IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 7, July 2021

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007243|

2-Layer Voting Classifier Model with Machine Learning Techniques for Crash Severity Prediction

Rajeshri Patil, Hemant Ambulgekar

Department of Computer Science, Shri Guru Gobind Singh Ji Institute of Engineering and Technology, Swami Ramanand Teerth Marthwada University, Nanded, India

Department of Computer Science, Shri Guru Gobind Singh Ji Institute of Engineering and Technology, Swami Ramanand Teerth Marthwada University, Nanded, India

ABSTRACT—Predicting the severity of accident injuries is critical for minimizing the effects of traffic incidents. Accurate severity prediction of road traffic crashes (RTCs) helps to the generation of critical data that can be utilized to implement suitable actions to mitigate the consequences of collisions. Traffic accidents are one of the world's most serious problems, since they result in several deaths, injuries, & fatalities, also economic losses each year. Accurate models for predicting traffic accidents severity are essential for transportation networks. This research endeavors to create models for the selection of important variables and the development of a model for categorizing and forecasting the severity of injuries. Numerous machine learning methods are used to create these models. In this work, we have implemented a 2-layered voting classifier model for this purpose. Supervised machine learning (ML) algorithms, like XGBoost (XGB), Decision tree (DT), Random Forests (RF), and Logistic Regression (LR) are implemented on traffic accident data in the first layer. Then use random forest and select from model to find feature importance and give the input to the second layer. In the 2nd layer, a voting classifier is applied with base machine learning models with Randomized Search CV with cross-validation. The national crash database of road collisions has been used to experiment. The findings of this research propose that the presented model is a viable tool for forecasting the severity of injuries sustained in traffic accidents, with a 96 percent accuracy rate.

KEYWORDS*:* Road Traffic Crashes, Crash Severity, Feature Importance, Randomized Search CV, Voting Classifier.

I. INTRODUCTION

Globally, the number of traffic accidents and their casualties has been increasing in recent years, owing to population growth and increased motorization. The many variables involved in traffic accidents have a significant impact on one another, making it impossible to examine any of the parameters separately when attempting to explain traffic crashes severity [1]. Traffic safety is a worldwide problem that is increasing. It has a devastating impact on both developing and developed nations.

Road Traffic Crashes (RTCs) are a major public health issue on a global scale. According to Peden et al. [2], RTCs are the leading cause of severe injuries & deaths in several nations. As per WHO's worldwide status report on road safety, 1.35 million people die every year in traffic accidents, with traffic accident injuries accounting for the majority of fatalities among young people. Additionally, vehicle accidents are believed to be the seventh leading cause of mortality [3]. Menckel&Andersson [4] determined that understanding the structure of the traffic safety matrix is critical for identifying the main causes of accidents and successful preventive strategies. The severity of RTCs is of great concern to policymakers and safety specialists owing to their enormous economic and social consequences. Precise prediction of traffic accident severity helps to the generation of critical data that may be utilized to implement suitable actions to mitigate the consequences of collisions. Additionally, precise traffic collisions severity prediction enables hospitals to offer medical treatment more promptly in the event of a collision. Numerous models for predicting the severity of RTCs have been suggested based on environmental, accident, vehicle, driver, and roadway factors.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007243|

Road Traffic Accidents (RTAs) are a result of miscoupling of static&dynamic elements (for example, vehicles, people, environment, & roads) [5]. Historical data on RTAs may provide a clear indication of the connection between these variables during the event. Previous research utilized a variety of statistical & machine learning (ML) models to traffic accidents severity prediction, with accuracy as the primary criterion. However, these vehicles' performance was usually poor, particularly in fatal and injury accidents. Additionally, focusing only on forecast accuracy is deceptive.

The ultimate objective of evaluating and researching expert data is to develop the most accurate and complete technique for forecasting the kind and quantity of accidents based on the road's geographical, physical, and human features. Limited accessible accident data, particularly for short-term events or pedestrian collisions, is regarded as primary engineering tests. Instead, the essence of road accident analysis is to deal with a small number of incidents [6]. Because governments use many preventative and corrective procedures, the number of accidents should be kept to a minimum. Thus, accurate and early forecast of traffic accidents severity in work zones is serious for reducing the response time of emergency units (For example, medical assistance), thus improving traffic safety & reducing congestion. Previous research utilized a variety of statistical &ML models to traffic accidents severity prediction, with accuracy as the primary criterion [7]. However, these vehicles' performance was usually poor, particularly in fatal and injury accidents. Additionally, focusing only on forecast accuracy is deceptive. The ultimate objective is to identify &estimate factors that have the greatest influence on accident severity & to provide the best precise prediction model for vehicle &pedestrian accidents separately.

The remainder of the paper is organized in the following manner: Section 2 literature review related to traffic crash severity prediction. Section 3 introduces a study about utilized methodology and problem statement in this paper. Section 4 discusses obtained results of the simulation model. Lastly, Section 5 presents the main conclusions & suggestions are given for future work.

II. LITERATURE REVIEW

Mamlook et al. (2019) The purpose of this study is to assess and compare various methods to modeling accident severity, and impact of risk variables on fatality consequences in traffic crashes, withML-based driving simulations. They created prediction models to determine which variables contribute to traffic collisions and how they might be targeted to minimize accidents. With an accuracy of 82.6 percent, the Random Forest model outperformed the other six methods [8].

Kamble et al. (2019) Importance of vehicle trajectory in road hazard notification, accident avoidance, scheduling & routing, also efficient vehicle movement is addressed with smart vehicle identification. Prediction & grouping of paths isfully reliant on the road's geometry & kinematics. Additionally, the research uses a machine learning model to forecast future vehicle uncertainty, also predicted trajectories are utilized to assess time to collision and collision warning, as well as the quality of service while routing and scheduling vehicle routes[9].

A. Ashtaiwi (2019) The present study proposes IRCA (Intelligent Road Collision Avoidance) system based on ANN (Artificial Neural Network) and DT algorithms. IRCA's prediction model is trained on a large dataset comprised of 1.6 million rows (car accidents) & 23 characteristics (information) covering 14 years of United Kingdom (UK) data gathering. Via prediction accuracy of 72 percent for ANN & 74 percent for TD algorithms, the IRCA system may forecast automobile accident risk levels in 941 districts throughout the United Kingdom. The IRCA system may be used in both human-driven and self-driving cars. Prediction accuracy may be increased further by training on a newly acquired dataset that has fewer missing data and outliers [10].

A. Ji & D. Levinson (2020) The purpose of this research is to leverage the prediction power of crash mechanism-related data via the use of ensemble machine learning models. The findings show that some models may accurately predict severity. For the proposed ensemble combination, a stacking model with the linear blender is recommended. The majority of stacking, bagging, and boosting algorithms perform well, suggesting that ensemble models may outperform individual models [11].

