury and death accidents among the different significant features are compared in Table 6.

Table 6. Comparing the percentages of accident severity (PDO, injury and death accidents) among a sample of significant features (geometric, environmental, and traffic flows) in rural areas

Features	Categories	Total	PDO	Injuries	Deaths	P-val.
Geometric Features						
Geometric Properties	Level	26809 (87.94%)	24661 (91.988%)	1661 (6.196%)	487 (1.817%)	0.002
	Not Level	3675 (12.06%)	3442 (93.66%)	179 (4.871%)	54 (1.469%)	
Environmental Features						
Road surface conditions	Dry	28553 (93.67%)	26294 (92.088%)	1744 (6.108%)	515 (1.804%)	0.039
	Not Dry	1931 (6.33%)	1809 (93.682%)	96 (4.972%)	26 (1.346%)	
Traffic Flow Features						
AM Peak Hour Volume (veh/hr)	<500	4425 (14.5)%	4188 (94.644%)	192 (4.339%)	45 (1.017%)	<0.001
	500-700	12177(39.9%)	11277 (92.609%)	693 (5.691%)	207 (1.7%)	
	700-900	5545 (18.2%)	5085 (91.704%)	354 (6.384%)	106 (1.912%)	
	>900	8336 (27.3%)	7552 (90.595%)	601 (7.21%)	183 (2.195%)	

Average AADT (veh/day)	<1100	6726 (22.06%)	6216 (92.417%)	378 (5.62%)	132 (1.963%)	<0.001
	1100-1700	8351 (27.39%)	7914 (94.767%)	367 (4.395%)	70 (0.838%)	
	1700-2300	7756 (25.44%)	6926 (89.299%)	637 (8.213%)	193 (2.488%)	
	>2300	7651 (25.1%)	7047 (92.106%)	458 (5.986%)	146 (1.908%)	
Speed(km/hr)	<40	7101 (23.29%)	6503 (91.579%)	484 (6.816%)	114 (1.605%)	0.001
	40-60	12056 (39.55%)	11080 (91.904%)	733 (6.08%)	243 (2.016%)	
	60-80	6825 (22.39%)	6346 (92.982%)	363 (5.319%)	116 (1.7%)	
	>80	4502 (14.77%)	4174 (92.714%)	260 (5.775%)	68 (1.51%)	

As shown in Figure 3, the bar plots show different categories of traffic flow features and the percentages of different injury levels in rural areas.

