ICIC Express Letters ICIC International ©2022 ISSN 1881-803X
Volume **16**, Number **12**, December **2022** pp. 1301–1307

ENSEMBLE AND BASE LEARNER ML TECHNIQUES FOR TRAFFIC ACCIDENT SEVERITY PREDICTION: A COMPARISON STUDY

Mohammed Akour[1,*], Osama Al Qasem[2], Mamdouh Alenezi[1]
Abeer Bataineh[2] and Feras Hanandeh[3]

[1]Software Engineering Department
Prince Sultan University
Riyadh 11586, Saudi Arabia
malenezi@psu.edu.sa
*Corresponding author: makour@psu.edu.sa; mohammed.akour@yu.edu.jo

[2]Information Systems Department
Yarmouk University
Irbid 21163, Jordan
osamahalqasem@yahoo.com; abeer.bataineh@yu.edu.jo

[3]Computer Information System
Hashemite University
Zarqa 13133, Jordan
feras@hu.edu.jo

Received March 2022; accepted May 2022

Abstract. *Based on several reports, one of the main causes of human injuries and death is traffic accidents. Many communities are suffering from the accidents at different levels of severity. Traffic accident severity prediction might play a role in enhancing the management and controlling the safety of traffic. By utilizing existing road accident data, more accuracy of accident severity prediction can be performed. This research paper aims to build an accurate traffic accident severity prediction model. The proposed model is mainly based on ensemble and base learner machine learning algorithms, i.e., Random Forest, XGBoost and decision tree. For comparison purposes, the performance of the studied ensemble methods is compared with the base learners. Five measurements are recorded and used for comparison. The findings of this paper show that Balanced Random Forest, XGBoost and decision tree provide a promising tool for predicting the injury severity of traffic accidents. Moreover, the voting (hard) has an advantage over the other two representative classifiers. Compared with other classifiers, voting (hard) has a good ability to predict fatal/serious injury.*

Keywords: Traffic accident severity prediction, Injuries, Base learner algorithms, Ensemble algorithms

1. Introduction.

Traffic accidents are considered as the main source of daily injury and death. These accidents cause property losses at two economic and social levels. Accident prediction and traffic safety assessment is playing a crucial role in building an effective traffic safety policy that led to reducing the rate of traffic accidents and losses [1]. Too many studies report the adverse influence of traffic accidents on countries' economies, traffic jams, environment pollution and the worst impact was human death. Although of the noticeable growth of the smart transportation systems that are produced by researchers and governments, traffic accident prediction is still a big challenge for these systems [2]. Several factors should be addressed to assess and enhance the existing traffic policies. Traffic accident severity prediction could participate in enhancing the management and

DOI: 10.24507/icicel.16.12.1301

controlling the safety traffic. By utilizing existing road accident data, more accuracy of accident severity prediction can be performed.

World Health Organization (WHO) [11] showed that more than 1.3 million died because of traffic accidents while around 50 million suffered from non-fatal injuries. As a conclusion of their study, the report reveals that traffic accidents are placed in the ninth cause of death worldwide. The traffic accidents can happen at any moment during the day, but if there is a system that can help in predicting these accidents and severity, then the harm might be prevented or at least minimized their impact. Studying the factors that cause the accidents can help in navigating and predicting the accidents severity. Several researchers addressed these factors and their relationship with the accident severity. They try to predict the accident severity by utilizing different techniques and mechanisms on existing traffic data. Machine learning classifiers were one of the techniques used to predict the traffic accident severity [4-8]. The main contribution of this study is to address the effectiveness of using ensemble learning methodology with respect to base learner performance in traffic severity prediction. The studied ML performance is evaluated by calculating 5 measurements, i.e., accuracy, recall, precision, true negative rate and true positive rate. Twelve classifiers were used, and the results were recorded. The comparison results reveal that the voting achieved highest performance among stacking models and other individual classifiers. The remainder of the paper is organized as follows: Section 2 discusses the literature related to the prediction traffic accident severity and its influencing factors; Section 3 explains the research methodology used; Section 4 discusses the experimental work and results; finally, Section 5 provides a conclusion of the work.