Mamlook et al. (2020) examine the risk variables that contribute to the severity of accident injuries in senior drivers. This is achieved via the development of precise prediction models based on ML. They recommend NB (Naive Bayesian), DT, LR, Light-GBM, and RF models (RF). The findings demonstrate that the Light-GBM model

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007243|

accomplishedthe highest accuracy, at 87 percent, of the five models tested. The 3 most significant factors influencing the severity of injuries, as perthe Light-GBM model, are driver's age, quantity of traffic, & vehicle's age[12].

U. Mansoor et al., (2020) focuses on correctly forecasting the severity of crashes based on easily accessible data from the accident site, such as the kind of intersection control, weather conditions, kind of vehicles involved in the collision, & area's speed limit. The research evaluated the predictive ability of different machine learning models for forecasting the severity of RTCs. 5 base models (DT, KNN, SVM, FNN & AdaBoost) &2-layer ensemble model was constructed using data from the UK's Department of Transport for six years (2011–2016). Additionally, the 2-layer ensemble model beats all other models on the Canadian dataset, according to the findings. The proposed 2-layer ensemble model willhelp forecast accident severity correctly using limited original crash data from the event location. Depends upon this information, trauma centers may prepare for appropriate &rapid medical treatment[13].

S. Elyassami et al. (2020) On the statewide vehicle crashes dataset supplied by Maryland State Police, they compared three machine learning algorithms: DT, RF, and GBT. The gradient boosting model had the greatest prediction accuracy also included the most predictive variables. Results indicate that ignoring traffic signals & stop signs, low visibility, poor road design, & inclement weather are the most significant factors in the predictive RTA model. Utilization of recognized risk factors is critical for developing measures that may mitigate the hazards associated with such variables [14].

Liu et al. (2020) To help intelligent cars reduce the severity of accident injuries in an emergency, it is essential to develop an appropriate crash injury severity model. This article proposes an ensemble model (CSSV-AGX) that dependsupon Classifier-specific Soft Voting that combinesAdaptive Boosting (AdaBoost), Gradient Boosting Decision Tree (GBDT), &eXtreme Gradient Boosting (XGBoost). The experimental outcomeshows that their CSSV-AGX model (50.96 percent accuracy) outperforms traditional MLtechniquesSVM (49.05 percent accuracy), MLP (48.22 percent accuracy), &RF (49.96 percent accuracy) in multi-classification problems. Through sensitivity analysis, they determined that vehicle's weight, speed, & occupants' age, all of which are critical variables for intelligent vehicles, had a significant effect on accident injuries severity. [15].

III. RESEARCH METHODOLOGY

A. Problem Statement

Globally, the number of traffic accidents and their casualties has been increasing in recent years, owing to population growth and increased motorization. The many variables that contribute to traffic crashes have an importanteffect on one another, making it impossible to examine any of the parameters separately when attempting to explain traffic crashes severity. Highway work zones are the most congested and dangerous highway stretches for traffic accidents. Thus, accurate and early forecast of traffic accidents severity in work zones is critical for reducing the response time of emergency units (for example, medical assistance), thus improving traffic safety and reducing congestion. Previous research utilized a variety of statistical &ML models to traffic accidents severity prediction, with accuracy as the primary criterion. However, these vehicles' performance was usually poor, particularly in fatal and injury crashes. Additionally, focusing only on forecast accuracy is deceptive.

B. Proposed Methodology

This work has proposed a novel predictive model to overcome the crash severity of road traffic problems.Firstly, a National Collision Database (NCDB) is used to collect variables (data elements) relating to fatal & injury collisions. It is six years dataset. This dataset has initially contained many entries so need to preprocess this dataset. The data is filtered to include only collisions that occurred at road intersections. Data cleaning entails removing redundant, unnecessary, or missing data fields. After filtering and cleaning the information, they extracted 6000 accident reports for road intersections. The data set included an array of input characteristics of the route, environment, and vehicle. Five types of characteristics were chosen as inputs from the total number of features: roadway features, crash features, vehicle features, environmental features, & area features. These chosen input features are those that may be immediately & readily retrieved from the crash site. Now, they've scaled the data to get correct results from machine learning models. To do this, they utilized the standard scale method to align all variables on the same scale. This scaling results in a transformation of variables that has a mean of zero and a variance of one. Input crash data is then

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007243|

randomly separated into testing& training datasets at 30 percent and 70 percent ratio, respectively. Then, four base models were employed to predict traffic accident severity: XGBoost, Decision tree, Random forest, and logistic regression using RandomizedSearchCV with cross-validation. After emerging base models, the voting classifier model has been applied to integrate four base ML models to improve the performance.

This predictive model consisted of two layers. The four base models – XGBoost, Decision tree, Random forest, and logistic regression with RandomizedSearchCV with cross-validation have been trained & validated in 1st layer. The four fundamental models were chosen for their variety. A voting classifier is utilized in the second layer to perform classification on outputs of 4 basic models from 1st layer. Though by feeding the 2nd layer just outputs from the 1st layer, the model suffered from underfitting. To circumvent this, certain characteristics, in addition to the outputs of layer one, are utilized as additional inputs. This is done to simplify the model by utilizing fewer input features in layer two than in the base models. To reduce the number of features, they utilized RandomForestClassifier and SelectFromModel to determine the significance of each feature.

Together with the output of layer one, these best features provide input for layer 2 of the voting classifier model. A voting classifier is used to train the input features in 2nd layer. It consists of XGBClassifier, Decision Tree, RandomForestClassifier, LogisticRegression with VotingClassifier with RandomizedSearchCV with crossvalidation. The 2-layer voting classifier model was evaluated on the test set once all of these parameters were established.

Numerous methods are used to model a classifier. To predict the severity of RTCs in this research, they used four fundamental machine learning classification models: XGBoost, Decision tree, Random Forest, and Logistic regression classifier. Following that, a two-layer voting classifier model is constructed utilizing these basic models to improve performance. This part provides an overview of various methods conceptually.

1) XGBoost Classifier

XGBoost is an abbreviation for eXtremeGradient Boosting, which was invented by Tianqi Chen [16]. This is a variant of gradient boosting. Because XGBoost is greedy by nature, it takes greedy method. It is very fast & performs well. We reviewed XGBoost and developed a basic baseline XGBoost model that incorporates XGBoost, k-fold CV, and hyperparameter tuning.

XGBoost algorithm now supports a wide variety of hyperparameters.

1) The XGBoost hyperparameters have been classified into four broad groups. They are listed below.
   - Booster parameters
   - General parameters
   - Command-line parameters
   - Learning task parameters
2) Before executing an XGBoost model, we must configure three kinds of parameters: general, booster, and task.
3) Command-line parameters are the fourth kind of parameter. They are only utilized in XGBoost's console version. As a result, we will skip over these parameters and focus only on the first three.