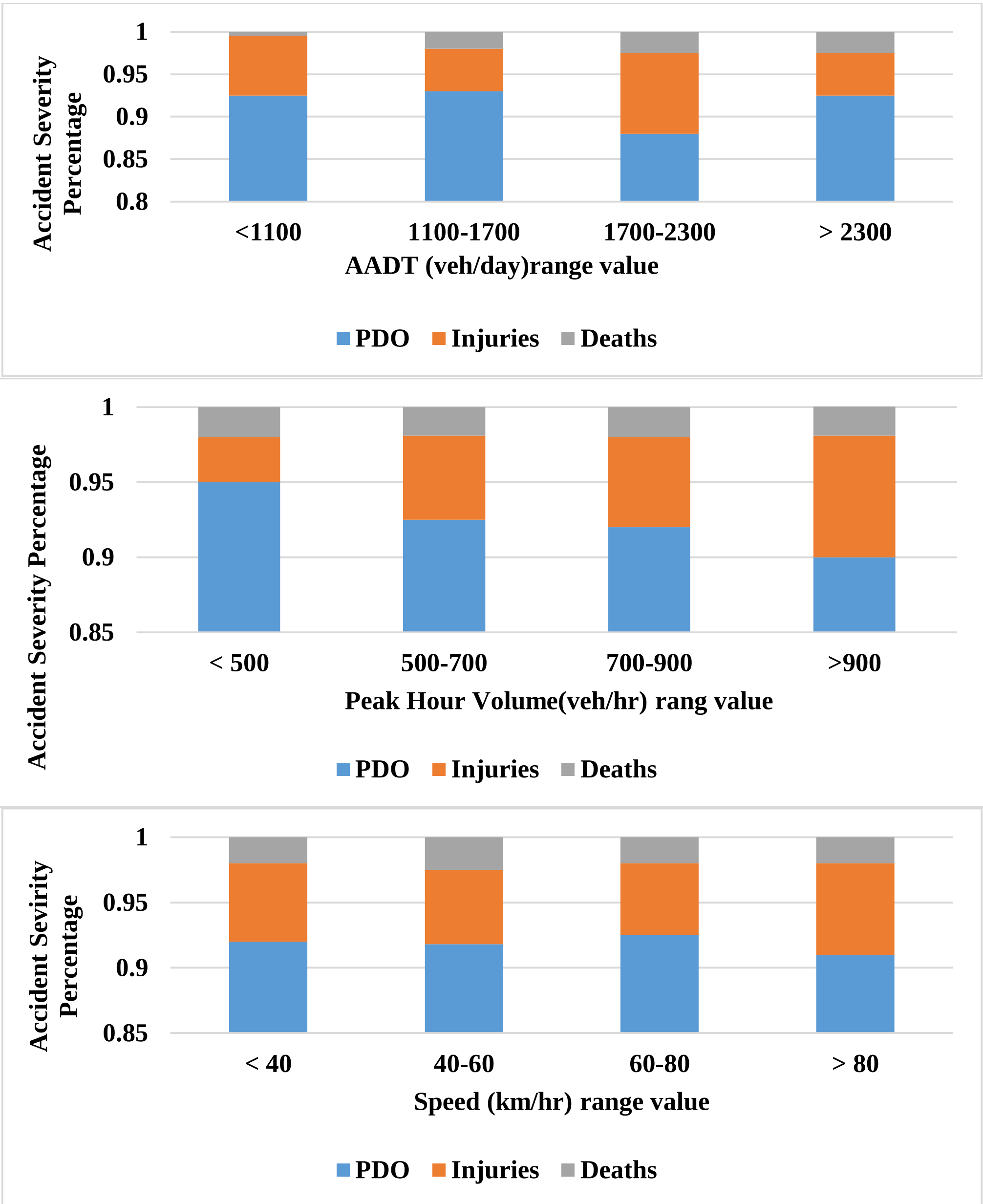

Figure (3) Bar plots showing the significance of traffic flow features and the percentages of different severity levels in rural data

The data imbalance was successfully overcome by down sampling, which managed to produce a more balanced dataset of 1233 records for PDO, 1233 for injuries, and 1121 for fatalities. The performance metrics for the machine learning classifiers on accident data in rural areas are compared.

Performance comparisons with different classification models, Gradient Boosting, LGBM, Random Forest, Bagging, and Extra Trees, were carried out against imbalanced accident data and hyper parameter tuning. Tables 7 and 8 present the performance measurements of the different classifiers on rural data before down sampling, with and without hyper-parameter adjustment.

Table 7. Rural data before down sampling (without hyper-parameter tuning)

	Accuracy	Balanced	Recall			Precision			FI-score		
			PDO	Injuries	Deaths	PDO	Injuries	Deaths	PDO	Injuries	Deaths
Gradient Boosting	0.973	0.333	0.999	0.014	0.001	0.876	0.18	0.333	0.933	0.02	0.001
LGBM	0.973	0.335	0.999	0.003	0.01	0.876	0.09	0.181	0.933	0.006	0.003
Random Forest	0.973	0.346	0.986	0.041	0.024	0.879	0.22	0.203	0.93	0.071	0.044
Bagging	0.973	0.347	0.985	0.026	0.031	0.88	0.148	0.209	0.93	0.044	0.054
Extra Trees	0.973	0.349	0.9888	0.035	0.021	0.879	0.19	0.221	0.931	0.06	0.039

Table 8. Rural data before down sampling (with hyper-parameter tuning)

