IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁝⁝ Peer-reviewed / Refereed journal ⁝⁝ Vol. 11, Issue 9, September 2024

DOI: 10.17148/IARJSET.2024.11901

ANALYSIS OF SOKOTO STATE ROAD ACCIDENT AND PREDICTION OF ACCIDENT SEVERITY USING MACHINE LEARNING TECHNIQUE

Muhammad Garba[1], Umar Sharif[2], Mairo Danjumma[3], Sulaiman Umar S.Noma[4], Muhammad Abdurrahman Usman[5], Mustapha Abubakar Giro[6]

Department of Computer Science, Kebbi State University of Science and Technology, Aliero, Nigeria[1,2,3,5,6]

Department of Computer Science, Kebbi State Polytechnic Dakingari. Nigeria[4]

Abstract: The leading cause of death that halts socioeconomic advancement in society is a traffic accident. Nigeria is one of the nations that has experienced a rise in road accidents as a result of a number of contributing factors. In order to predict the severity of road crashes in Sokoto and identify the factors that produce accurate predictions, a comparative analysis will be done using four machine learning techniques, including Decision Tree (DT), Support Vector Machine (SVM), K Nearest Neighbor (KNN), and Nave Bayes (NB). This study will employ data from the Federal Road Safety Corps (FRSCN), Sokoto command. Using the Waikato Environment for Knowledge Analysis (WEKA), the experiment will be carried out. The final result for the experiments shows that Random forest (RF) has the highest accuracy score with 98.11% followed by Support Vector Machine (SVM) with the accuracy score of 94.33%, followed by K Nearest Neighbour (KNN) with the accuracy of 92.45% and the last model with the lowest score is Naïve Bayes with accuracy score of 84.90%.

Keywords: Accident Severity, Machine Learning, Naïve Bayes, Weka, Data mining.

I. INTRODUCTION

Industrial Trust Fund (ITF) (2018) states that the likelihood of a fatal car accident in Nigeria is 47 times higher than it is in Britain. In addition, there are 2.65 crashes for every fatality, which is a high ratio. In contrast, South Africa has one fatality for every 47 crashes, the Czech Republic has one for every 175 crashes, and France has none. In order to classify what combinations of factors can be used to predict whether a crash severity is fatal, serious, or minor, the three main factors of RTC severity—vehicle, environment, and human—need to be thoroughly highlighted (Radzi, Gwari, Mustaffa, & Sallehuddin, 2019). This will make it simple to classify or predict the severity of future road traffic crashes using data mining techniques.

Data mining is a technique for isolating and turning knowledge from a huge dataset by extracting useful information from vast chunks of data. The ultimate goal of this branch of computer science and statistics is to extract information from data and transform it into a form that may be used for other purposes (Leszek, Maciej, & Piotr, 2020). Machine learning algorithms may accurately forecast accident severity by identifying patterns and connections between multiple elements like weather conditions, road characteristics, and driver behaviour by using historical accident data (Abdullahi et al., 2021).

Numerous categories of smart vehicles have emerged as a result of the quick development of new technologies, which has increased the rate of road traffic crashes (RTC) around the world (World Health Organization [WHO], 2018). Road users (vehicles, motorbikes, or pedestrians) collide with one another in a road traffic crash (RTC) when there are technological, human, or environmental shortcomings. According to a World Health Organization (WHO) report, RTC has resulted in up to 50 million injuries and 1.3 million murders worldwide. It is crucial to identify and treat any causes of RTC severity.

Komol, Hasan, Elhenawy, Yasmin, Masoud, and Rakotonirainy (2021) performed a research that focuses on employing machine learning-based classification approaches for modelling injury severity of vulnerable road users; pedestrian, bicyclist, and motorcyclist. The study aims to analyse critical features associated with different (vulnerable road users)

DOI: 10.17148/IARJSET.2024.11901

VRU groups for pedestrian, bicyclist, motorcyclist and all VRU groups together. The supervised machine learning algorithms considered for the empirical analysis includes the K-Nearest Neighbour (KNN), Support Vector Machine (SVM) and Random Forest (RF). The result of the comparative analysis, motorcyclists are found to be more likely exposed to higher crash severity, followed by pedestrians and bicyclists.