2) Decision Tree (DT)

The most widely used and powerful method in knowledge finding and data mining is decision tree [17] learning. This is used to investigate huge and complicated sets of data to identify valuable patterns. Decision tree learning makes use of DT as a prediction model to connect observations about an object to inferences about its target value. A classification algorithm is used to process a training set that contains a collection of characteristics. The primary goals of decision tree classifiers are to 1) correctly classify as much of the training data as possible; 2) generalize beyond the training data to accurately classify unseen samples; 3) be easy to update as new training data becomes available; and 4) have as simple a structure as possible. Then, the task of developing a decision tree classifier may be divided into the following subtasks:

1) The tree structure is appropriately chosen.
2) The subsets of features to utilize for each internal node.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007243|

3) The decision rule or strategy to be applied at each internal node is selected.

3) Random Forest (RF)

It is a technique for machine learning that may be used for classification, regression, and some other applications. During training, several DTs are produced, &the result is a mean or average prediction for each tree. Tin Kam Ho [18] proposes the algorithm. Random forest is created in the following manner:

1) A sampling of the training dataset is performed using the bagging method. It specifies the number of trees.
   a) Nodes are partitioned according to a set of partitioning criteria.
   b) As a result of the splitting criterion, data is separated into nodes.
   c) Classification occurs at the leaf node level.
2) Once the data has been trained for trees, it is sampled for testing purposes. Each sample is provided to all tress. Classification takes place at the leaf node level.
3) Finally, the test dataset's class is determined using either a majority voting or an average method.

4) Logistic Regression (LR)

The statistical technique of Logistic Regression [19] is used to analyze the dataset and generates a binary result. The dataset may have included one or more autonomous variables. These dichotomous factors influence the outcome. Which implies that there are only two potential outcomes. It is a subcategory of regression that is optimal for predicting binary & categorical output. LR technique is utilized to control the effect of the large number of autonomous variables that are imparted concurrently. Additionally, this approach predicts variables from any of the 2 independent categories. LR employs the greatest likelihood technique to determine the best-fitting function to optimize the probability of categorizing known data into appropriate divisions.

5) Voting Classifier

A voting classifier is a classification technique that makes predictions by combining several classifiers. It is particularly useful in instances when a data scientist or machine learning engineer is unsure which categorization technique to employ. A Voting Classifier is a kind of ML model that is trained using an ensemble of several models and predicts an output (class) based on the class having the highest probability of being chosen as the output. The voting hyperparameter is set to either hard or soft. A voting classifier is applied baseline models with RandomizedSearchCV with k-fold CV on Feature Importance.Cross-validation is a widely used technique due to its simplicity and seeming universality. It is used to assess an estimator's risk or to conduct model selection. Numerous studies have been conducted on the cross-validation methods' model selection performance. This article demonstrates experimentally and theoretically why randomly selected trials are more efficient for optimising hyper-parameters. To begin, do a random search. Consider the fact that attempting every conceivable combination requires a significant amount of brute force computation. We used a more efficient technique: we randomly selected parameters from a range. The aim is that it will cover the set of parameters that is near optimum quicker than grid search. However, this is a naïve approach. It is unaware of and has no recollection of prior runs.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007243|

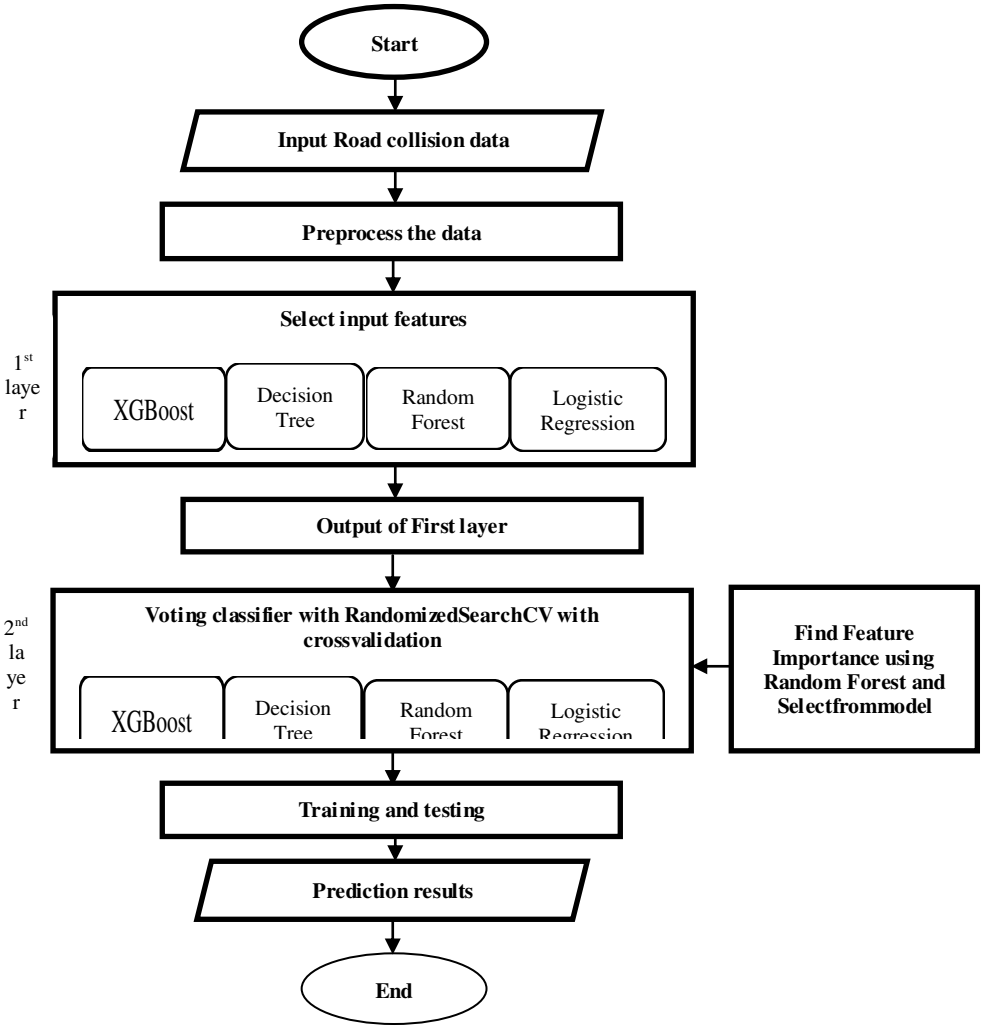


Fig. 1.The proposed 2-layer voting classifier model

IV. RESULTS AND DISCUSSION

The simulation has been done using Jupyter notebook in python programming. The dataset utilized in this study is the National Collision Database (NCDB), which contains information on accident features, road features, vehicle features, area features, & environmental features. Models are trained and evaluated on a dataset comprised of 70% training data and 30% testing data. Each model's performance is assessed using the data produced by the confusion matrix, which includes the outcomes of the model's original and predicted classifications.The efficiency of the proposed classification modelsis simulated with different metrics & test results are represented in subsection.

