Journal of Theoretical and Applied Information Technology
30th April 2023. Vol.101. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SEVERITY PREDICTION FOR TRAFFIC ROAD ACCIDENTS

AYOUB ESSWIDI[1*], SOUFIANE ARDCHIR[2], ABDERRAHMANE DAIF[1], MOHAMED AZOUAZI[1]

[1]Laboratory Of Information Technologies And Modeling, Ben M'sick Faculty of Science University Hassan II Casablanca, Morocco

[2]National School of Commerce And Management University Hassan II, Casablanca, Morocco.

E-mail: *esswidi125@gmail.com,

ABSTRACT

Automobile manufacturing, transport and other social and economic fields are impacted by road security. So, the number and the severity of traffic road accidents are problems faced by all these fields to continue their development, this problem can be considered also as an obstacle that slows down the development of countries. Regarding the fast development of data science, those in charge of road management must use this science to minimize the number of accidents or reduce their severity. This study employed exploratory data analysis (EDA) and machine learning (ML) algorithms to predict traffic road accident severity and show the factors that affect them. To this end, we used a dataset proposed by the department for transport in the United Kingdom (UK). This dataset contains information about almost 91200 accidents within more than 70 features in 2020. It was collected by each police force independently, and it is distributed on three CSV files; data about accidents, data about the casualties and data about vehicles crushed. the tasks were, firstly, exploring data and giving insights. Secondly, cleaning and preparing the data for machine learning algorithms, then building and evaluating models using four ML algorithms, and finally, comparing these models to choose the accurate one. As a result, artificial neural networks (ANN) demonstrate their performance. In which, the accuracy reached 0.83 with an average precision of 0.76 and an average f1-score of 0.77 based on the 26 most significant factors.

Keywords: *Traffic Road Accidents, Severity Prediction, Gravity of Accidents, Machine Learning.*

1. INTRODUCTION

Traffic road accidents (TRAs) are one of the important causes of death worldwide. According to the world health organization (WHO), every year almost 1.35 million people die because of road accidents, and between 20 and 50 million people survive and live out the rest of their lives with injuries and disabilities caused by TRA. Moreover, the losses were not limited to this, but several other social and economic domains. That can be considered more destructive than epidemics or other reasons, besides, it can be also considered an obstacle to the development of countries, especially developing ones.

Generally, accidents are classified into three classes: Slight, Serious and Fatal. Dealing with accidents does not mean reducing the number of accidents, but also minimizing their gravity as much as possible. Despite the efforts made by governments and concerned authorities, the number of accidents and the proportion of serious and fatal accidents are still scary numbers. This is due to several reasons such as "road and weather conditions", "the state of the vehicles" and "casualty behaviours and characteristics". Therefore, humanity nowadays needs real solutions. Besides, decisions must be made from a logical analysis of the existing data and scientific studies.

Artificial intelligence (AI) is widely used to provide personalized recommendations to people in several fields. The field of road management could be one of these fields. AI allows concerned authorities to manage roads, based on historical data, and it allows them also to build models for predictions or classifications. Therefore, AI is appreciated for dealing with the raised problem, especially data analysis and machine learning algorithms. Indeed, with data analysis, we provide Charts, graphs, and diagrams, or generally, we give visualization and statistical analysis. In addition, ML helps for building reliable models for TRAs severity prediction.

The remaining portions of the essay are structured as follows: the next section briefly reviews related works on the technics and the approaches used to detect traffic accident severity. Section 3 describes the steps followed from data acquisition to model building step. Section 4 represents the results and evaluation metrics of the experiments. Finally, Conclusions are then drawn in Section 5.

2. RELATED WORK

This work is focused on accident severity prediction, by analyzing historical information about accidents. For this aim, researchers around the world tried different techniques in various domains, one of the most used domains is data science (data analysis and ML model building). They applied statistical analysis and ML algorithms for prediction and classification. Generally, they followed several steps such as data collection and data preparation for ML model building. They finished by comparing the results to provide an accurate model according to the raised problem.