A comparative research was conducted between deep learning model and five other data mining technique in order to find which technique can have the most accurate prediction percentage. The techniques used to perfume the comparative research include; logistic regression, XGBoost, K Nearest Neibour, Random forest and Support vector machine. The final findings show that logistic regression algorithms show the best performance among others with an accuracy of 88% in classifying accident severity (ÇELİK & SEVLİ, 2022). Machine learning techniques were used in this study's case study of the Federal Road Safety Sokoto Command in Nigeria's Sokoto State to assess and forecast the severity of accidents. These results highlighted how machine learning has the potential to increase traffic safety.

In conclusion, road accident analysis and prediction of accident severity using machine learning techniques offer promising solutions for enhancing road safety in Nigeria. By leveraging historical data and advanced algorithms, these models can provide valuable insights to inform evidence-based decision-making, reduce accident risks, and improve emergency response strategies. However, continuous research and data collection efforts are necessary to refine and improve the accuracy and reliability of these machine learning models.

II. METHODOLOGY

2.1 Research Design

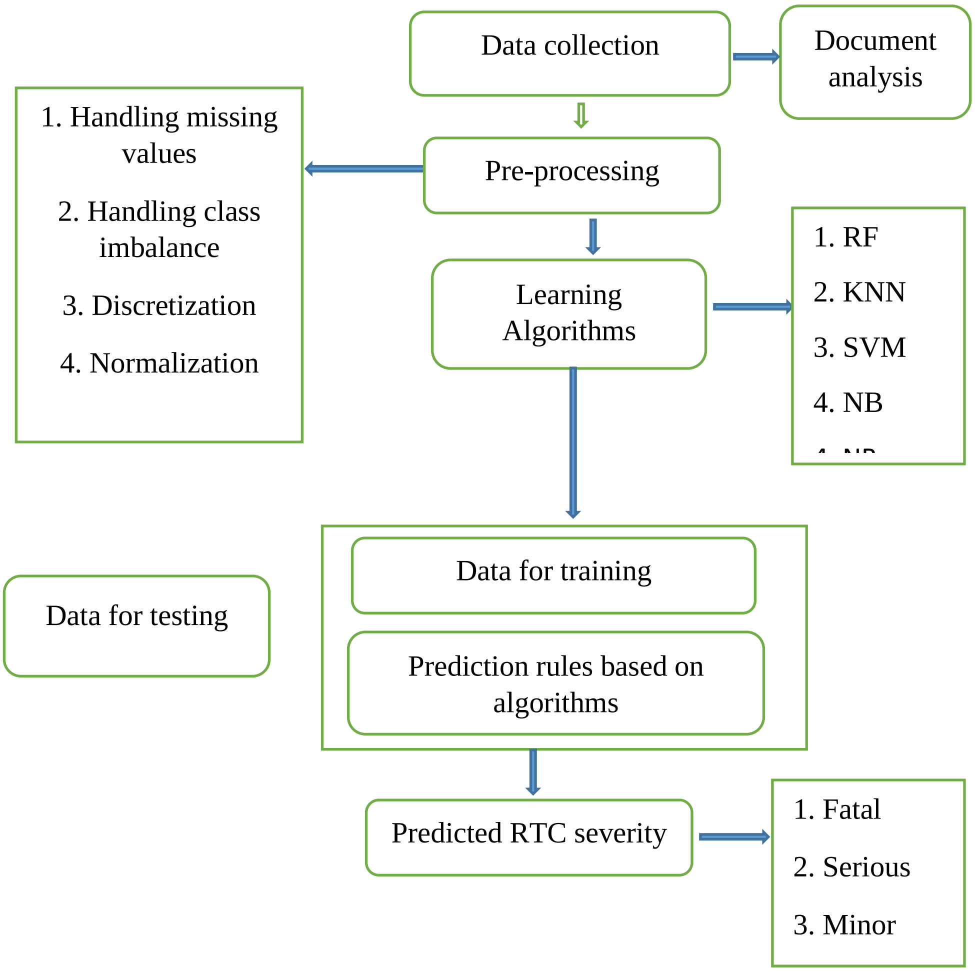

Figure 1: Research Design

2.2 Data Collection

The data for this research is collected from federal road safety corps, Sokoto state command. The data is for the duration of 36 months from January, 2020 to December 2022. The data contain 24 attribute that may lead to accident which are either environmental factors, human factors or mechanical factors. Three targeted classes from RTC (Road traffic crash) were presented which are; fatal, serious and minor cases. Fatal cases from the data is 139, the serious cases are 205 and minor cases have 62 cases. Table 1 describes the attributes (Name, Factor, Description, and Data type), while Table 2 gives the distribution of each of the target classes.

Table 1: RTC Dataset Attribute Description

Attribute Name	Factors	Descriptions	Data Type
Report Time	Human	The time the crash reported	Numeric
Arrival Time	Human	Rescue team arrival time	Numeric
Crash Year	Environmental	Year the crash occurrence	Year type
Crash Time	Environmental	Time the crash happen sharp bend, black spot, etc.	Valid time of crash
Arrival Time	Human	Rescue team arrival time	Numeric
Vehicle Ttype	Vehicle	Type of vehicle involved in the crash (Bus, Lorry, Car etc.)	Categorical
Vehicle Category	Vehicle	Different categories of vehicles involve in the crash (Private, Commercial or Government)	Categorical