A. Dataset Description

Models of road traffic crashes are heavily reliant on accident data availability;also, therefore their accuracy is directly proportional to the quality of the available crash dataset. To verify the trustworthiness of accident models, this study uses 6 years' worth (2011 to 2016) of RTC data from the Canadian Department of Transport [20]. National Collision Database (NCDB) Is a database of all motor vehicle accidents on public highways in Canada as reported by police. Selected variables (data components) related to fatal and non-fatal crashes for collisions occurring between 1999 and the most current data available. Open Data will consolidate 59,515 Canada Lands Survey records from Natural Resources Canada's Survey Plan Search Service into one dataset.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007243|

B. Performance Metrics

The following are the most often used performance measures for classification problems:

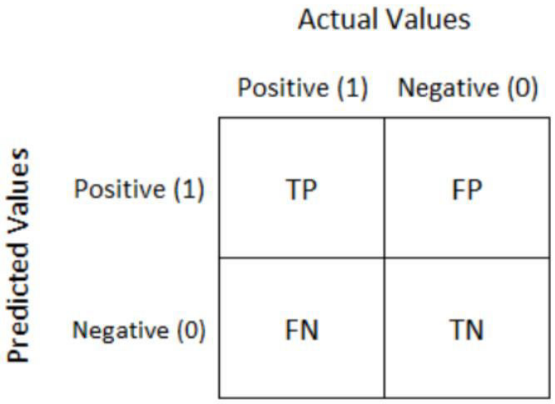
- Accuracy
- Confusion Matrix
- ROC AUC
- Recall, Precision, and F1 score

1) Accuracy

It is a simple ratio calculated by dividing the total no. of points by no. of properly categorized points.

2) Confusion Matrix

It is a summary of projected outcomes in particular table arrangement that enables visualization of machine learning model's performance measure for a binary classification issue with two classes or multi-class classification problem with more than two classes (over 2 classes)



Binary classification's confusion matrix

- TP is an abbreviation for True Positive.It is true if the model predicted a positive class.
- FP is an abbreviation for False Positive.Although this might be taken as a model predicting positive class, it is false.
- FN is an abbreviation for False Negative.It is possible to read this as a model predicting negative class; however, this is False.
- TN is an abbreviation for True Negative.It is true if the model anticipated a negative class.

3) Precision

Precision is defined as the percentage of properly categorized occurrences relative to the total no. of classified instances.

$$Precision = \frac{TP}{TP+FP}$$

4) Recall

A recall is the percentage of properly categorized occurrences in comparison to the total no. of classified instances.

$$Recall = \frac{TP}{TP+FN}$$

5) F1 Score

It is calculated as a harmonic average of recall & precision. It is stated as,

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007243|

$$F1\ score = \frac{2*Precision*Recall}{Precision+Recall}$$

6) ROC AUC

ROCcurve or Receiver Operating Characteristic curveis constructed by graphing TP against FP at different threshold settings. ROC curve is constructed by comparing TP's cumulative distribution function against FP's cumulative distribution function on the y-axis. The area under ROC AUCis a single-valued statistic utilized to assess performance.

C. Screenshots of the Results

This section visualizes the screenshots of all results obtained from the simulation for road traffic collision data. Figure 2shows the collected dataset information with their attribute's values.

```
df.head()
```

	C_YEAR	C_MNTH	C_WDAY	C_HOUR	C_SEV	C_VEHS	C_CONF	C_RCFG	C_WTHR	C_RSUR	C_RALN	C_TRAF	V_ID	V_TYPE	V_YEAR	P_ID	P_SEX	P_AGE	P_PSN	P_ISEV	P_SAI
0	1999	1	1	20	2	02	34	UU	1	5	3	03	01	06	1990	01	M	41	11	1	U
1	1999	1	1	20	2	02	34	UU	1	5	3	03	02	01	1987	01	M	19	11	1	U
2	1999	1	1	20	2	02	34	UU	1	5	3	03	02	01	1987	02	F	20	13	2	U
3	1999	1	1	08	2	01	01	UU	5	3	6	18	01	01	1986	01	M	46	11	1	U
4	1999	1	1	08	2	01	01	UU	5	3	6	18	99	NN	NNNN	01	M	05	99	2	U

Fig.2.Data Information

	C_WDAY	P_SAFE	C_VEHS	C_CONF	C_RCFG	C_WTHR	C_RSUR	C_RALN	C_TRAF	V_TYPE	P_SEX	P_AGE	P_ISEV	P_USER
0	5	2	2	36	3	1	1	1	18	1	1	46	1	1
1	5	2	2	31	1	1	1	4	18	1	0	58	3	2
2	1	2	1	6	1	1	1	1	18	1	0	43	1	1
3	2	2	2	22	1	4	3	1	18	1	1	41	2	1
4	2	2	2	21	3	3	2	1	18	1	1	7	2	2

Fig. 3. Preprocessed data

The preprocessed dataset is shown in Figure 3. Data cleaning was shown by removing any irrelevant, duplicate, or partial data fields. After filtering and cleansing the information, we extracted 6,000 accident reports for road intersections. The data set included input characteristics about the road, the environment, and the vehicle. Fourteen child characteristics were chosen as input from the total; each child feature corresponded to one of 5 parent features, as demonstrated in figure 3: (1)environmental features, (2) collision features, (3)road features, (4) area features, &vehicle features (5). 14 input characteristics chosen are those that may be rapidly & readily retrieved from the accident site. A randomized class balancing method was used to obtain a random sample of crashes comprising 6000 crash points from the filtered dataset. This approach removed any potential of bias toward certain severity levels; the original dataset classified severity levels as mild, severe, and deadly. The severity level indicates the extent of harm suffered by those involved in a collision. Due to the dataset's low fatal accident rate, fatal & serious crashes were combined also classified as severe crashes.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007243|