	Accuracy	Balanced	Recall			Precision			FI-score		
			PDO	Injuries	Deaths	PDO	Injuries	Deaths	PDO	Injuries	Deaths
Gradient Boosting	0.973	0.338	0.999	0.02	0.003	0.876	0.263	0.34	0.933	0.03	0.001
LGBM	0.973	0.344	0.99	0.004	0.014	0.878	0.154	0.181	0.931	0.007	0.035
Random Forest	0.973	0.347	0.988	0.043	0.024	0.879	0.157	0.21	0.93	0.073	0.044
Bagging	0.973	0.347	0.984	0.03	0.034	0.88	0.151	0.223	0.929	0.53	0.06
Extra Trees	0.973	0.35	0.99	0.038	0.021	0.879	0.22	0.228	0.931	0.071	0.039

The classifiers have a high accuracy of 0.973 in rural-data prediction, although this is with low balanced recall and precision. These factors are little affected by hyper-parameter tuning. In addition, PDO accident rate is higher for injuries than for deaths, suggesting intrinsic better detection before down sampling.

Tables 9 and 10 present the performance measures of different classifiers on rural data following down sampling, both with and without hyper-parameter adjustment.

Table 9. Rural data after down sampling (without hyper-parameter tuning)

	Accuracy	Balanced	Recall			Precision			FI-score		
			PDO	Injuries	Deaths	PDO	Injuries	Deaths	PDO	Injuries	Deaths
Gradient Boosting	0.973	0.81	1	0.514	0.916	0.934	0.952	0.693	0.966	0.668	0.791
LGBM	0.973	0.795	0.984	0.579	0.822	0.933	0.804	0.696	0.958	0.673	0.754
Random Forest	0.973	0.782	0.984	0.611	0.752	0.936	0.728	0.693	0.959	0.664	0.721
Bagging	0.973	0.778	0.978	0.643	0.71	0.935	0.695	0.7	0.952	0.657	0.699
Extra Trees	0.973	0.776	0.978	0.64	0.71	0.936	0.693	0.694	0.956	0.665	0.704

Table 10. Rural data after down sampling (with hyper-parameter tuning)

	Accuracy	Balanced	Recall			Precision			FI-score		
			PDO	Injuries	Deaths	PDO	Injuries	Deaths	PDO	Injuries	Deaths
Gradient Boosting	0.973	0.811	1	0.498	0.935	0.931	1	0.696	0.964	0.665	0.796
LGBM	0.973	0.821	1	0.543	0.894	0.931	0.918	0.704	0.964	0.682	0.788
Random Forest	0.973	0.813	1	0.511	0.927	0.931	0.982	0.697	0.965	0.672	0.795
Bagging	0.973	0.778	0.978	0.643	0.71	0.935	0.695	0.7	0.952	0.657	0.699
Extra Trees	0.973	0.809	1	0.504	0.921	0.922	0.987	0.697	0.959	0.668	0.792

For instance, the hyper-parameters in the Gradient Boosting classifier improved slightly in its recall after hyper-parameter tuning, allowing it to become better at predicting injuries and deaths at a more balanced recall, precision trade-off, although the overall accuracy stayed constant at 0.973. Similarly, a little bit more increased is the LGBM classier in balanced recall and precision, mainly in predicting the death class. Another classifier that gained in balanced recall and precision is the Random Forest classifier, thus achieving better class discrimination. Not much change can be noted with the Bagging classifier metrics, while the Extra Trees classifier improved a little bit in the balanced recall, mainly for injury class prediction. Generally, hyper-parameter tuning increased mainly the balanced recall and precision, suggesting more balance in class predictions, hence improving the capability of the classifier to handle imbalanced data without altering the overall accuracy.

