Car Accident Severity Classification Using Machine Learning

Abdulrahman Atwah
Computer Engineering Department
Princess Sumaya University for Technology
Amman, Jordan
Abd20160508@std.psut.edu.jo

Amjed Al-Mousa
Computer Engineering Department
Princess Sumaya University for Technology
Amman, Jordan
a.almousa@psut.edu.jo

2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT) | 978-1-6654-4032-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/3ICT53449.2021.9581646

Abstract— **Car accidents have always been a terrible and extremely dangerous phenomenon. It caused the loss of many lives. The delay of the needed medical treatment for injuries at accident locations puts lives at risk. In this work, machine learning was used to predict the severity of accidents that occurred in the United Kingdom between the years 2005 - 2014. The combination of this AI solution and other systems to report to relevant authorities when accidents occur will preserve more lives. The medical support that will reach the accident location will depend on the severity of the accident. Several machine learning models were used, including Support Vector Machine (SVM), Artificial Neural Network (ANN), and Random Forest (RF). The best accuracy has been achieved was using the RF model with an accuracy of 83.9 %.**

Keywords—Machine learning; Classification; car accidents; Support Vector Machine (SVM); Artificial Neural Network (ANN); Random Forest (RF).

I. INTRODUCTION

Car accidents are considered one of the most dangerous phenomena around the world. According to World Health Organization (WHO), approximately 1.3 million people around the world die due to traffic accidents every year [1]. This number of losses is affected by many factors, such as speeding, damaged road infrastructure, and late medical treatment. This statistic reflects the seriousness of this dilemma and calls on society to study this phenomenon in detail and try to determine the reasons behind these accidents.

This work studies UK traffic and tries to analyze the causes of traffic accidents. The study is based on a dataset collected by the UK government [2], for accidents that occurred between the years 2005-2014, in which many features related to humans and vehicles were collected such as the number of casualties and number of damaged cars. Other features are related to the accident's environment such as light conditions, speed limit, and road surface conditions.

The use of machine learning to study and analyze the mentioned phenomenon is an effective way to get excellent results and to highlight common patterns in dangerous accidents, such as accident locations and speed limits. By working on those patterns and trying to improve them, we could save lives and reduce expenses.

The followed approach to predict the accident severity was to produce more than one supervised classification model such as Support Vector Classifier (SVC), Random Forests Classifier (RFC), and short-list the promising models (the ones with the best results) to be used in the ensemble learning technique. The ensemble technique combines multiple classifier models to get more accurate results by applying the hard-voting technique that chooses the class based on the highest number of votes as determined by each model.

Section II of this work shares what has been reported in the literature in applying machine learning on traffic datasets. Section III explains the steps followed to prepare and clean up the dataset and to visualize it by using different charts. Section IV shows the used models and the best hyperparameter values found using the grid search. Section V shows the best-reached results and shows the feature scaling over the accuracy graph. And finally, section VI includes the conclusion for this work.

II. LITERATURE REVIEW

Machine learning has been used to solve many classification problems like how likely is it to be admitted to a certain university [3], which books you might like based on previous reading history [4], or predicting who wrote a certain tweet [5]. In the domain of predicting the car crash severity, many studies share similar objectives and managed to achieve comparable results, although the studies subsequently mentioned used different datasets. The used dataset in [6] depends primarily on hybrid feature vectors. These features are extracted by applying some image processing techniques such as Hu moments, Histograms of Oriented Gradients (HOG), and Local Binary Pattern (LBP) over the accident images. The results of [6] show that the best result it achieved was 75% using the Random Forest classifier.

The authors of [7] used a dataset collected from the Accident Research Institute of BUET for the accidents that happened in Bangladesh between the years 2001-2015. The dataset contains many similar features to this work such as Traffic control, Weather, and Light Conditions. The accident severity values used were unique, these are Fatal, Grievous, Simple Injury, and Motor Collision. A set of classifier models were used to predict the output such as K-Nearest Neighbors (KNN), Decision Tree, Naïve Bayes, and AdaBoost. The best accuracy the authors achieved was using AdaBoost classifier with 80%.