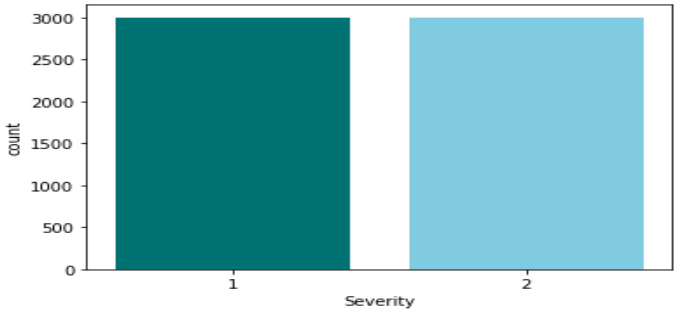
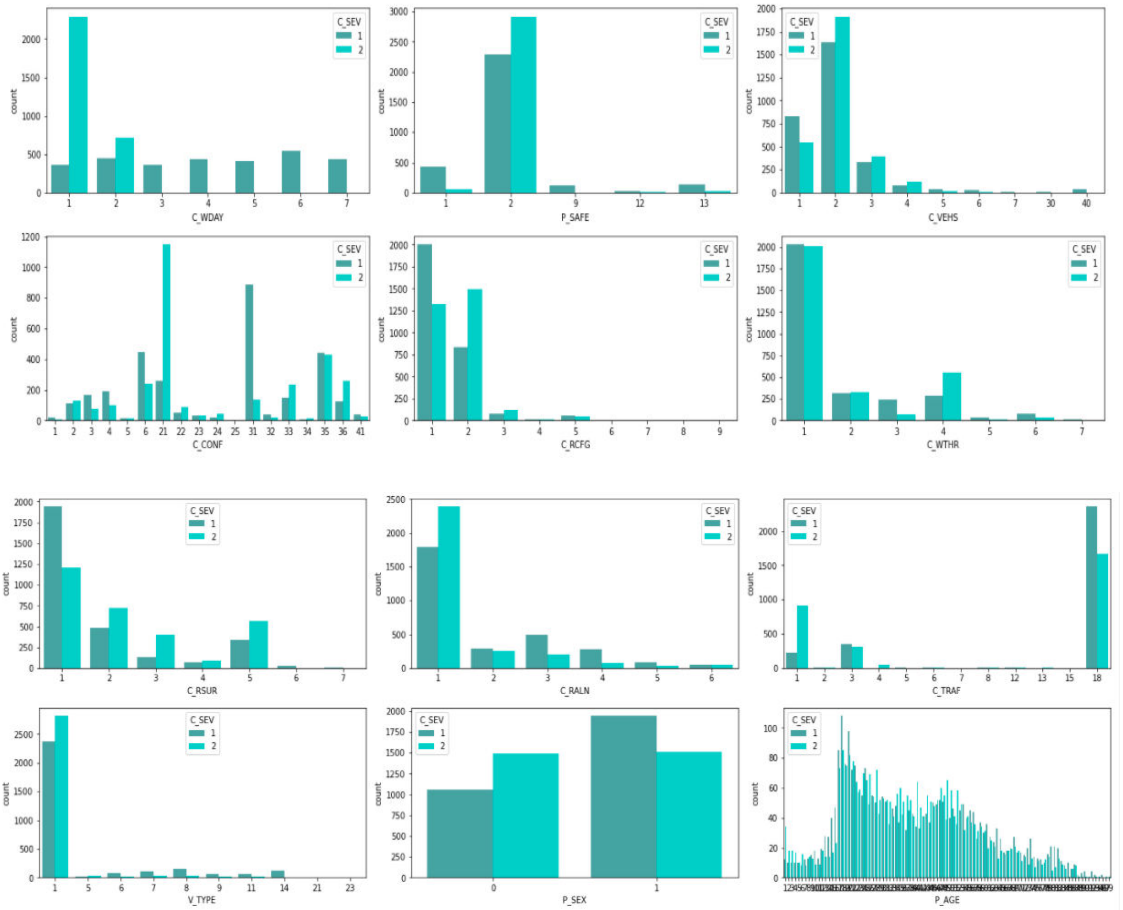


Fig. 4. Severity count distribution

As we have taken equally for both the classes, fatal and non-fatal state, the sample count of both classes isthe same as shown above in figure 4.



| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007243|

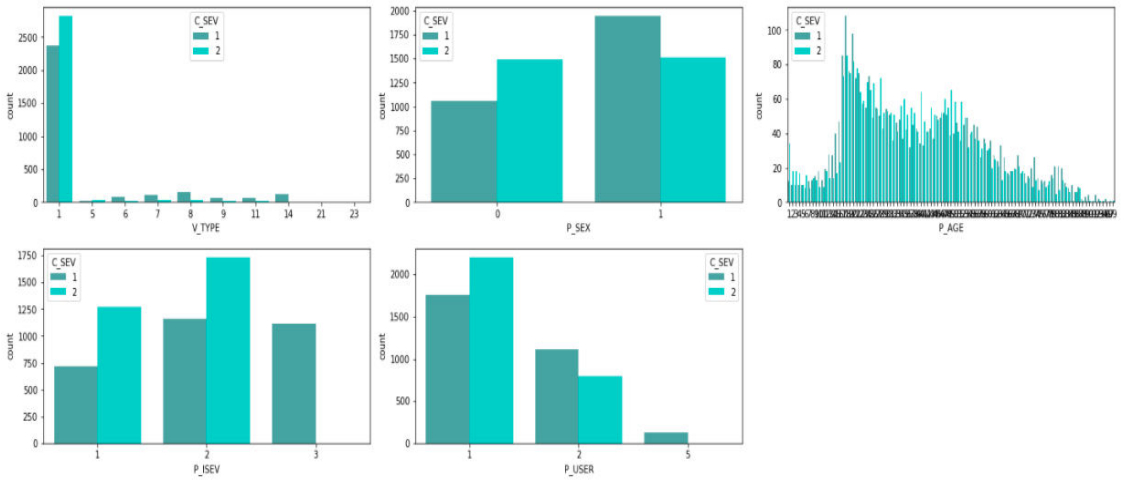


Fig. 5. Histogram visualization for all attributes of road traffic collision data

Figure 5visualize the histogram visualization for all attributes of road traffic collision data. A histogram is a bar graph-like depiction of data in which a range of possible outcomes are bucketed into columns along the x-axis (i.e., attributes). The y-axis displays the count or percentage of occurrences of each column in the data and may be used to visualize data distributions.

	C_WDAY	P_SAFE	C_VEHS	C_CONF	C_RCFG	C_WTHR	C_RSUR	C_RALN	C_TRAF	V_TYPE	P_SEX	P_AGE	P_ISEV	P_USER	C_SEV
0	1.121973	-0.194853	-0.090825	1.115511	1.752009	-0.595432	-0.742139	-0.563178	0.691666	-0.367139	0.858555	0.379477	-1.214759	-0.571228	1
1	1.121973	-0.194853	-0.090825	0.696980	-0.676839	-0.595432	-0.742139	1.996721	0.691666	-0.367139	-1.164748	0.993724	1.624569	0.835738	1
2	-0.829285	-0.194853	-0.394756	-1.395676	-0.676839	-0.595432	-0.742139	-0.563178	0.691666	-0.367139	-1.164748	0.225915	-1.214759	-0.571228	2
3	-0.341470	-0.194853	-0.090825	-0.056376	-0.676839	1.699103	0.614812	-0.563178	0.691666	-0.367139	0.858555	0.123541	0.204905	-0.571228	2
4	-0.341470	-0.194853	-0.090825	-0.140082	1.752009	0.934258	-0.063664	-0.563178	0.691666	-0.367139	0.858555	-1.616827	0.204905	0.835738	1

Fig.6.Dataset after Standard Scaling

Figure 6 shows the dataset after applied Standard Scaling. This data scaling is accomplished by removing the mean value of every variable from the observation's original score also then dividing by the variable's standard deviation.