Brake Failure	Vehicle	Crash caused by brake failure	Numeric
Mechanical Deficiency	Vehicle	Crash because the vehicle is mechanically deficient	Numeric
Sign Light Violation	Vehicle	Accidents caused by vehicles not having a good working light sign in the vehicle	Numeric
Over Speeding	Human	Accident due to Over speeding	Numeric
Dangerous Over take	Human	Over takes in a corner, sharp bed without seeing his front.	Numeric
Use Of Phone	Human	Accidents caused due to the use of the phone by a driver on the road	Numeric
Sleeping On Steering	Human	Accidents caused due to dangerous overtaking such as overtake in a corner	Numeric
Overloading	Human	The crash occurs as a result of excess overloading of the vehicle with either	Numeric

		passenger or load by the driver.	
Tyre Burst	Vehicle	Accidents resulted as a result of flat tires.	Numeric
Dangerous Driving	Human	Accident due to dangerous driving.	Numeric
Lost Control	Human	Road crash occurred due to loss of control from the driver	Numeric
Over loading overloading Numeric	Human	Accident due to load or passenger	Numeric
Light_conditions	Environmental	Daylight and Darkness	Numeric
Weather_conditions	Environmental	Normal Weather or Raining Weather	Numeric
Type_of_collision	Human	Collision with roadside-parked vehicles, Vehicle with vehicle collision Collision with animals	Numeric
Vehicle_ movement	Human	Going straight, Moving Backward, Moving Backward etc	Numeric

Table 2: RTC severity Distribution

Severity	Number of cases
Fatal	139
Serious	205
Minor	62

2.3 Pre-processing Data

Before applying any machine learning algorithm to a dataset, it is recommended that you carry out data pre-processing. Pre-processing is the process of cleaning the data before further analysis is carried out. It involves several processes that include, missing values handling, normalization, attribute selection or extraction transformation, and/or handling categorical. To handle missing values, we used the mean imputation technique, convert the categorical attribute into numeric using the encoding technique, and solve the problem of class imbalance using the Class-Imbalance technique in W. Attributes selection was also performed to select the most important attribute in RTC severity classification (Hayatu, Mohammed, Baroon, Ali, & Mohammed 2020).

2.3.1 Missing Values Handling

A WEKA filter function called "ReplaceMissingValues" was employed to address the missing values in our dataset. The missing values for nominal attributes were replaced using the mode of the attributes, while the mean value of numerical attributes was used to replace the missing values (Hayatu et al., 2020).