On the other hand, all classifiers, Gradient Boosting, LGBM, Random Forest, Bagging, and Extra Trees classifiers, were able to maintain their accuracy after down sampling with hyper-parameter tuning. Gradient Boosting classifier improved the most in the balanced recall. The performance is improved by hyper-parameter tuning, more so after down sampling. Down sampling also improves the overall accuracy, especially in Gradient Boosting classifier, since it handles imbalanced classes. Down-sampling is a fine starting point toward reducing size and improving accuracy; hyper-parameter tuning has additional advantages.

Finally, it was noted that all classifiers in rural and urban areas were very accurate, with an accuracy of over 0.97, before down sampling and hyper-parameter tuning. The Gradient Boosting classifier returned the highest score for injuries in rural areas, but was the weakest in precision. Decision Tree classifier returned the highest recall and the lowest precision in urban areas. LGBM classifier returned the highest precision, but the lowest recall. Random Forest classifier returned the highest F1-score for injuries in both rural and urban settings. This is where down sampling made a huge difference in the performance of the classifier, and that was in the rural areas, whereas hyper-parameter optimization did very little.

Comparing the results to those of other studies generally reflects some important differences and similarities. Some of those studies, such as Sahin (2020) and Aldhari et al. (2022), have found that Gradient Boosting really holds up in the prediction of traffic accident severity with quite high precisions of 85.01% and 94%, respectively. Li et al. (2019) further reiterated that Gradient Boosting provided a precision value from 70.02% to 73.65%, showing reliability in forecasting accidents. These findings were consistent with our results such that Gradient Boosting shall have improved the accuracy and balanced recall after down sampling and hyper-parameter tuning. However, Bagging resulted in the highest recall and precision rates when accident prediction is made in comparison with other studies in this field which didn't put weight on the comparative advantage of Bagging. Compared to previous studies, the good performance of Random Forest classifier with respect to the avoidance of overfitting and noisy data, our results found Random Forest classifier to have the lowest F1-score with respect to injuries, suggesting a gap in its efficiency in different contexts. It was further reported that LGBM classifier returned the highest precision and recall for injuries, which deviates strongly from the general trend of existing literature focusing on Gradient Boosting and Random Forest classifiers. Much of the past research, according to Ketha and Imambi (2019) and Alqudah et al. (2019), indicated that Decision Tree classifier was highly effective, whereas our results show it to have the worst scores for both recall and precision. These differences stipulate the importance of context-specific analysis, since differences in dataset characteristics (the case of Jordan's accident data) may impact classifier performance. It is the combination of down sampling and hyper-parameter tuning that improved the classifier accuracy and balanced recall for Gradient Boosting and LGBM classifiers. Thus, both are observed to be necessary in improving model performance and correcting class imbalance in accident severity prediction.

CONCLUSIONS

This study focuses on the accident report for practical analysis, which should be comprehensively documented. The objective is to come up with the optimal model for identifying key factors influencing accidents through the use of several ML classifiers, such as LGBM, RF, Gradient Boosting, Extra Tree, and Bagging classifiers. Road surface characteristics, traffic flow characteristics, and geometric road section property conditions are significant in influencing safer driving behaviors that lower the risk of accidents. A test/training set division was followed in testing the accuracy of the models based on the dataset. The performance of different machine learning classifiers was estimated for predicting the severity of accidents in urban and rural areas. The model assessment before and after down-sampling to deal with the class imbalance on accuracy, balanced accuracy, recall, precision, and F1-score metrics with and without hyper-parameter tuning was performed based on distinguishability between PDO, injury and death cases.

The classifiers perform well in both rural and urban areas. Before down sampling, the choice of classifier depends on the prioritization of recall (Gradient Boosting for rural areas and Decision Tree for urban areas), precision (LGBM for both), or F1-score (Random Forest for both). After down sampling, the preferred classifier performed better in all precision, recall, and F1-scores than before down sampling.

The study also places much focus on class imbalance handling as a very relevant aspect of traffic accident prediction and shows how down sampling, when combined with hyper-parameter tuning, can improve classifier performance. One possible line of future studies could be other techniques in dealing with imbalanced datasets, which might further improve the predictive accuracy and reliability of traffic accident models.