The authors of [8] made a comparison between supervised models to predict the traffic accident severity. The work was applied over a dataset that collects information for traffic accidents that occurred in Michigan from 2010-2016. Many supervised models were used such as Nanoforest, Logistic Regressor, and Naïve Bayes. The Random Forest managed to achieve an accuracy of 75.5%.

The main objective of this work is to predict the severity of the accidents that occurred in the UK. The best models mentioned in [6], [7], and [8] will be used individually and in the ensemble learning approach to compare which model will have the best accuracy.

III. Experimental Setup

The used dataset was downloaded from Kaggle website. It contains a lot of open-source datasets that can be used for regression and classification problems. The dataset used to predict the accident's severity was named Accidents0514. However, another two related datasets named Casualties0514 and Vehicles0514 were used to extract some related statistics. The Casualties0415 dataset contains information related to casualties in each accident such as casualty age and gender, while the Vehicle0415 dataset contains information related to damaged vehicles in each accident such as the age of the vehicle, engine capacity (CC), the driver's age, and gender.

The Accidents0514 dataset was collected by the UK government using the STATS19 accident reporting form. This form includes the features used in the classification process. A set of steps was followed to set up the environment and prepare the data to satisfy the classification objective. The following subsections explain these steps in detail.

Table 1: Features Information

Feature Name	Type	Correlation
Accident Severity (Label)	Discrete	1.00
Police Officer Attendance	Discrete	0.121
Number of Vehicles	Discrete	0.074
Junction Control	Discrete	0.070
2nd Road Class	Discrete	0.066
Junction Detail	Discrete	0.047
2nd Road Number	Discrete	0.028
Weather Conditions	Discrete	0.025
1st Road Class	Discrete	0.017
Road Surface Conditions	Discrete	0.012
Pedestrian Control	Discrete	0.006
Pedestrian Facilities	Discrete	0.004
Special Conditions at Site	Discrete	0.004
Longitude	Continuous	0.002
Day of Week	Discrete	0.002
Location Easting OSGR	Continuous	0.002
Carriageway Hazards	Discrete	0
1st Road Number	Discrete	-0.001
Local Authority (District)	Discrete	-0.024
Police Force	Discrete	-0.026
Latitude	Continuous	-0.029
Location Northing OSGR	Continuous	-0.029
Road Type	Discrete	-0.038
Light Conditions	Discrete	-0.069
Speed limit	Continuous	-0.081
Urban or Rural Area	Discrete	-0.082
Number of Casualties	Discrete	-0.098

A. Exploring Data

The Accidents0514 is the dataset that will be explored in detail in this section, the Casualties0415, and Vehicle0415 datasets will be mentioned later to show some related statistics.

As a start, some information about the dataset's features should be known. The Accident Severity is the classification problem (the label) for this work. The possible accident severity values are fatal (represented by 1), serious (represented by 2), and slight (represented by 3). The correlation coefficients were calculated for all the features with the label (the Accident Severity). Table1 shows the features' names, types, and their correlation coefficient values.
The shaded cells from Table 1 are the features that have been excluded from the Accident0415 dataset, the *Preprocessing Data* section discusses the reasons for dropping these features in detail.

As previously mentioned, the UK dataset contains many features. Having longitude, latitude, and accident severity features provided the author to get deeper inside the dataset and opened a way to estimate the diversity of the accidents' locations based on the aforementioned features. Figure 1 shows the distribution of the accidents that occurred across the UK.

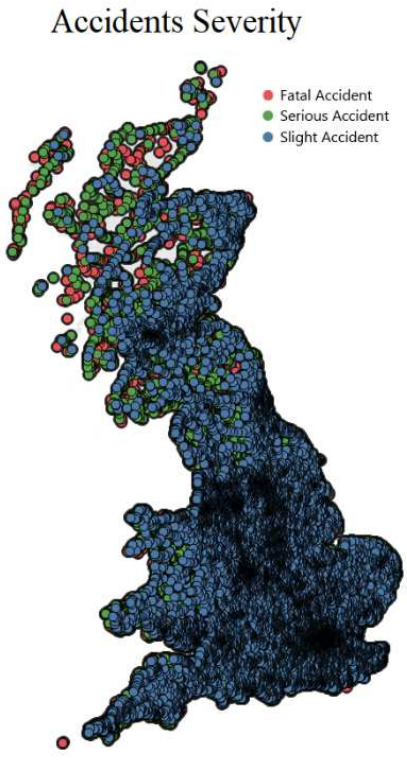

Figure 1: Longitude vs Latitude for accidents depending on the severity.