2.3.2 Handling Class Imbalance

The class imbalance is a major cause for concern in a classification or prediction problem. The dataset's unbalanced distribution among the target classes is primarily to blame. When given a dataset with an imbalance in the number of classes, the majority of classification algorithms will prioritize classifying the majority class while disregarding the minority class, which lowers the effectiveness of the classification model (Al-Radaideh & Daoud, 2018). To handle this type of problem, sampling techniques are used which involves the resampling of the original imbalance dataset. This can be achieved in different ways; by oversampling the minority class, by under-sampling the majority class, or by using a hybrid of the two previous methods (Luque, Carrasco, Martin, & Heras, 2019).

In this research study, a WEKA filter function “weka.filters.supervised.instance.Resample” method will be used to oversample the minor cases and under-sample the serious and the fatal cases, and over-sample the Minor cases.

2.3.3 Discretization

Data discretization techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values. This leads to a concise, easy-to-use, knowledge-level representation of mining results. Data discretization can perform before or while doing data mining. Most of the real data set usually contains continuous attributes (Rajalakshmi, Vinodhin, & Bibi, 2016)**.**

2.3.4 Normalization

It’s the process of casting the data to the specific range, like between 0 and 1 or between -1 and +1. Normalization is required when there are big differences in the ranges of different features. This scaling method is useful when the data set does not contain outliers (Jamal et al., 2014).

2.4 CLASSIFICATION ALGORITHMS

Four alternative machine learning methods are applied to the RTC dataset to determine which one performs best in terms of marginal accuracy and recall for classifying RTC severity using pre-processed data. Clustering, association rules, and classification are three categories of machine learning problems. Classification methods assign cases to a predefined target class. Classification in machine learning is divided into two phases: training and testing. Four classification algorithms (DT, KNN, SVM, and NB) will be utilized to identify which algorithm should be used to classify RTC severity in the study area based on accuracy, precision, recall, and F1 Score metrics.

2.4.1 Support Vector Machine (SVM)

Support Vector Machines (SVM) can handle both classification and regression problems. In this method hyperplane needs to be defined which the decision boundary is. When there are a set of objects belonging to different classes then decision plane is needed to separate them. The objects may or may not be linearly separable in which case complex mathematical functions called kernels are needed to separate the objects which are members of different classes (Ray, 2019).

2.4.2 K Nearest Neighbor (KNN)

K Nearest Neighbor (KNN) Algorithm is a classification algorithm it uses a database which is having data points grouped into several classes and the algorithm tries to classify the sample data point given to it as a classification problem. KNN does not assume any underlying data distribution and so it is called non-parametric (Ray, 2019).

2.4.3 Naïve Bayes (NB)

This algorithm is simple and is based on conditional probability, In this approach there is a probability table which is the model and through training data it is updated, The "probability table" is based on its feature values where one needs to look up the class probabilities for predicting a new observation, The basic assumption is of conditional independence and that is why it is called "naive". In real world context the assumption that all input features are independent from one another can hardly hold true (Ray, 2019).

2.4.4 Decision Tree (DT)

Decision Tree is a Supervised Machine Learning approach to solve classification and regression problems by continuously splitting data based on a certain parameter. The decisions are in the leaves and the data is split in the nodes. In Classification Tree the decision variable is categorical (outcome in the form of Yes/No) and in Regression tree the decision variable is continuous (Ray, 2019).

2.5 PERFORMANCE EVALUATION

2.5.1 Performance evaluation metrics

Using a confusion matrix, the performance of the research will be evaluated. A confusion matrix is a visual table that aids in assessing how well a classification algorithm is performing. In a confusion matrix, the expected instances and the actual cases are represented in turn by each row and each column (and vice versa). Ivo and Gunther (2020) define a confusion matrix as a NN matrix in which the goal output is indicated and N helps determine how well a prediction or classification model performs. According to, (Raihan-Al-Masud & Rubaiyat, 2020), each column and each row in the matrix represent the actual target instances and the forecasted cases, respectively. The confusion matrix is shown in from the following table the accuracy precision, and recalls are computed.