In [1] authors combined "Random Forest and Convolutional Neural Network"; the approach is called RFCNN. For predicting TRA severity, they applied FCNN to a dataset of accident records collected in the USA from 2016 to 2020. Then, they were able to build a model with high accuracy. This model allows for enhancing the process of decision-making and road management.

Reference [2] utilized ANN to forecast the severity of harm in traffic accidents. The dataset used in this work is a set of records about accidents that occurred in Abu Dhabi, with initially 48 attributes including the target variable. This last is a categorical variable of four classes: minor, moderate, severe, and death. After a data pre-processing step, the dataset was reduced to 16 features. Furthermore, ANN classifier achieved an accuracy rate of 0.74.

Researchers in [3] analyzed traffic accidents using ML techniques on a dataset of incidents from Bangladesh. to determine the severity of accidents. Moreover, they compared the results of four ML algorithms in two experiments: the first is on four accident severity classes "Fatal, Grievous, simple Injury, Motor Collision", and the second is on the transformed target variable into (Fatal / Grievous); the Grievous represents the last three classes.

In [4], the authors developed a framework based on six different ML algorithms, which are Naïve Bayes, Logistic Regression, Decision Tree, Random Forest, Bagging, and AdaBoost. Mainly to forecast the severity of crashes if they are fatal, serious, or slights. Hence, the aim behind this implementation is to help enhance safety on roads. Moreover, to keep traffic under control by road authorities.

Sharma B, Katiyar VK, and Kumar K [5] conducted a study to extract the main factors responsible for the majority of accidents. They employed support vector machines (SVM) with Gaussian kernel and Multilayer perceptron (MLP). These approaches were trained and tested on a collected and generated dataset from a questionnaire filled by different kinds of people like drivers, pedestrians, etc. This work showed that a person with high alcohol consumption driving at a high speed has more probability of making an accident.

3. METHODOLOGY FOR TRAFFIC ROAD ACCIDENTS PREDICTION

3.1 dataset

The dataset used in this research was proposed by the department of transport in the UK. The data in the Table 1 were collected by each policeman independently. Every year the government publishes a dataset of three .CSV files distributed as follows.

Table 1: variables of the dataset

.CSV file name	Variable name	Number of possible values (if categorical)
Accidents	Police force	52
	Day of the week	7
	Daytime	Time
	Local authority district	416
	Road type	8
	Speed limit	Non-categorical
	Junction detail	11
	Junction control	8
	Light conditions	6
	Weather conditions	10
	Road surface conditions	9
Casualties	Sex of the casualty	2
	Age band of casualty	12
	Pedestrian movement	11
	Casualty home area type	3
Vehicles	Sex of driver	2
	Age band of driver	12
	Engine capacity	Non-categorical
	Age of vehicle	non-categorical

Journal of Theoretical and Applied Information Technology
30th April 2023. Vol.101. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In this work, we are satisfied only with the data for 2020, it contains pieces of information about almost 91200 accidents, and initially, 70 columns combined in the three files. Figure 1 shows the proposed methodology Flowchart which we have adopted to predict TRAs severity. It shows that the decisions or the recommendations could be provided two times; the first based on EDA, and the second based on a compared and an evaluated ML models.

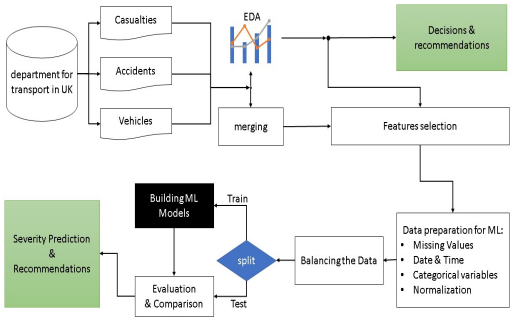

Figure 1: Proposed flowchart to predict TRAs severity and provide recommendations.