The intersections of accidents' longitude and latitude features look very similar to the UK's map. The severity for each accident is represented by a unique color, red means fatal, green means serious, and blue means slight.

The dataset also included important information such as the gender of the casualties and the gender of the drivers. It was found that the number of males is greater in both cases than the number of females. Figure 2 and Figure 3 show the percentages of males and females of the casualties and drivers.

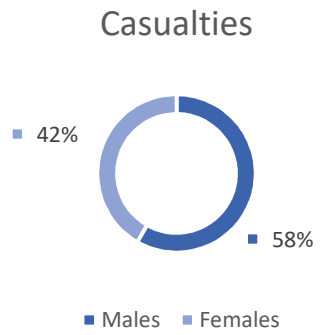

Figure 2: Male vs Female casualties' percentages.

Figure 2 shows the gender percentage of casualties in the UK for the years 2005 - 2014. Males percentage (58%) is greater than the female percentage (42 %).

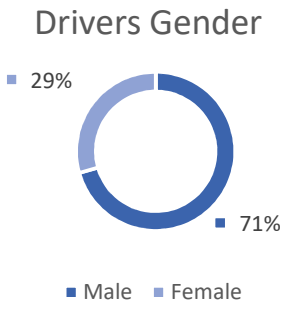

Figure 3: Male vs Female drivers' percentages.

Figure 3 shows the gender percentages of drivers in the UK. It shows that the male percentage (71%) is greater than the female percentage (29 %).

B. Preprocessing Data

This section includes a sequence of steps that prepares the Accident0514 dataset to be used for training, validating, and testing the machine learning models.

Data cleaning is considered an initial step in preprocessing. The cleaning process was first started by checking the number of null values. It was found that the number of these values equals 111. This problem was solved by dropping the rows with the null values. Another problem related to data cleaning, discovered after checking the lookup table (the lookup table describes the possible values for each feature), was the -1 values. The cells that were lost while filling in the accident tables were filled with -1 values. This has resulted in adding an extra task to resolve the issue of the -1 values. This problem was solved by removing the rows in which the -1 values were few, while the features with many records of -1 values were excluded from the dataset.

This approach was followed because deleting rows of features that have many -1 values would have resulted in deleting a significant number of rows. The gray-colored cells in Table 1 represent the features that were dropped due to the -1 values and the low correlation index.

Obtaining the massive number of instances (1,048,576) was very helpful for increasing the accuracy while generating Figure 1. However, after delving deeper into the problem, it appears that the number of casualties for each class is important and will affect the Machine Learning models' behaviors. Having a skewed dataset will cause difficulty for models while classifying the accident's severity.

Since the difference between class casualties is enormous, the models tend to classify the new instances as the most common class most of the time. Thus, the Accidents0514 dataset is considered a skewed dataset. Figure 4 shows the number of casualties for each severity class (Slight, Serious, and Fatal).

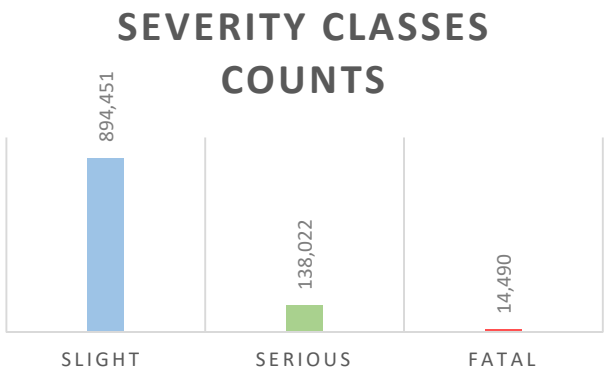

Figure 4: Number of casualties for each class

The followed solution for this skewness was by taking an equal number of instances for each severity class. Since fatal accidents were the fewest (14,490) in the dataset, an equal number of instances for both slight and serious accidents were taken. This approach has resulted in having a dataset with 43,470 records.