Table 3: Performance Evaluation Matrix.

	Actual Label	Predicted Label
	+(1)	-(0)
+(1)	True Positive	False Negative
-(0)	False Positive	True Negative

TP, TN, FP, FN metrics can be described as follows

- True Positive (TP): instances that are positive and classified as positive (Aci & Özden, 2018).
- True Negative (TN): instances that are negative and classified as negative (Aci & Özden, 2018).
- False Positive (FP): instances that are negative but classified as positive (Aci & Özden, 2018).
- False Negative (FN): instances that are positive but classified as negative instances that are negative and classified as negative (Aci & Özden, 2018).

Confusion matrix table is used to calculate different performance metrics as discussed below:

2.5.2 Accuracy (ac): is calculated as the sum of all cases that were correctly categorised or forecasted, divided by the sum of cases that were correctly and wrongly predicted. (Raihan-Al-Masud & Rubaiyat, 2020). This can be expressed arithmetically as in eqn.3.

$$ac = \frac{TP+TN}{TP+TN+FP+FN} \qquad (3)$$

2.5.3 Recall (re): is defined as the number of correctly classified (+) cases divide by the number of (+) cases present in the dataset or the number of correctly classified (-) cases divide by the number of (-) cases present in the dataset (Raihan-Al-Masud & Rubaiyat, 2020). This can be expressed arithmetically as in eqn.4.

$$re = \frac{TP}{TP+FN}, for\ (+)class\ 0r\ \frac{TN}{TN+TP}, for\ (-)class \qquad (4)$$

2.5.4 Precision (pr): is defined as the number of cases the model classified/predicted in the class, and are in the class (Raihan-Al-Masud & Rubaiyat, 2020). This can be expressed arithmetically as in eqn.5.

$$pr = \frac{TP}{TP+FP} \qquad (5)$$

2.5.5 F1 Score ($f1$): The precision and recall are combined to generate the weighted harmonic mean, or f1, which measures overall performance. (Raihan-Al-Masud & Rubaiyat, 2020). This can be expressed arithmetically as in eqn.6.

$$f1 = \frac{2 \times pr \times re}{pr+re} \qquad (6)$$

III. RESULTS AND DISCUSSION

3.1 Pre-Processing Result

The analysis was performed in WEKA using 10-fold cross-validation for splitting the RTC dataset into training and testing to measure the performance of the work.

For the purpose of obtaining good analytic results, the pre-processing results contained the 24 attribute that were either nominal or numerical in the study. Figure 2 displays the attribute visualization that displays the full output of the 24 attributes that were utilized in this research.