Our findings demonstrated that down sampling and hyper-parameter tuning are effective means of improving the accuracy in traffic accident predictions. Down sampling had an accuracy of 0.973 in rural areas, while tuning improved performance metrics, like balanced recall and F1-score, in Gradient Boosting, LGBM, and Random Forest classifiers. Balanced precision and recall improved accuracy in the mentioned classifiers. LGBM increased recall for minority classes, while Random Forest, Bagging, and Extra Trees maintained a good balance between precision and recall. All the five classifiers: Gradient Boosting, LGBM, Random Forest, Bagging, and Extra Trees classifiers, maintained accuracy after down sampling and tuning hyper-parameters. Down sampling improved the accuracy of Gradient Boosting, LGBM, Random Forest, Bagging, and Extra Trees classifiers. In this respect, the study shows that among these classifiers, the Decision Tree classifier has the highest recall and precision. All classifiers gave an accuracy above 0.97 with a detection rate of PDO accidents from 0.972 to 0.973. This was a very important impact in accident severity classifiers.

Several limitations may be due to underreporting, data quality, class imbalance, geographic and temporal coverage, external factors, and model interpretability. Poor or missing data might also cause biased models. Accident severity balance will have a negative effect if one class is dominant over another in the dataset, which possibly badly affects the predictive ability of the model for low-frequency events and temporal coverage can also be a limiting factor when generalizing the findings. Other external factors are driver behavior and enforcement of traffic laws, which might not be fully represented in the dataset. This can then become a challenge to model interpretability in identifying the key factors influencing accidents.

The authors are alive to the limitation created by 2022 data and go ahead to recommend ways of improving models used in the prediction of traffic accidents. They further propose regular updating, cross-validation methods, and exogenous variables. Sensitivity analysis and comparative analysis with multi-year studies should also be conducted to identify informative variables and address anomalies.

The authors also recommend incorporating variables of temporal patterns, driver behavior, demographic data, motorcycle *vs.* car, and geographic location into the models of accident prediction to make them more efficient and stronger. Other factors, like speeding, consumption of alcohol, and distraction, may also contribute to a perfect modeling system. This approach refines the accuracy and efficiency of road safety measures.

REFERENCES

Afshar, F., Seyedabrishami, S., and Moridpour, S. (2022). "Application of extremely randomised trees for exploring influential factors on variant crash severity data". Scientific Reports, 12 (1), 11476.

Aldhari, I., Almoshaogeh, M., Jamal, A., Alharbi, F., Alinizzi, M., and Haider, H. (2022). "Severity prediction of highway crashes in Saudi Arabia using machine learning techniques". Applied Sciences, 13 (1), 233.

Al-Masaeid, H. (2009). "Traffic accidents in Jordan". Jordan Journal of Civil Engineering, 3 (4), 331-343.

Al-Mistarehi, B.W., Alomari, A.H., Imam, R., and Mashaqba, M. (2022). "Using machine learning models to forecast severity level of traffic crashes by R Studio and ArcGIS". Frontiers in Built Environment, 8, 860805.

Al-Omari, B., Ghuzlan, K., and Hasan, H. (2013). "Traffic accident trends and characteristics in Jordan". International Journal of Civil & Environmental Engineering IJCEE-IJENS, 13 (05), 9-16.

Alqudah, A.M., Alquraan, H., Qasmieh, I.A., Alqudah, A., and Al-Sharu, W. (2019). "Advanced trends in computer science and engineering". International Journal, 8 (6).

Al-Rousan, T.M., Umar, A.A., and Al-Omari, A. A. (2021). “Characteristics of crashes caused by distracted driving on rural and suburban roadways in Jordan”. Infrastructures, 6 (8), 107.

Ampomah, E.K., Qin, Z., and Nyame, G. (2020). “Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement”. Information, 11(6), 332.

Baykal, T., Ergezer, F., Eriskin, E., and Terzi, S. (2023). “Accident severity prediction in big data using auto-machine learning. Scientia Iranica.