Having an equal number of instances from each class has significantly affected the correlation of the features. Figure 5 shows the new correlation index for each feature.

Since all categorical features are coded with numbers rather than containing textual strings, the one-hot encoding technique was applied over the nominal categorical features while the ordinal categorical features were left as is. Using this technique increased the number of features to 58 columns. Having this number of columns is not considered a huge number compared to the famous MINST dataset [9] ($784 \times 70{,}000$).

The features' values of the dataset range along with different scales. The difference may cause an issue over the accuracy of the trained models when performing the prediction process. This problem can be solved using feature scaling. In this work, both feature normalization (shown in equation 1.1) and feature standardization (shown in equation 1.2) was used.

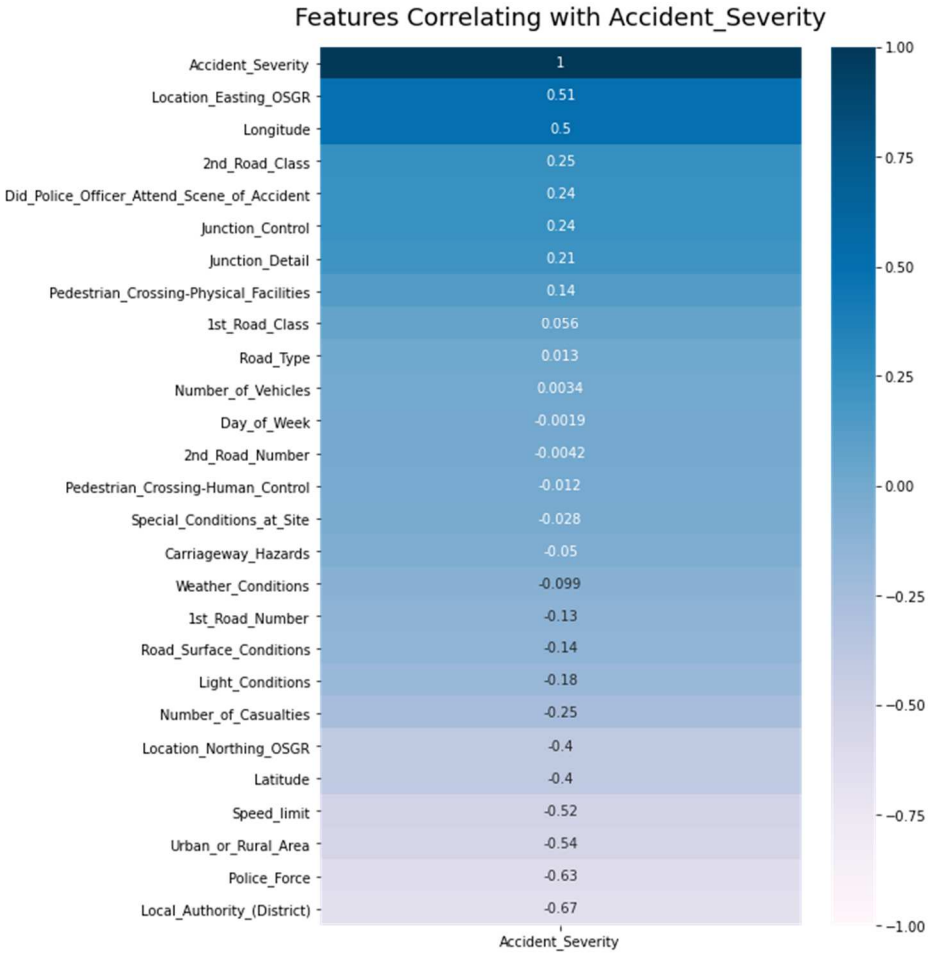

Figure 5: The new correlation indices

Section V discusses the effect of normalization and the standardization of the results.

$$x(i)_{new} = \frac{x(i)_{old} - x_{min}}{x_{max} - x_{min}} \qquad (1.1)$$

$$x(i)_{new} = \frac{x(i)_{old} - \mu}{\sigma} \qquad (1.2)$$

Where μ is the mean of the feature values, and σ is the standard deviation of the feature values.