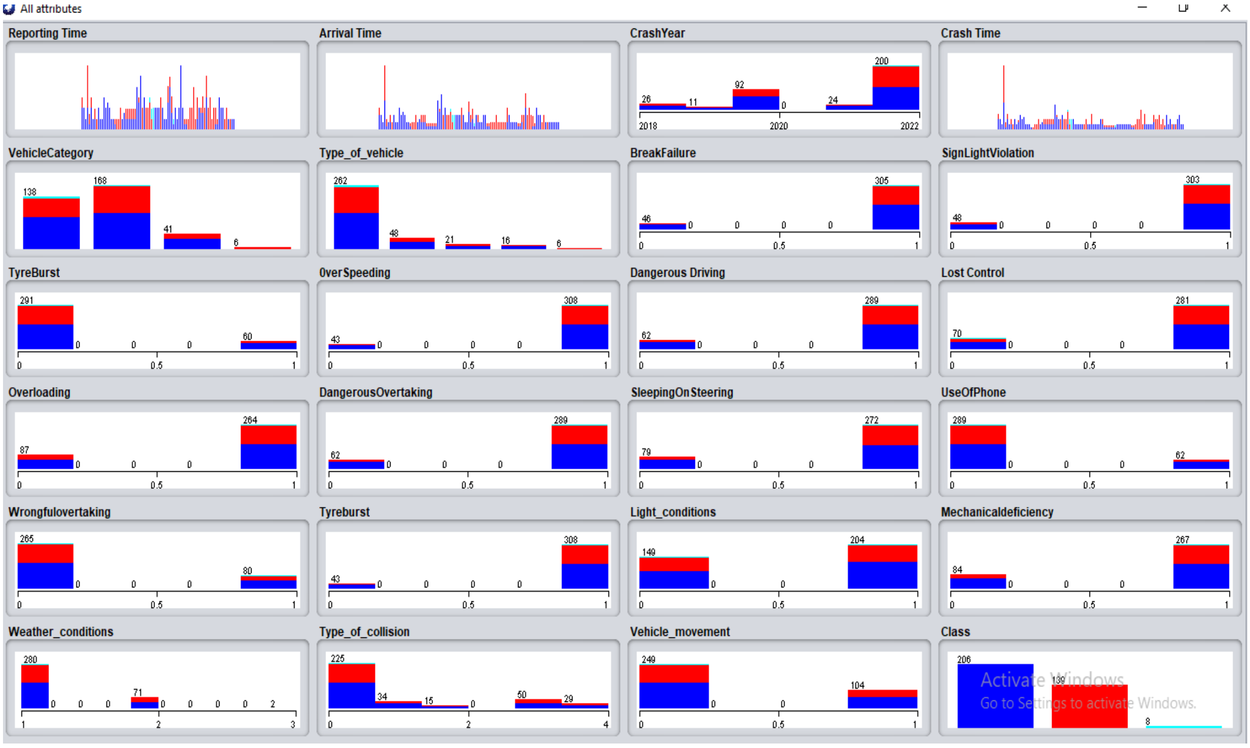

Figure 2: Attributes visualization

3.2 Performance Result For The Classifiers

The experimental results of the data mining algorithms executed on the RTC dataset are shown in this section. The four selected algorithms used in this study are: SVM, K-NN, RF, and NB algorithm.

In the final experimentation of the models which include training and testing the models, Naïve Bayes has the accuracy of 84%, Support Vector Model has the accuracy of 94%, K Nearest Naighbour has the accuracy of 92% and Decision Tree (Radom Forest) has the accuracy of 98%.

The Table 4 summarizes clearly the result of the four selected algorithms that were used to predict RTC severity as fatal, serious or minor using the 24 attributes and the number instances of 354.

Table 4: The summarized result of the four selected algorithms.

S/N	ALGORITHM	CLASS	TP	FP	PRECISION	RECALL	F1-SCORE	RESULT (%)
		Fatal	0.500	0.000	1.000	0.500	0.667	
1	Naïve Bayes	Serious	0.833	0.098	0.714	0.833	0.769	84.9057
		Minor	0.892	0.250	0.892	0.892	0.892	
		Fatal	1.000	0.000	1.000	1.000	1.000	
2	KNN	Serious	0.833	0.049	0.833	0.833	0.833	92.4528
		Minor	0.946	0.125	0.946	0.946	0.946	
		Fatal	0.500	0.000	1.000	0.500	0.667	
3	SVM	Serious	0.917	0.000	1.000	0.917	0.957	94.3396
		Minor	1.000	0.188	0.925	1.000	0.961	
		Fatal	1.000	0.000	1.000	1.000	1.000	
4	RF	Serious	0.917	0.000	1.000	0.917	0.957	98.1132
		Minor	1.000	0.063	0.974	1.000	0.987	