2. **Related Works.** Various studies have employed several methodologies to discover the relationship between traffic accident severity and its influencing factors. Moreover, these techniques are used to predict the severity of the accidents after determining a set of important factors. In this section some results of the previous works in accident severity prediction are presented. Wahab and Jiang [10] studied the effectiveness of using four machine learning in the field of predicting the motorcycle crash severity, i.e., decision tree, J48, instance-based learning and Random Forest (RF). This study is conducted on Ghana traffic data in 2019. The best accuracy is achieved by Random Forest (RF) with 73.91%. The cause of the accidents is very important to determine its severity. Mohammed [7] in 2014, used traffic data in Dubai to predict the causes of traffic accidents. Several ML techniques were used, and the results were compared. The best accuracy was around 75%. The study shows that the most frequent causes of road traffic in Dubai were neglecting other vehicles on the road or over-speeding. Jamal et al. [5] performed a comparative study between four ML techniques. The performance of the eXtreme Gradient Boosting (XGBoost) technique outperforms all other studied algorithms. Although the XGBoost is a new algorithm, it has achieved 95% accuracy which is considered a high rate in comparison with the body of art in this field. Although this prediction is mainly conducted on Riyadh city traffic data within two years, this is still a promising result, and the model could be used on other data sets to generalize the result.

Investigating the influence of key factors that mainly cause traffic accidents is very important. AlMamlook et al. [1] like other studies addressed these impacts and developed a model to predict the accident traffic severity. Five ML are used to build their model, and accuracy was the mainly used measurement to assess the effectiveness of these algorithms. The results show that Random Forests model outperforms other studied algorithms in predicting the severity of traffic accidents. The highest accuracy was 75.5% by RF algorithms. Gan et al. [4] predicted the traffic accident severity based on the Deep Forests algorithm. Their model employed the Deep Forests algorithm, while a dataset from United Kingdom road traffic is used to evaluate their proposed prediction model. In this study, the performance of the proposed model is compared with other addressed ML algorithms. The results show the superiority of the proposed model as it shows promising stability. Yassin [13] before developing the proposed model, tried to address and extract the most significant influencing factors for accident severity prediction. To achieve this target, Random Forest and Hybrid K-means approaches are developed. The developed technique is mainly evaluated in comparison with deep neural networks. Based on the comparison result, the proposed approach outperforms the studied classifiers in terms of prediction accuracy.

3. **Research Methodology.** Machine Learning (ML) techniques are applied in this paper for the purpose of crash severity prediction. ML algorithms used in classifying datasets can produce promising results due to their flexibility in implementation, multi-dimensional data processing capability. Our solution will follow a framework consisting of three main steps: preparing the data, features selection, and classification executed in order to obtain the severity prediction. The methodology will be further discussed in the following sections. Figure 1 illustrates the main steps of the proposed research methodology.