3.2 Exploratory Data Analysis (EDA) of Significant Factors of TRAs

This section illustrates graphs to present ideas about the main characteristics of the data. It helped us in the features selection task by visualizing relationships or correlations between features and the target variable. Moreover, it identifies the important variables for classification. In addition to that, it could be considered a statistical analysis task. It allows taking decisions directly by looking at the representations. Furthermore, with EDA road managers were able to provide recommendations behind the decision which is the advantage issue compared with ML classifiers.

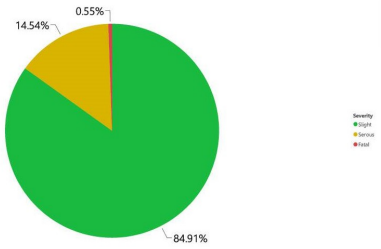

Figure 2: TRAs distribution

The Figure 2 shows that most of the accidents in the dataset are Slights. Otherwise, the percentage of the fatal accident is low which raise a problem known as the unbalanced dataset.

The curves in Figure 3 represents the progress of the number of accidents by the age of drivers. It appears that most of the accidents are made by young people between 20 and 40 years old. However, the age of drivers does not affect the severity of accidents, because of the distribution of the severity on the dataset; the number of accidents considering severity is parallel with their percentages in the dataset.

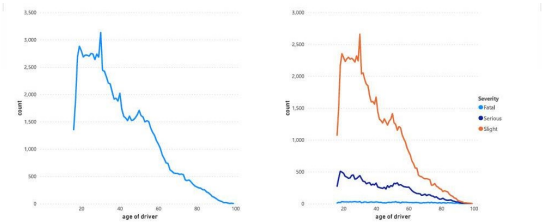

Figure 3: TRAs per age of drivers

Besides, to give a clear insights, several graphs representing the progression of the number of accidents by other variables and regarding severity should be analyzed. These graphs can help for providing recommendations to decrease the number of accidents. For example, when and where we must improve monitoring and raise awareness of the importance of respecting the traffic law. Figure 4Figure 5Figure 6.

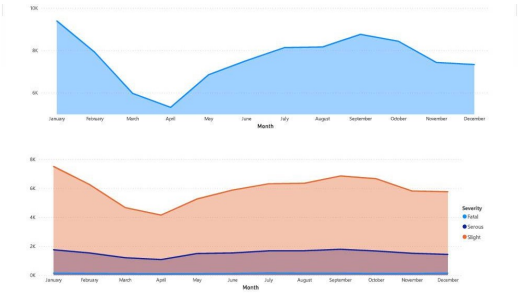

Figure 4: TRAs per months

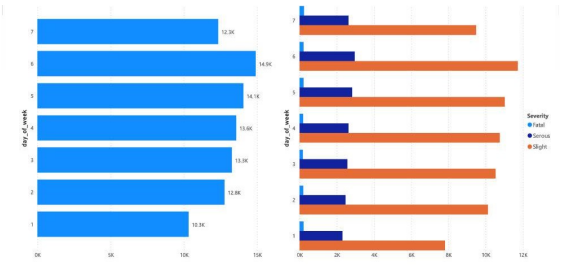

Figure 5: TRAs per days of week

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

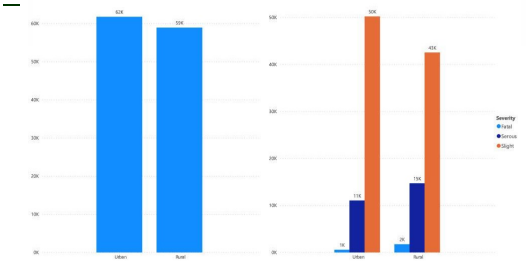

Figure 6: TRAs per regions

By analyzing the graph of the number of accidents in terms of regions Figure 6, it is noticeable that the number of accidents in urban areas is greater than in rural areas. But when we consider the severity of accidents, we noticed that the accidents in rural roads are more severe than in the urban roads, which indicates the importance of the variable for predicting the severity of accidents.