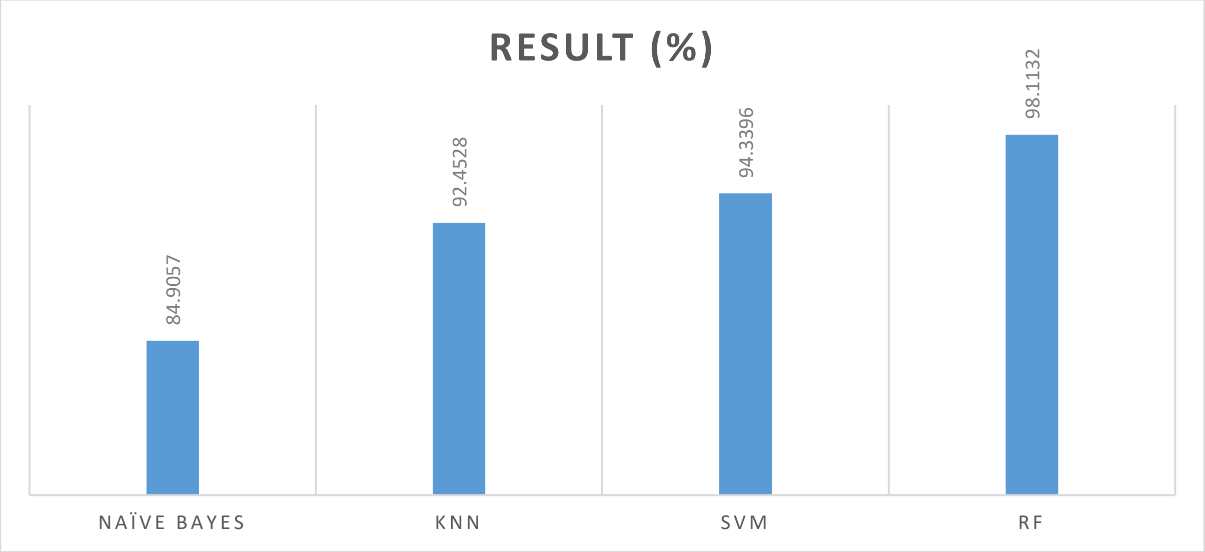

Figure 3: Highest Marginal Accuracy for each of the models.

3.3 Performance Evaluation

This research was conducted based on data mining which focused of predicting RTC specifically on FRCN sokoto state command, which include three main classes of accident which are; fatal, serious and minor. The study was performed in order to explore data mining domain and it aims to achieve an interesting or wonderful result which are related to the RTC classification. The outcome results were in line with the related literature. The research also provide the review of the related literature which were conducted by other researchers. In the final result of the research after experiment on the FRCN sokoto command RTC data was perfumed, the best performing classifier as Fatal, Serious or minor was Decision Tree (Random Forest) with 98.11% of accuracy. The lowest score of accuracy goes to Naïve Bayes with 84.90% of accuracy. The Table 5 shows the performance of other related works compared to this study.

Table 5: Performance Evaluation with Related Studies.

S/N	AUTHOR	CLASSIFIER	DATASET	METHOD	SOFTWARE	RESULT%
1	Hayatu et al.	IG + RF,	Kaduna state FRSCN	Feature Selection	WEKA	97.2
2	Tariq et al.	RF	Yeman Hospital	SMOTE	WEKA	94.8
3	Enviekpaefe and Umar	KNN	Kaduna FRSCN	Feature Selection	WEKA	96.1
4	Rabia et al.	RF	MTA	SMOTE	WEKA	75.5

IV. CONCLUSION AND FUTURE WORK

The WEKA experiment was carried out on the FRCN Sokoto command RTC dataset, which had 24 attributes and 354 instances. Using a variety of machine learning methods, this study divided the RTC dataset into three severity categories: fatal, serious, and minor. At the conclusion of the investigation, the study identifies the most common factors that cause accidents in Sokoto state, which include overspeeding, risky overtaking, and overloading.

There are also certain elements that contribute to the cause of an accident, such as environmental conditions and drug use.

Finally, the study can help law enforcement authorities and the government handle the problem of road accidents in Sokoto State by drafting laws and regulations for motorists.

Feature work; because the study was conducted over a 36-month period using the FRCN Sokoto State command RTC dataset, more data is required for feature work, as well as the use of different FRCN command data from neighboring states such as Kebbi and Zanfara.