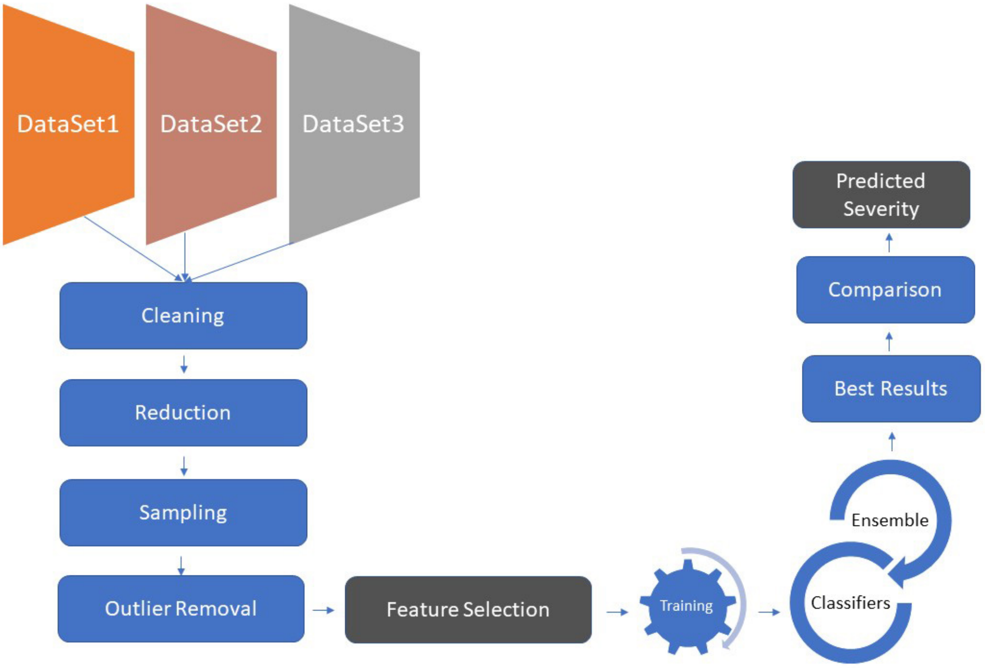

Figure 1. Research methodology

3.1. **Dataset.** For the accurate prediction of the crash severity, a huge number of accident records with detailed information are needed to be applied on the proposed approaches. In this work, the dataset collected from the Leeds City Council consists of a total 27,540 traffic road accidents recorded from the year 2009-2019 in England. Data includes location, number of vehicles and people involved, weather and lightning conditions, road surface and severity of any casualties. Table 1 shows samples from Leeds dataset.

The traffic data recorded has two types of injury severities. The fatal and serious injury (0) is an injury where the casualty died, or the victim of the crash was admitted at the hospital for medical attention. Whereas the slight injury (1) is a victim that was admitted at the hospital for less than 24 hours. The features and their descriptions are presented in Table 2.

3.2. **Preprocessing.** Some preprocessing is an important step in which raw data is processed in a way that the system can understand it efficiently before applying ML algorithms. Real data is generally incomplete and missing values. For that, preprocessing is

Table 1. Dataset sample

Class	0	1	0	0
Easting	429093	434723	441173	428487
Northing	436258	435534	433047	431364
Number of vehicles	1	1	1	1
Time (24hr)	55	2335	1645	1723
1st Ro3d 5l3ss	6	6	6	3
Road surface	1	1	1	1
Lighting conditions	4	4	4	4
Weather conditions	1	1	1	1
Casualty class	3	1	3	3
Sex of casualty	1	2	2	1
Age of casualty	44	23	12	15
Type of vehicle	9	9	9	9

a crucial phase to solve these problems and improve the quality and accuracy of the data. During this step, raw data is transformed into a dataset for knowledge discovery. The data preprocessing stages include the following.

– Data Integration: in this step merging 11 datasets into a single and combined view.
– Data Cleaning: Real data has the tendency to be incomplete, noisy and uncertain. Data cleaning aids in filling the missing values, smooths out noise and detects outsider and precise unpredictability in the data.
– Reduction: consist of removal of few repeating data, dimensionality reduction and aggregation.
– Sampling: is a method of converting data from several similar samples into a single labelled dataset in order to reduce the amount of variation in the datasets in Leeds dataset, the sampling becomes as described in Table 3.

3.3. **Outlier removal.** Outlier is an anomaly, abnormalities or discordance, also known as an observation which differs so much from other observations. Outlier detection is to find those anomalies data and remove it. There are various outlier removal techniques available, in this paper using Zscore based outlier detection.

3.4. **Feature selection.** Feature selection is a useful method that enhances the performance of the model, by removing inconsistent, irrelevant, and redundant features. This helps in reducing the computational time and complexity of the model. Therefore, in the Leeds dataset eliminating four attributes does not affect the accuracy of the model.

3.5. **Model training.** The main reason to conduct this study is to evaluate the performance of ensemble algorithms and compare it with individual classifiers. The algorithm frequency utilized in the previous study is dependent on selection of base classifier. This phase involves training of the classifier using the Leeds dataset. In this phase, various algorithms were analyzed and selected based on their accuracy, recall, precision, true positive rate and true negative rate score. The algorithms applied and the results are shown in Table 4.

3.6. **Ensemble learning.** Final stage is using the stacking and voting ensemble method to predict the severity. Balanced Random Forest, XGBoost and decision tree are the algorithms selected for creating stacking and (hard and soft) voting models. The ensemble learning model is created as illustrated in Figure 2.