3.3 Features selection for model building

Using as input a dataset with 70 is a huge number, especially since the data has been collected randomly, not for severity problem analysis. To produce a reliable model, a features-selection task is required. To this end, we have adopted two approaches:

Firstly, we manually selected fields; we dropped non-significant columns like indexes, years, and references after merging the three files. These fields seem that they have no impact either on analysis or on building ML models. Along with this, it could negatively impact our models. This approach allowed us to reduce the number of columns from 70 to 57.

Secondly, we adopted the filter method using Pearson's Correlation [6] Equation (1). It calculates the level of linearity between two variables Figure 7. It varies from -1 to 1; where 1 corresponds to Positive linear correlation, 0 to non-linear correlation, and -1 to negative linear correlation Fig. 7, this method allows us to conserve just 28 columns from 57 columns.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (1)$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the **x**-variable
y_i = values of the **y**-variable in a samp
$\bar{y}$ = mean of the values of the **y**-variab

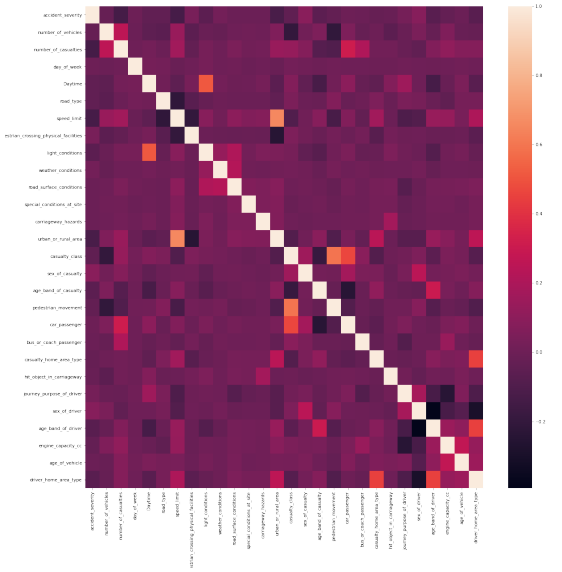

Figure 7: correlation matrix

3.4 Data Preparation

After selecting the subset of relevant variables, it is time to prepare this dataset for ML algorithms. As mentioned before, the data was collected manually, bearing in mind that the target behind this collection was not fixed. Hence, a data cleaning step is demanded. In addition, the application of ML algorithms requires a specific format of inputs to train models.

Firstly, dealing with missing values. From 91200 records, 14 records containing missing values, at least in a single dimension. So, deleting these 14 records won't impact the results of the models. Also, it kept the data intact from any unprofessional tuning.

Secondly, according to EDA, some of the most important variables are the time of accidents and the date. We have transformed the time into a categorical variable with 5 values. Each value represents a part of the day, for example, value 1 represents the morning; the period between 5:00 and 10:00. Additionally, we have done the same for the date; we transformed it into a categorical variable within 7 values, each value represents a day of the week.

To handle categorical variables, by looking at the original data set, we have reached that we have categorical variables with a huge number of values, for example, we have 50 possible values for the Police-force variable, in such a case, we have decided to delete the entire column. On the other hand, some categorical variables were transformed using the one-hot-encoder method. This method is converting each categorical column

into new categorical columns and assigns a binary value of either 1 or 0 to those columns. All the values are zeros, and the index is marked with a 1. Moreover, by comparing speed-limit and engine-capacity values with the other values, we have noticed that the length of the interval for these couple of variables is too long. Therefore, we have reduced this length using normalization methods. We have obtained values between 0 to 9 using the MinMaxScaler transformer Equations (2) and (3). Thus, most of the variables became close in the terms of their intervals of values.