The final step in preprocessing the data was splitting the test set from the training set. The dataset was split into 34,776 records (80% of the dataset) as a training set and 8,694 records (20% of the dataset) as a test set. The training set will be used to train the models and to validate their scores, this can be done by further splitting the training set into several validation sets. The separation process would result in splitting the training set into 27,820 records (80% of the old training dataset) and 6,956 records (20% of the dataset) as a validation set. Figure 6 shows the steps followed to prepare the data.

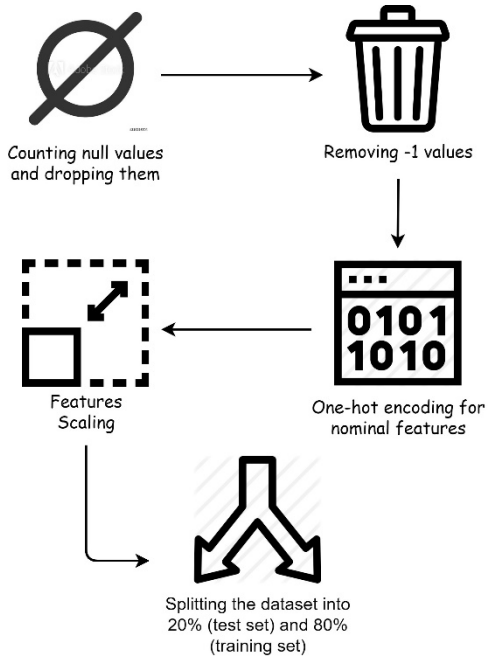

Figure 6: Preprocessing data steps

IV. ALGORITHM

After completing the exploring and preprocessing process, the dataset became ready to be used on classifier models.

The training and validating sets were fit to several classifiers such as SVC, KNN, Naive Bayes, and many other models. The models with the best accuracy scores were short-listed. For increasing accuracy, the hyperparameters of the short-listed models were tuned, using grid search, to get the optimal values for achieving higher accuracy. Then, the tuned models were used in the ensemble learning technique to maximize the accuracy to the most.

The following subsections briefly describe the main principle for the short-listed models, their hyperparameters, and the ensemble learning technique:

A. Artificial Neural Network (ANN)

Artificial Neural Network (ANN) Classifies the model's input instances using neurons. Each neuron determines the class for each instance based on the installed activation function inside that neuron (such as RELU and step function).

A neuron may take a single input or may take several inputs, each input will be multiplied by a weight that will affect the decision of the neuron. The training process includes determining the weights that will be used to multiply the input in the next time step. Having many neurons at the same level forms a layer. Each neural network consists of three layers: The input layer, the output layer, and in between layer named the hidden layer. If the neural network contains more than one hidden layer, then this network is called Deep Neural Network.

ANN has many hyperparameters that can be used to increase the model accuracy such as the number and the size of the hidden layers. Table 2 shows the best hyperparameter values found using grid search.

Table 2: Hyperparameter values for ANN

Model Name	Alpha	Hidden layer sizes	Max Iteration	Solver
ANN	0.05	(10, 37, 10)	1000	SGD

B. Support Vector Classifier (SVC).

Support Vector Machine (also known as large margin classification) is a model that can be used in classification and prediction problems. It is called a Support Vector Classifier (SVC) or Support Vector Regressor (SVR) based on the desired task. SVC works by drawing a dividing line between two different classes of instances. It will be followed by two other sidelines called support vectors. When drawing these three lines to separate the different classes, the shown result resembles a road is formed for the observer. The wider the street margin, the better SVC performance becomes. The two most important parameters of this model are C and Gamma (γ). The first one controls the width of the street while the second controls its curvature level. The best hyperparameter values for this model were found using the grid search. Table 3 shows these values.