Table 2. Dataset feature description

Feature	Description
Road class	1 Motorway 2 A(M) 3 A 4 B 5 C 6 Unclassified
Road surface	1 Dry 2 Wet/Damp 3 Snow 4 Frost/Ice 5 Flood (surface water over 3cm deep)
Lighting conditions	1 Daylight: street lights present 2 Daylight: no street lighting 3 Daylight: street lighting unknown 4 Darkness: street lights present and lit 5 Darkness: street lights present but unlit 6 Darkness: no street lighting 7 Darkness: street lighting unknown
Weather conditions	1 Fine without high winds 2 Raining without high winds 3 Snowing without high winds 4 Fine with high winds 5 Raining with high winds 6 Snowing with high winds 7 Fog or mist – if hazard 8 Other
Casualty class	1 Driver or rider 2 Vehicle or pillion passenger 3 Pedestrian
Type of vehicle	1 Pedal cycle 2 M/cycle 50cc and under 3 Motorcycle over 50cc and up to 125cc 4 Motorcycle over 125cc and up to 500cc 5 Motorcycle over 500cc 8 Taxi/Private hire car 9 Car 10 Minibus (8-16 passenger seats) 11 Bus or coach (17 or more passenger seats) 14 Other motor vehicle 15 Other non-motor vehicle 16 Ridden horse 17 Agricultural vehicle (includes diggers etc.) 18 Tram/Light rail 19 Goods vehicle 3.5 tonnes and under 20 Goods vehicle over 3.5 tonnes and under 7.5 tonnes 21 Goods vehicle 7.5 tonnes and over 22 Mobility scooter 90 Other vehicle 97 Motorcycle – Unknown CC

Table 3. Feature sampling

	Feature			
	Road surface	Lighting conditions	Weather conditions	Type of vehicle
Description	1 Dry 0 Not dry	1 Daylight 0 Darkness	1 Fine 2 Raining 3 Snowing 4 Other	1 Pedal cycle/Motorcycle 2 Taxi/Car 3 Larger vehicle

Table 4. Algorithms prediction performance

Classifier	Accuracy	Recall	Precision	TNR	TPR
LogisticRegression	.874	1	.874	0	1
GaussianNB	.863	.979	.874	.055	.979
k-nearest neighbor	.863	.979	.878	.061	.979
RandomForest	.870	.985	.880	.068	.985
DecisionTree	.792	.873	.887	**.227**	.873
AdaBoost	.873	.997	.875	.013	.978
CNN	.874	1	.874	0	1
CNN-LSTM	.874	1	.874	0	1
CascadeForest	.87	.873	.873	.027	.996
BalancedRandomForest	.783	.79	.81	**.320**	.854
XGBoost	.869	.98	.883	**.112**	.98
LGBM	.871	.992	.876	.053	.992

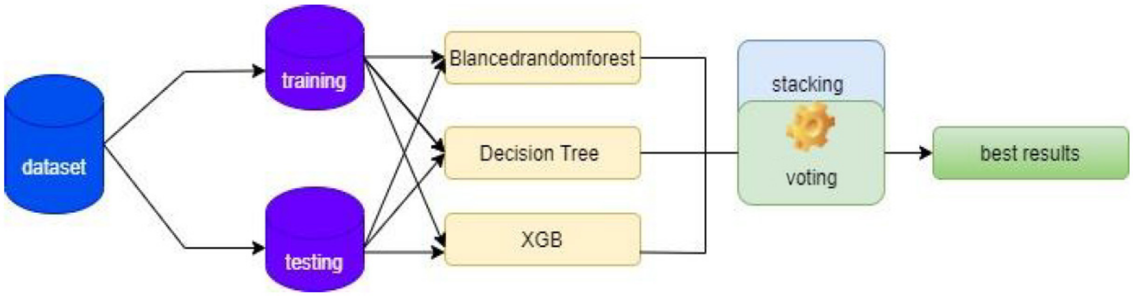

Figure 2. Ensemble learning

Table 5. Ensemble algorithms performance

Classifier	Accuracy	Recall	Precision	TNR	TPR
Voting (hard)	.831	.906	.90	**.330**	.906
Voting (soft)	.826	.904	.897	.289	.904
Stacking	.872	.995	.875	.029	.995

4. **Result and Discussion.** This section presents and discusses the experiments and the results for the different ensemble algorithms. Comparisons and exploration were discussed to see which model provides the best prediction for traffic accident severity. We evaluated the performance of the models using accuracy, recall, precision, true positive rate and true negative rate. Evaluation measures for each model are summarized in Table 5.