$$X_{scaledValue} = X_{std} * (\max - \min) + min \quad (2)$$

$$X_{std} = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (3)$$

In the end, the data was cleaned. Thus, we are ready to build and train ML models. As it is known, a reliable model must pass through two steps: training and evaluation. Therefore, we have split the dataset into 80% - 20% parts, the larger portion is for training, and the lowest is for the test.

the goal of this study is to predict the severity of accidents, accordingly while looking at the distribution of the data, we have noticed that the observations in the class of Slight accidents are higher than the observations in other classes, especially Fatal accidents Fig. 2. This raises a problem known as unbalanced data. To deal with this problem there are three possibilities:

•Under-sampling: delete or select a subset of samples from the majority class.

◦ Random Under-sampling

◦ Condensed Nearest Neighbour Rule (CNN)

◦ Near Miss Under-sampling.

•Oversampling: duplicate samples in the minority class or synthesize new samples from the samples in the minority class.

◦ Random Oversampling

◦ Synthetic Minority Oversampling Technique (SMOTE)

◦ Adaptive Synthetic Sampling (ADASYN)

•Hybrid: combinations of methods

◦ SMOTE and Random Under-sampling

◦ SMOTE and Tomek Links

◦ SMOTE and Edited Nearest Neighbours Rule.

In this work, we have adopted the SMOTE technique. we have duplicated the labelled observations as Slight and Serious in the training dataset. Thus, the number of observations in each class became the same.

3.5 Machine Learning Classifiers to Predict Severity

To predict severity, we have used ML algorithms. It can be considered as a subset of artificial intelligence. These algorithms make the machine learn from historical data. It trains models to build a reliable classifier. Initially, it is based on mathematic equations. ML algorithms themselves can be classified into traditional algorithms or as simply ML and novel algorithms usually known as deep learning (DL). In this research, we have adopted the four famous ML classifiers.

3.5.2 Decision Tree:

Decision Tree (DT) is a supervised ML algorithm. Both classifications problems and regressions can be solved using DT. Generally, it is based on two measures:

- **Entropy (E)**: to calculate the impurity of the dataset. It allows selecting the best feature to split on and finding the optimal decision tree Eq. (4).

$$E(S) = \sum_{i=1}^{n} -p_i(c) \log_2 p_i(c) \quad (4)$$

$p_i(c)$: the probability of randomly selecting an example in class i

- **Information gain (IG)**: represents the difference in entropy before and after a split on a given attribute Eq. (5).

$$IG = E(Y) - E(Y|X) \quad (5)$$

3.5.3 Random Forest Classifier:

It is a multi-decision tree [7]. These decision trees are selected randomly through an attribute. Then, it combines the results of those decision trees and selects the best one based on votes.

3.5.4 KNN (k-Nearest Neighbor) Classifier:

It makes decision-Based on the k observations near to the entry data point, the proximity from the data point is calculated using one of the famous following distance metrics. See the Table 2.

Table 2: Useful distances to calculate the proximity between two variables.

The name of The distance	The distance $d(x; y) =$	Notes
Minkowski distance	$(\sum_{i=1}^{n} \lvert y_i - x_i \rvert^p)^{\frac{1}{p}}$	The parameter, p, in the formula allows for the creation of other distance metrics.
Euclidean distance (p=2)	$\sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$	Minkowski distance with p=2
Manhattan distance (p=1)	$\sum_{i=1}^{n} \lvert y_i - x_i \rvert$	Minkowski distance with p=1

3.5.5 Artificial Neural Networks (ANN)

It is an efficient computing system inspired by biological neural networks. Or

technically is a perceptron of multi-layers, all inputs pass through a network of nodes connected with each other and organized on layers. Each layer contains several neurons. Generally, ANN architecture is a suit of layers with a specific number of nodes in each layer and is composed of three steps: the first, is forward propagation in which all data vectors travel in only one direction from the first layer to the output. In this step, the algorithm computes a weighted sum and applies an activation function to the result. Then the step terminates with the calculation of the probability that the data input belongs to a given class. After applying the forward propagation on a sample of inputs known as a batch, it is time to compare the predicted classes with the actuals, along with providing a level of truth by calculating the accuracy and loss values. Finally, an optional step called Backward propagation could be adopted, as opposed to the first step, which starts from the loss value to the input, then this last allows editing weights.