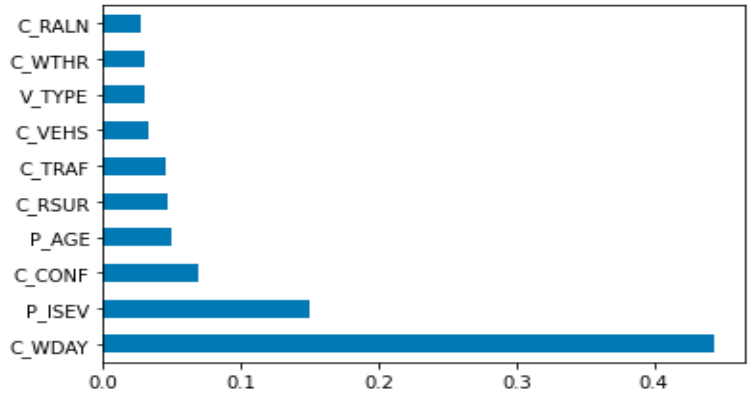


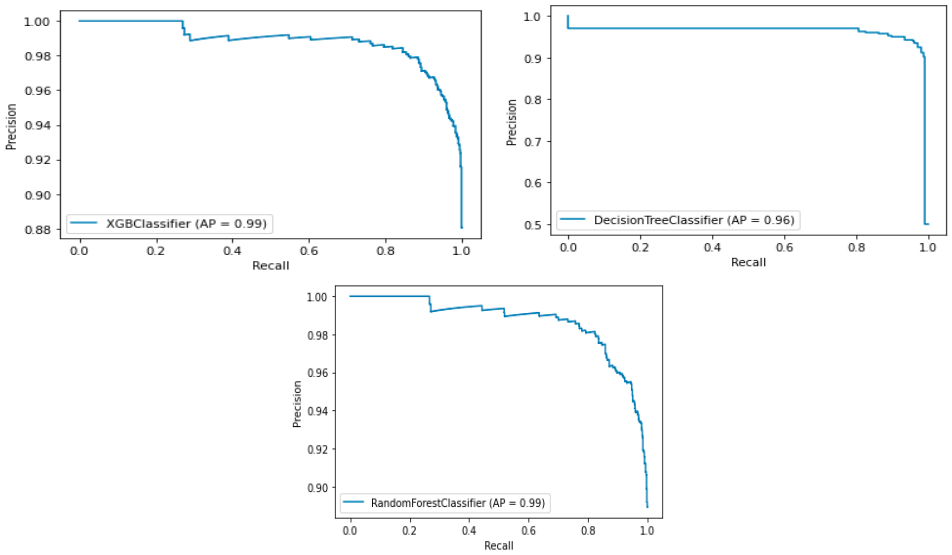
Fig. 7. Bar graph for Feature importance

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

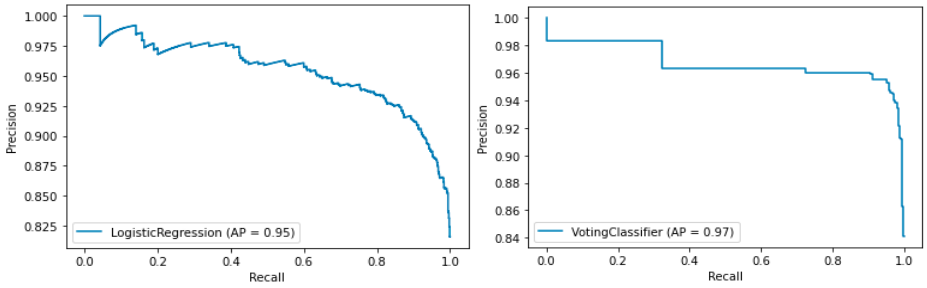
|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007243|

According to the Feature importance rule, only features with select from model were selected. It has provided the ten best features from all classifiers, as shown by a bar graph in Figure 7. These ten features are then input to the voting classifier with Randomized CV search with cross folding in the data.



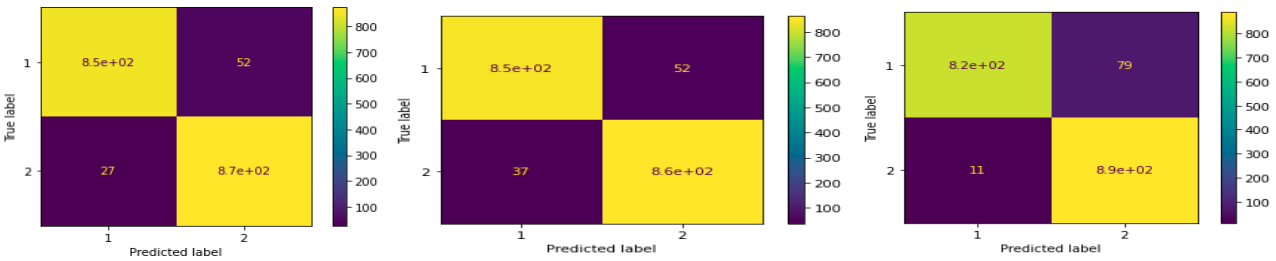
(a) XGB Classifier (b) Decision Tree (c)RandomForest



(d) Logistic Regression (e) Voting Classifier

Fig.8.ROC Area under Curve for all classification algorithms

The ROC curve in Figure 8 is obtained by graphing the cumulative distribution function of Precision against the cumulative distribution function of Recall on the y-axis againstthe cumulative distribution function of Recall on the x-axis.

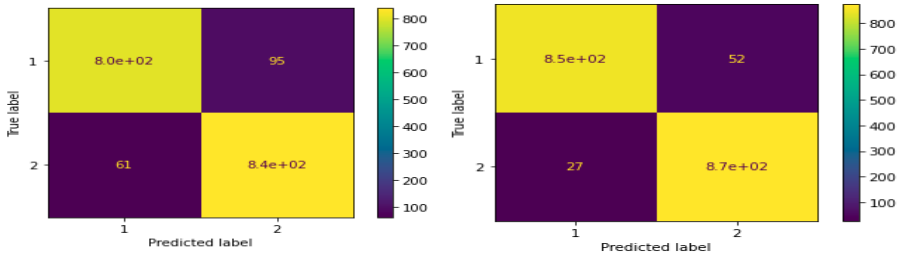


(a) XG Boost Classifier (b) Decision Tree (c) Random Forest

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007243|



(d) Logistic Regression (e) Voting Classifier

Fig. 9. Confusion matrix for all classification algorithms

Figure 9 shows the Confusion matrix that is summarizing the performance of different classification algorithms. On the y-axis, it displayed the true label's cumulative distribution function and, on the x-axis, the predicted label's cumulative distribution function. It displayed a table that is generallyutilized to illustrate the performance of the classification model (or "classifier") onset of test data for which true values are identified.

TABLE I. Performance Measures of Models for Severe Crashes.