In this study authors adopt the specificity (TNR) to make the prediction performance comparison, for example, when the test-data that identifies all persons as being negative for a particular injury is very specific. As shown in Table 5, the voting (hard) ensemble method achieves the highest TNR among all the others. Therefore, there is no doubt that the voting (hard) has an advantage over the other two representative classifiers. Compared with other classifiers, voting (hard) has a good ability to predict fatal/serious injury.

5. **Conclusion and Future Work.** The analysis of road accident severity is a promising research area. The present study investigated the efficiency of the three ML classifiers creating ensemble methods to build reliable classifiers. This includes Balanced Random Forest, XGBoost and decision tree. The test results show that the voting seemed to perform better than stacking models and other individual classifiers. In the future work, the authors aim to search for a bigger dataset where the key factor causing traffic crashes can be studied, and the performance of the algorithms can be compared. The main limitation of this study is that the addressed dataset may not contain all important factors such as passenger information, and traffic conditions, which may impact the accident severity.

Acknowledgements. The authors would like to acknowledge the support of Prince Sultan University for paying the Article Processing Charges (APC) of this publication.

REFERENCES

[1] R. E. AlMamlook, K. M. Kwayu, M. R. Alkasisbeh and A. A. Frefer, Comparison of machine learning algorithms for predicting traffic accident severity, *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, Amman, Jordan, pp.272-276, 2019.

[2] B. Dadashova, B. A. Ramírez, J. M. McWilliams and F. A. Izquierdo, The identification of patterns of interurban road accident frequency and severity using road geometry and traffic indicators, *Transportation Research Procedia*, vol.14, pp.4122-4129, 2016.

[3] J. De Oña, G. López, R. Mujalli and F. J. Calvo, Analysis of traffic accidents on rural highways using latent class clustering and Bayesian networks, *Accident Analysis and Prevention*, vol.51, pp.1-10, 2013.

[4] J. Gan, L. Li, D. Zhang, Z. Yi and Q. Xiang, An alternative method for traffic accident severity prediction: Using deep forests algorithm, *Journal of Advanced Transportation*, vol.2020, 1257627, 2020.

[5] A. Jamal, M. Zahid, M. Tauhidur Rahman, H. M. Al-Ahmadi, M. Almoshaogeh, D. Farooq and M. Ahmad, Injury severity prediction of traffic crashes with ensemble machine learning techniques: A comparative study, *International Journal of Injury Control and Safety Promotion*, vol.28, no.4, pp.408-427, 2021.

[6] X. H. Meng, L. Zheng and G. M. Qin, Traffic accidents prediction and prominent influencing factors analysis based on fuzzy logic, *Journal of Transportation Systems Engineering and Information Technology*, vol.9, no.2, pp.87-92, 2009.

[7] E. A. Mohamed, Predicting causes of traffic road accidents using multi-class support vector machines, *Journal of Communication and Computer*, vol.11, no.5, pp.441-447, 2014.

[8] R. O. Mujalli, G. López and L. Garach, Bayes classifiers for imbalanced traffic accidents datasets, *Accident Analysis and Prevention*, vol.88, pp.37-51, 2016.

[9] Z. Pu, Z. Li, Y. Jiang and Y. Wang, Full Bayesian before-after analysis of safety effects of variable speed limit system, *IEEE Transactions on Intelligent Transportation Systems*, vol.22, no.2, pp.964-976, 2020.

[10] L. Wahab and H. Jiang, A comparative study on machine learning based algorithms for prediction of motorcycle crash severity, *PLoS One*, vol.14, no.4, e0214966, 2019.

[11] *WHO|Road Traffic Injuries*, WHO, 2017.

[12] Z. Yan, X. Lu and W. Hu, Analysis of factors affecting traffic accident severity based on heteroskedasticity ordinal Logit, *ICTE 2019*, Chengdu, China, pp.422-435, 2020.

[13] S. S. Yassin, Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach, *SN Applied Sciences*, vol.2, no.9, pp.1-13, 2020.

[14] X. F. Zhang and L. Fan, A decision tree approach for traffic accident analysis of Saskatchewan highways, *2013 26th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, Regina, Saskatchewan, Canada, pp.1-4, 2013.