4. EXPERIMENTS & RESULTS

This section evaluates the performance of the previous algorithms with different parameters. To obtain the reliable one, for evaluation we used the famous metrics which are accuracy, precision, recall and f1-Score [8]. Equations. (6), (7), (8) and (9).

$$Accuracy = \frac{truePositive+TrueNegative}{Total\ predictions} \quad (6)$$

$$Precision = \frac{True\ Positive}{TruePositive+FaslePositive} \quad (7)$$

$$Recall = \frac{TruePositive}{TruePositive+FalseNegative} \quad (8)$$

$$f1-Score = 2 * \frac{Precision * Recall}{Precision+Recall} \quad (9)$$

TruePositive: the predicted value is positive and its positive.
TrueNegative: the predicted value is negative and its negative.
FalsPositive: the predicted value is positive but it negative.
FalseNegative: the predicted value is negative but its positive.

Besides, for ANN algorithm, we have compared loss and accuracy curves through epochs for validation and train chunks of dataset and, the previous metrics for the test dataset. In addition to collecting and preparing the input dataset, each ML algorithm requires several parameters. Hence, training and validating the ML algorithm means tuning those parameters to obtain the most accurate model [9]. Table 3 represents the appreciative results of the algorithms with the chosen values of the parameters.

Based on the results in Table 3, the accurate model is achieved when using the ANN algorithm. It reached an accuracy of 0.83 with an average precision of 0.76. In addition to that, according to the curves of training and validation accuracy and loss through epochs, these results could be obtained almost after epoch 40, see Figure 8, and Figure 9.

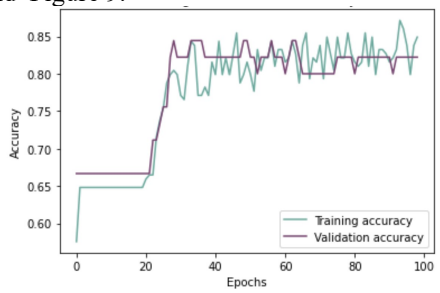

Figure 8: training and validation accuracy

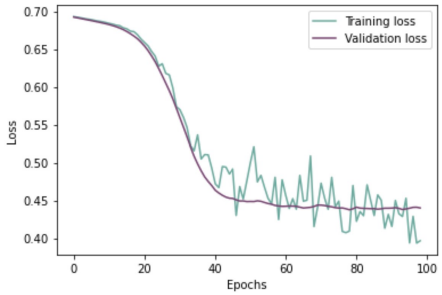

Figure 9: training and validation loss

5. CONCLUSION

Losses caused by traffic road accidents became unbearable. Its cost is unacceptable in

Table 3: Comparison of models regarding the chosen parameters using the average (avg.) of evaluation metrics.

	Chosen parameters	Avg. Precision	Avg. Recall	Avg. F1-score	Accuracy
Decision tTee	max_depth= 9, min_samples_split = 0.5	0.48	0.52	0.39	0.70
Random Forest	max_depth = 16, n_estimators=50	0.58	0.69	0.60	0.80
KNN	Number of neighbors=5, Euclidean distance (p=2)	0.54	0.62	0.57	0.71
ANN	Epochs = 100, optimizer =Adam(learning_rate=0.01), loss = categorical_crossentropy	0.76	0.77	0.77	0.83

several domains, especially in public health and safety. The interests involved in road management in every country are required to find solutions, in order to reduce these losses. Most countries adopt the punishment of offensive drivers. However, regarding the number of accidents, this solution is still insufficient. To tackle this problem there is a need to analyze previous traffic accident data, and to exploit the power of data science. In this paper, with EDA we were able to detect important factors for TRAs and their severities, by the visualization of either the number of accidents per other variables or the number of accidents per other variables considering the severity of the accidents. Furthermore, our goal was to use machine learning for predicting the severity of accidents, which require data preparation tasks starting by reducing the number of features by selecting significant variables and ending by adopting the Synthetic Minority Oversampling Technique (SMOTE) to handle the problem of unbalanced data.