Model	Precision	Recall	F1-score
XGBoost Classifier	0.97	0.94	0.96
Decision Tree	0.96	0.94	0.95
Random Forest	0.99	0.91	0.95
Logistic Regression	0.93	0.89	0.91
2-layer Voting Classifier	0.97	0.94	0.96

Tables I and II represent performance measures results for all the modelsfor both severe crashes & non-severe crashes, respectively, in this study. As a consequence of the findings, it is clear that all models performed nearly equally well at both severity levels. For all severity levels, the Logistic regression model has the lowest F1 score of basic models, whereas XGBoost has the highest F1 score. Additionally, DT and Random forest fared equally well at all levels of severity.

TableII. Performance Measures of Models for Non-Severe Crashes.

Model	Precision	Recall	F1-score
Decision Tree	0.94	0.96	0.95
XGBoost Classifier	0.94	0.97	0.96
Logistic Regression	0.98	0.93	0.91
Random Forest	0.92	0.99	0.95
2-layer Voting Classifier	0.94	0.97	0.96

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007243|

TableIII. Accuracy of Models

Baseline Approach					Proposed Approach				
Adaboost	Decision Tree	K-Nearest Neighbors	SVM	2nd Layer (FNN with other classifiers)	XGBoost Classifier	Decision Tree	Random Forest	Logistic Regression	2-layer Voting Classifier
94	95	91	90	51	96	95	95	91	96

Table III is the tabular representation of the Accuracy results obtained by both approaches on the collected National collision database.

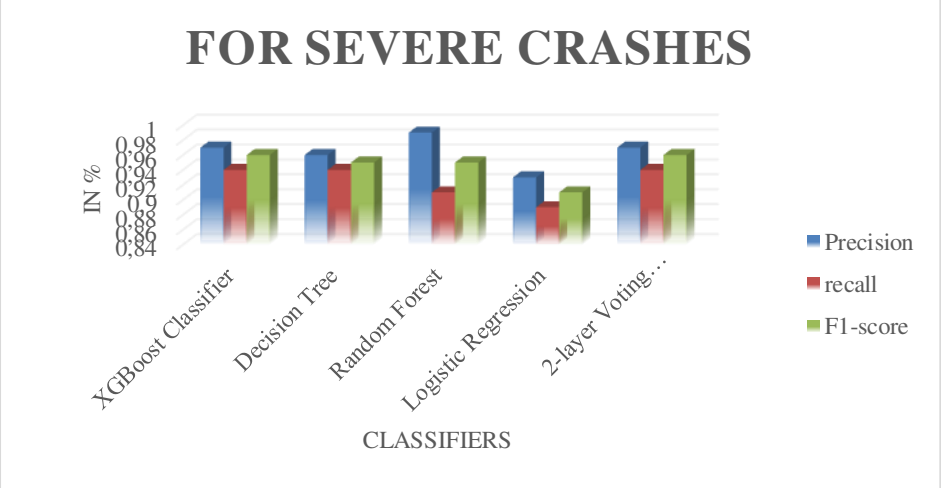


Fig. 10. Bar graph representations ofall metrics for severe crashes

Figure 10 visualize the graphical representation of precision, recall,& f1-score performance metrics for severe crashes for the proposed approach.

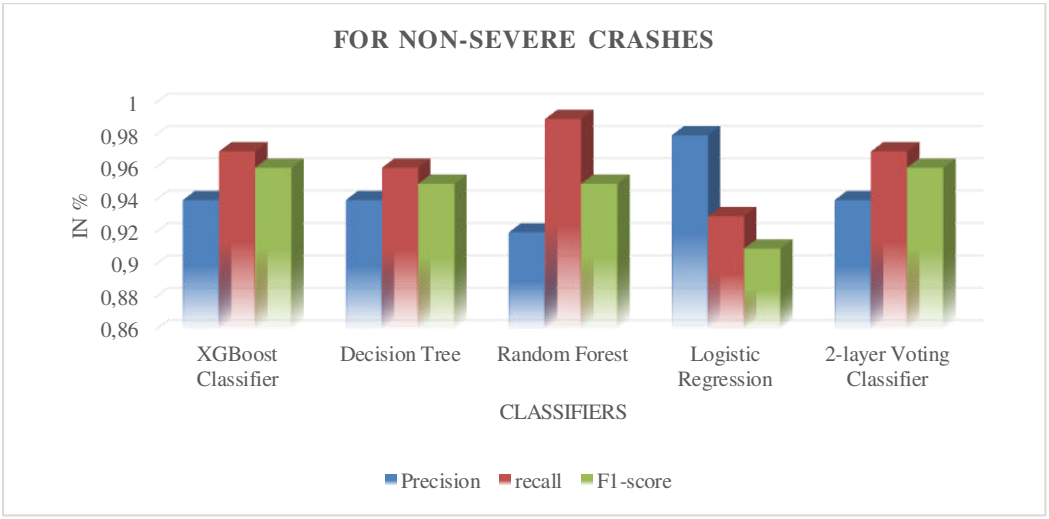


Fig. 11. Bar graph representations of all metrics for non-severe crashes

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007243|

Figure 11illustrated the bar graph representation of precision, recall, and f1-score performance metrics for non-severe crashes for the proposed approach.

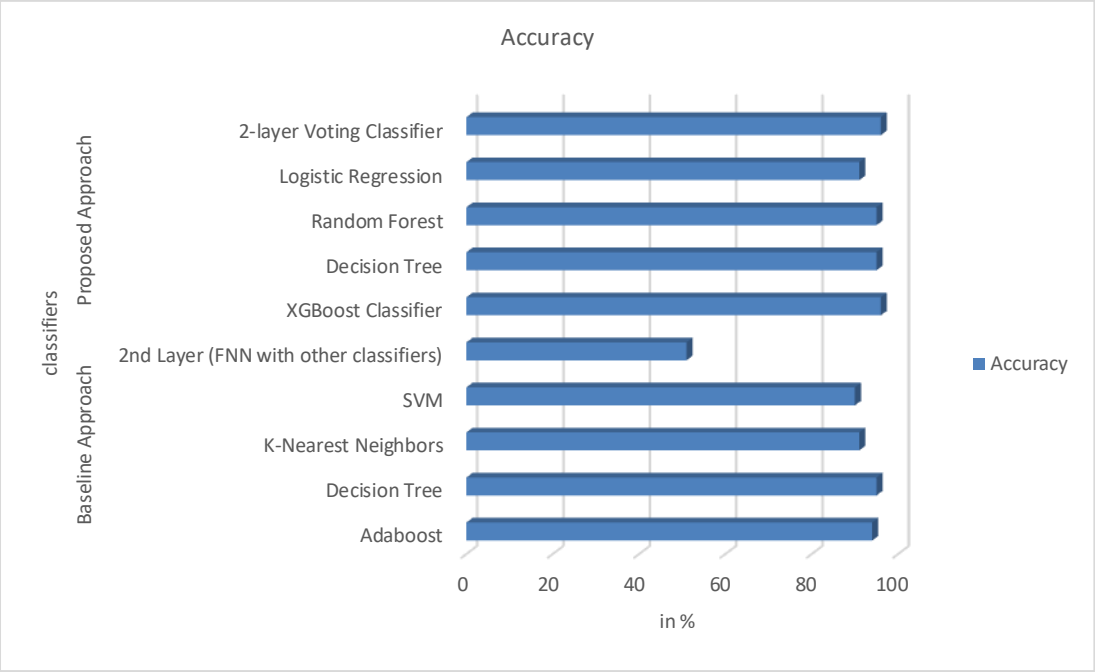


Fig. 12. Comparative graph representations of Accuracy

Figure 12 visualize the comparative graphical representation of accuracy for baseline predictive classifiers models and proposed predictive classifiers models. The proposed approach shows improvement in accuracy in comparison to the baseline approach.