Table 3: Hyperparameter values for SVC

Model Name	C	Gamma (γ)
SVC	95	0.01

C. Random Forest Classifier (RFC)

Random Forest is based on a set of Decision Trees. The decision trees use the dataset features to separate classes from one another by checking if the current instance has the current feature (yes) or it does not have it (no). Based on the answers (yes or no), decision trees try to split the answered instances with 'Yes' to be similar grace as much as possible, and that the instances answered with 'No' as similar as possible either. While the difference between each of the two classes is completely different from each other. In Random Forest, more than one decision tree is used to classify instances, and the decision is made based on the highest voting value. The best hyperparameter values found for the RFC model, using grid search, are shown in Table 4.

Table 4: Hyperparameter values for RF

Model Name	Max Features	Minimum Samples Leaf	Minimum Samples Split
RF	12	5	5

D. Ensemble Learning

The ensemble technique combines multiple classifier models to get high accuracy results. Using more than one model for the classification process may produce better results than using each of them separately. Therefore, this technique was used as an attempt to raise the ceiling of accuracy. The used ensemble technique was hard voting. The results section compares the accuracy for the three models and the ensemble learning.

E. Overfitting Analysis.

One of the common overfitting cases in machine learning happens when a model achieves a very high score on the training set but performs weakly on the test set. This happens because the model learns too many details about the training set resulting in saving the outcomes (labels) rather than predicting them. The RFC accuracy score indicated overfitting. The score on the training set hit 97.3% while the score on the validation set was 75% which is less than expected.

This was solved by tuning the hyperparameters of RFC. It increased the validation score, and it achieved a very good accuracy on the test set. Table 5 shows the score of the validation set before and after tuning the hyperparameters. The results of the test set will be discussed in section V.

Table 5: Random Forest Classifier (RFC) Validation Accuracy Scores

	Training Set	Validation Set (Before tuning)	Validation Set (After tuning)
Accuracy Score	97.3%	75%	78%

V. RESULTS

This part shows the achieved results. The used performance metric for all models was accuracy. The accuracy performance metric is the number of the correct predictions the model makes for all categories over the number of the total predictions. This performance metric is more general than the precision and recall and it might be calculated using the confusion matrix. The rows of the confusion matrix represent the actual values, while the columns represent the predicted values that the model made. The main diagonal has an equal number of the i^{th} row and the j^{th} column. So, the higher the main diagonal values, the better accuracy the model achieves. In other words, when the predictions approach the actual values, the model accuracy increases.

The accuracy for ANN was good to some extent. The accuracy of ANN before using the grid search was 74.8% while after using the grid search the accuracy was raised to 75.6%. Figure 7 shows the matrix with the result.

The performance of the SVC model is nearly the same as the previous model, this model got an accuracy rate of 74 % before using the grid search, while it got 75% after using the grid search. Figure 8 shows the confusion matrix for this model.

And finally, the RF model scored the best result. Although the result before the grid search was 74.7 %, the result after using the search improved to become 78.2 %. Figure 9 shows the confusion matrix for this model.