Contrary to the literature, the previous works use only ML to predict the severity of accidents. As known ML allows for building accurate models for prediction. But it is limited to detecting the factors that impact the results. this work combines ML algorithms and EDA, this last allows the identification of factors that causes the accidents and understanding of the data through visualization and analysis.

After that, the Decision tree, random forest, KNN, and ANN algorithms were trained and evaluated. Regarding the results of these algorithms, ANN demonstrates its performance for severity prediction. It has achieved good results, in which with ANN we were able to build a model with an accuracy of more than 0.8. This study is based on data from the UK. Such a study can be applied to other countries. It allows the authorities who work on road safety to take decisions, manage traffic, and detect factors that cause TRAs.

REFERENCES:

[1] M. Manzoor *et al.*, "RFCNN: Traffic Accident Severity Prediction Based on Decision Level Fusion of Machine and Deep Learning Model," *IEEE Access*, vol. 9, pp. 128359–128371, 2021, doi: 10.1109/ACCESS.2021.3112546.

[2] S. Alkheder, M. Taamneh, and S. Taamneh, "Severity Prediction of Traffic Accident Using an Artificial Neural Network: Traffic Accident Severity Prediction Using Artificial Neural Network," *Journal of Forecasting*, vol. 36, Jan. 2016, doi: 10.1002/for.2425.

[3] Md. F. Labib, A. S. Rifat, Md. M. Hossain, A. K. Das, and F. Nawrine, "Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh," in *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, Sarawak, Malaysia, Malaysia, Jun. 2019, pp. 1–5. doi: 10.1109/ICSCC.2019.8843640.

[4] S. Malik, H. El Sayed, M. A. Khan, and M. J. Khan, "Road Accident Severity Prediction — A Comparative Analysis of Machine Learning Algorithms," in *2021 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT)*, Dec. 2021, pp. 69–74. doi: 10.1109/GCAIoT53516.2021.9693055.

[5] Bharti Sharma, V. K. Katiyar, and Kranti Kumar, "Traffic Accident Prediction Model Using Support Vector Machines with Gaussian Kernel," in *Proceedings of Fifth International Conference on Soft Computing for Problem Solving*, Singapore, 2016, pp. 1–10. doi: 10.1007/978-981-10-0451-3_1.

[6] A. Ly, M. Marsman, and E.-J. Wagenmakers, "Analytic posteriors for Pearson's correlation coefficient," *Statistica Neerlandica*, vol. 72, no. 1, pp. 4–13, 2018, doi: 10.1111/stan.12111.

[7] A. Singh, M. N. Halgamuge, and R. Lakshmiganthan, "Impact of Different Data Types on Classifier Performance of Random Forest, Naïve Bayes, and K-Nearest Neighbors Algorithms," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 8, no. 12, Art. no. 12, Jul. 2017, doi: 10.14569/IJACSA.2017.081201.

[8] Ž. Đ. Vujovic, "Classification Model Evaluation Metrics," *IJACSA*, vol. 12, no. 6, 2021, doi: 10.14569/IJACSA.2021.0120670.

[9] K. Thavasimani and N. K. Srinath, "OPTIMAL HYPER-PARAMETER TUNING USING CUSTOM GENETIC ALGORITHM ON DEEP LEARNING TO DETECT TWITTER BOTS," vol. 17, p. 18, 2022.