Belkebir, R., and Guessoum, A. (2015). “A supervised approach to Arabic text summarization using ADABoost”. New Contributions in Information Systems and Technologies, 1.

Berhanu, Y., Alemayehu, E., and Schröder, D. (2023). “Examining car accident prediction techniques and road traffic congestion: A comparative analysis of road safety and prevention of world challenges in low-income and high-income countries”. Journal of Advanced Transportation, 2023.

Dogru, N., and Subasi, A. (2018). “Traffic accident detection using random forest classifier”. 2018 15th Learning and Technology Conference (L&T).

Guo, M., Yuan, Z., Janson, B., Peng, Y., Yang, Y., and Wang, W. (2021). “Older pedestrian traffic crash severity analysis based on an emerging machine learning XGBoost”. Sustainability, 13 (2), 926.

Hala, H., Anass, C., Rajaa, B., Youssef, B., and Garza-Reyes, J. (2021). “Machine learning techniques for forecasting the severity of traffic accidents”. 2021 International Conference on Digital Age & Technological Advances for Sustainable Development (ICDATA).

Hareru, H.E., Negassa, B., Kassa Abebe, R., Ashenafi, E., Zenebe, G.A., Debela, B.G., ..., and Eshete Soboksa, N. (2022). “The epidemiology of road traffic accidents and associated factors among drivers in Dilla town, southern Ethiopia”. Frontiers in Public Health, 10, 1007308.

Hussain, S.F., and Ashraf, M.M. (2023). “A novel one-*vs.*-rest consensus learning method for crash severity prediction”. Expert Systems with Applications, 228, 120443.

Jamal, A., Zahid, M., Tauhidur Rahman, M., Al-Ahmadi, H.M., Almoshaogeh, M., Farooq, D., and Ahmad, M. (2021). “Injury severity prediction of traffic crashes with ensemble machine learning techniques: A comparative study”. International Journal of Injury Control and Safety Promotion, 28 (4), 408-427.

Jiang, M., Liu, J., Zhang, L., and Liu, C. (2020). “An improved stacking framework for stock index prediction by leveraging tree-based ensemble models and deep learning algorithms. Physica A: Statistical Mechanics and Its Applications, 541, 122272.

Ketha, T., and Imambi, S.S. (2019). “Analysis of road accidents to indentify major causes and influencing factors of accidents: A machine learning approach”. International Journal of Advanced Trends in Computer Science and Engineering, 8 (6), 3492-3497.

Kunwar, M. (2019). “Artificial intelligence in finance: Understanding how automation and machine learning transform the financial industry”.

Lee, J., Yoon, T., Kwon, S., and Lee, J. (2019). “Model evaluation for forecasting traffic accident severity in rainy seasons using machine learning algorithms: Seoul city study”. Applied Sciences, 10 (1), 129.

Li, Y., Yang, L., Yang, B., Wang, N., and Wu, T. (2019). “Application of interpretable machine learning models for the intelligent decision”. Neurocomputing, 333, 273-283.

Marcillo, P., Valdivieso Caraguay, Á.L., and Hernández-Álvarez, M. (2022). “A systematic literature review of learning-based traffic accident prediction models based on heterogeneous sources”. Applied Sciences, 12 (9), 4529.

Miqdady, T., and de Oña, J. (2020). “Identifying the factors that increase the probability of an injury or fatal traffic crash in an urban context in Jordan”. Sustainability, 12 (18), 7464.

Sahin, E.K. (2020). “Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest”. SN Applied Sciences, 2 (7), 1308.

Shen, X., and Wei, S. (2020). “Application of XGBoost for hazardous material road transport accident severity analysis”. IEEE Access, 8, 206806-206819.

Wu, S., Yuan, Q., Yan, Z., and Xu, Q. (2021). “Analyzing accident injury severity *via* an extreme gradient boosting (XGBoost) model”. Journal of Advanced Transportation, 2021, 1-11.