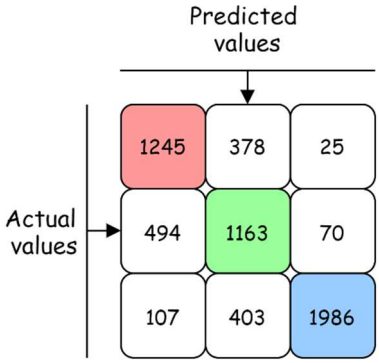

Figure 7:The ANN confusion matrix

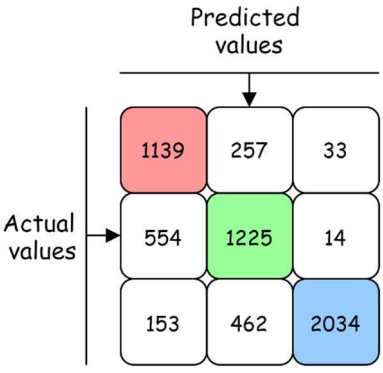

Figure 8:The SVC confusion matrix

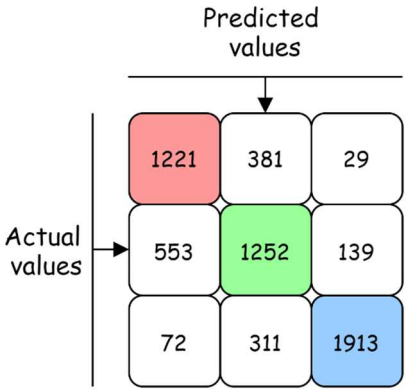

Figure 9: The RFC confusion matrix

The ensemble model got a good accuracy using the three mentioned models. Table 6 shows the accuracy for all models over the test set:

Table 6: Summary of Results

Model Name	Accuracy
Support Vector Classifier	80.4%
Random Forest Classifier	83.9%
Artifactual Neural Network	77%
Hard Voting Ensemble	76.5%

The results also show that the accuracy of standardization values was much better than normalization values. Figure 10 shows the Accuracy difference between standardization and normalization scaling for the ANN, SVC, and RF models.

VI. CONCLUSION

In this work, a machine learning solution we developed to classify accidents severity based on features related to injured people and the accident's environment. Several common machine learning models were developed to achieve this target. The hard-voting ensemble and artificial neural network (ANN) came last with accuracy scores of 76.5% and 77% respectively, the support vector classifier (SVC) came second with the accuracy of 80.4%, and the best-reached accuracy was achieved by random forest classifier (RFC), with the accuracy of 83.9%. The classifier system could be used in combination with other systems to assist relevant authorities to analyze such phenomena and to preserve more lives.

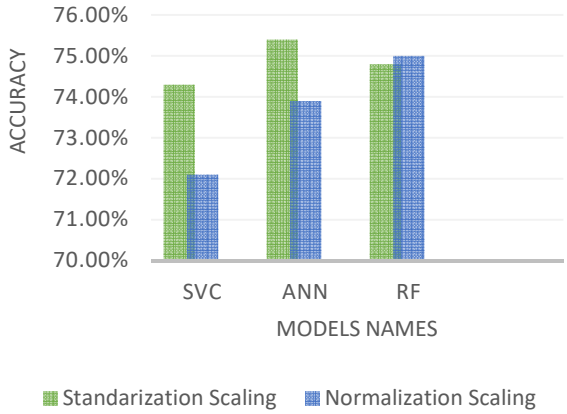

Figure 10: Accuracy difference between standardization and normalization scaling

REFERENCES

[1] "World Health Organization - Road Traffic Injuries," [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries. [Accessed 5 2021].

[2] "Open Government Licence," 5 2021. [Online]. Available: http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/.

[3] Z. Bitar and A. A. Al-Mousa, "Prediction of Graduate Admission using Multiple Supervised Machine Learning Models," in *IEEE SoutheastCon*, Raleigh, 2020.

[4] S. Khalifeh and A. A. Al-Mousa, "A Book Recommender System Using Collaborative Filtering," in *Data'21*, Petra, 2021.

[5] L. Ahmad and A. A. Al-Mousa, "Identification of Donald Trump's Tweets Using Machine Learning," in *Multi-Conference on Systems, Signals & Devices*, Monastir, 2021.

[6] P. J. B. Princess, S. Silas and E. B. Rajsingh, "Machine Learning Approach for Identification of Accident Severity from Accident Images Using Hybrid Features," in *2020 International Conference for Emerging Technology (INCET)*, 2020.

[7] M. Labib, A. Rifat, M. Hossain, A. Das and F.Nawrine, "Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh," in *The 2019 7th International Conference on Smart Computing & Communications (ICSCC)*, 2019.

[8] R. AlMamlook, K. Kwayu, M. Alkasisbeh and A. Frefer, "Comparison of machine learning algorithms forpredicting traffic accident severity," in *2019 IEEE Jordan InternationalJoint Conference on Electrical Engineering and Information Technology (JEEIT)*, Amman, 2019.

[9] H. Xiao, K. Rasul and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarkingmachine learning algorithms," arXiv preprint, arXiv:1708.07747 , 2017.