DOI: 10.17148/IARJSET.2024.11901

REFERENCES

[1]. Abdullahi U., et al. (2021). Analysis of Road Traffic Accidents in Nigeria Using Machine Learning. *Journal of Traffic and Transportation Engineering, 8(4), 689-702.*

[2]. Aci, Ç, & Özden, C. (2018). International Journal of Intelligent Systems and Applications in Engineering Predicting the Severity of Motor Vehicle Accident Injuries in Adana-Turkey Using Machine Learning Methods and Detailed Meteorological Data. *Original Research Paper International Journal of Intelligent Systems and Applications in Engineering IJISAE*, *6*(1), 72.

[3]. Al-Radaideh, Q. A., & Daoud, E. J. (2018). Data Mining Methods for Traffic Accident Severity Prediction.*International Journal of Neural Networks and Advanced Applications*, 1-12.

[4]. ÇELİK, A., & SEVLİ, O. (2022). Predicting Traffic Accident Severity Using Machine Learning Techniques. *Türk Doğa ve Fen Dergisi*, *11*(3), 79–83. https://doi.org/10.46810/tdfd.1136432

[5]. Hayatu, H. I., Mohammed, A., Barroon Isma'eel, A., Ali, Y. S., & Mohammed, U. S. (2020). Feature Relevance Analysis and Classification of Kaduna State Road Traffic Accident Data using Machine Learning Techniques. *International Journal of Information Processing and Communication (IJIPC*, *10*(1).

[6]. ITF. (2018). Road Safety Report 2018 | NIGERIA. *International Transport Forum.* Retrieved from https://www.itf-oecd.org/sites/default/files/nigeria-road-safety.pdf

[7]. Ivo, D., & Gunther, G. (2020). Indeces for rough set approximation and the application to confussion matrices. *International Journal of Approximate Reasoning*, 155-172.

[8]. Jamal, P., Ali, M., Faraj, R. H., Ali, P. J. M., & Faraj, R. H. (2014). 1-6 Data Normalization and Standardization: A Technical Report. *Machine Learning Technical Reports*, *1*(1), 1–6. https://docs.google.com/document/d/1x0A1nUz1WWtMCZb5oVzF0SVMY7a_58KQulqQVT8LaVA/edit#

[9]. Komol, M. M. R., Hasan, M. M., Elhenawy, M., Yasmin, S., Masoud, M., & Rakotonirainy, A. (2021). Crash severity analysis of vulnerable road users using machine learning. *PLoS ONE*, *16*(8 August). https://doi.org/10.1371/journal.pone.0255828

[10]. Leszek , R., Maciej , J., & Piotr , D. (2020). Stream Data Mining: Algorithms and Their Probabilistic Properties (Studies in Big Data). Springer.

[11]. Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, *91*, 216–231. https://doi.org/10.1016/j.patcog.2019.02.023

[12]. Radzi, N. H., Gwari, I., Mustaffa, N., & Sallehuddin, R. (2019). Support Vector Machine with PrincipleComponent Analysis for Road Traffic Crash Severity Classification. *Joint Conference on Green Engineering Technology & Applied Computing 2019* (pp. 1-5). IOP Publishing. doi:10.1088/1757-899X/551/1/012068.

[13]. Raihan-Al-Masud, M., & Rubaiyat Hossain Mondal, M. (2020). Data-driven diagnosis of spinal abnormalities using feature selection and machine learning algorithms. *PLoS ONE*, *15*(2), 1–21. https://doi.org/10.1371/journal.pone.0228422.

[14]. Rajalakshmi, A., Vinodhini, R., & Bibi, K. F. (2016). Data Discretization Technique Using WEKA Tool. *International Journal of Computer Science ,Enginering and Technology*, *6*(8), 293–298. https://ijcset.net

[15]. Ray, S. (2019). A Quick Review of Machine Learning Algorithms. *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing:Trends, Prespectives and Prospects, COMITCon 2019*, 35–39. https://doi.org/10.1109/COMITCon.2019.8862451

[16]. WHO. (2018). Global status report on road safety 2018. France: *World health organization .Retrievedfromhttps*://www.who.int/violence_injury_prevention/road_safety_status/2018/en/.