V. CONCLUSION

Because of the enormous economic and social costs associated with traffic accidents, policymakers and safety experts are very interested in their severity. Accurate prediction of traffic accident severity helps to the generation of critical data that may be utilized to implement appropriate measures to mitigate the consequences of collisions. Additionally, precise severity prediction for traffic collisions enables hospitals to offer medical treatment more promptly in the event of a collision.

This article presented a model for a two-layer voting classifier. The research evaluated the predictive abilities of different machine learning models for forecasting the severity of RTCs. 5 base models (AdaBoost, SVM, DT, FNN, & KNN) &2-layer voting classifier model were constructed using data from National Crash Database Online for six years (2011–2016). The two-layer voting classifier model was created by layering together XGBoost, DT, Random forest, Logistic regression, and a voting classifier using RandomizedSearchCV. All machine learning models have their parameters precisely adjusted to produce accurately predicted outcomes. The prediction findings show that the 2-layer voting classifier model beats four basic classification models on 2 performance indicators: testing accuracy and F1 score. Findings indicate that the two-layer voting classifier model outperforms all other models on the Canadian dataset.

As a general conclusion, ML performs better. Furthermore, this work can enhance using models of deep learning to producethe most precise results for the prediction of severity with different road traffic data for better performance.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007243|

REFERENCES

[1] Khaled Assi et al., "Predicting Crash Injury Severity with Machine Learning Algorithm Synergized with Clustering Technique: A Promising Protocol", International Journal of Environmental Research and Public Health (IJERPH), Int. J. Environ. Res. Public Health 2020, 17, 5497; doi:10.3390/ijerph17155497.

[2] Peden M., Scurfield R., Sleet D., Hyder A.A., Mathers C., Jarawan E., Hyder A.A., Mohan D., Jarawan E. World Report on Road Traffic Injury Prevention. World Health Organizatio; Geneva, Switzerland: 2004.

[3] World Health Organization. Global Status Report on Road Safety. World Health Organizatio; Geneva, Switzerland: 2018.

[4] Andersson R., Menckel E. On the prevention of accidents and injuries: A comparative analysis of conceptual frameworks. Accid. Anal. Prev. 1995; 27:757–768. doi: 10.1016/0001-4575(95)00031-3.

[5] Jing Gan, Linheng Li, Dapeng Zhang, Ziwei Yi, Qiaojun Xiang, "An Alternative Method for Traffic Accident Severity Prediction: Using Deep Forests Algorithm", Journal of Advanced Transportation, vol. 2020, Article ID 1257627, 13 pages, 2020. https://doi.org/10.1155/2020/1257627.

[6] Ghasedi, M., Sarfjoo, M. &Bargegol, I. Prediction and Analysis of the Severity and Number of Suburban Accidents Using Logit Model, Factor Analysis and Machine Learning: A case study in a developing country. SN Appl. Sci. 3, 13 (2021). https://doi.org/10.1007/s42452-020-04081-3.

[7] Md AdilurRahim et al., "A deep learning based traffic crash severity prediction framework", Accident Analysis & Prevention, volume 154, May 2021, 106090, https://doi.org/10.1016/j.aap.2021.106090.

[8] R. E. Al Mamlook, A. Ali, R. A. Hasan and H. A. Mohamed Kazim, "Machine Learning to Predict the Freeway Traffic Accidents-Based Driving Simulation," 2019 IEEE National Aerospace and Electronics Conference (NAECON), 2019, pp. 630-634, doi: 10.1109/NAECON46414.2019.9058268.

[9] S. J. Kamble and M. R. Kounte, "On Road Intelligent Vehicle Path Predication and Clustering using Machine Learning Approach," 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2019, pp. 501-505, doi: 10.1109/I-SMAC47947.2019.9032648.

[10] A. Ashtaiwi, "Intelligent Road Crashes Avoidance System," 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2019, pp. 280-286, doi: 10.1109/ICAIIC.2019.8668972.

[11] A. Ji and D. Levinson, "Injury Severity Prediction From Two-Vehicle Crash Mechanisms With Machine Learning and Ensemble Models," in IEEE Open Journal of Intelligent Transportation Systems, vol. 1, pp. 217-226, 2020, doi: 10.1109/OJITS.2020.3033523.

[12] R. E. Al Mamlook, T. Z. Abdulhameed, R. Hasan, H. I. Al-Shaikhli, I. Mohammed and S. Tabatabai, "Utilizing Machine Learning Models to Predict the Car Crash Injury Severity among Elderly Drivers," 2020 IEEE International Conference on Electro Information Technology (EIT), 2020, pp. 105-111, doi: 10.1109/EIT48999.2020.9208259.

[13] U. Mansoor, N. T. Ratrout, S. M. Rahman and K. Assi, "Crash Severity Prediction Using Two-Layer Ensemble Machine Learning Model for Proactive Emergency Management," in IEEE Access, vol. 8, pp. 210750-210762, 2020, doi: 10.1109/ACCESS.2020.3040165.

[14] S. Elyassami, Y. Hamid and T. Habuza, "Road Crashes Analysis and Prediction using Gradient Boosted and Random Forest Trees," 2020 6th IEEE Congress on Information Science and Technology (CiSt), 2020, pp. 520-525, doi: 10.1109/CiSt49399.2021.9357298.

[15] L. Liu, X. Zhang, Y. Liu, W. Zhu and B. Zhao, "An Ensemble of Multiple Boosting Methods Based on Classifier-Specific Soft Voting for Intelligent Vehicle Crash Injury Severity Prediction," 2020 IEEE Sixth International Conference on Big Data Computing Service and Applications (BigDataService), 2020, pp. 17-24, doi: 10.1109/BigDataService49289.2020.00011.

[16] Chen, Tianqi, and Carlos Guestrin. "XGBoost: Reliable largescale tree boosting system." Proceedings of the 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA. 2016.

[17] Shahrukh Teli and PrashastiKanikar, "A Survey on Decision Tree Based Approaches in Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 4, 2015.

[18] Ho, Tin Kam (1995). Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 1416 August 1995. pp. 278282. 8. Leo Breiman, Random Forests, Statis-tics Department, University of California Berkeley, CA 94720, January 2001.

[19] S. Celine, M. Maria Dominic, M. Savitha Devi, "Logistic Regression for Employability Prediction", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-3, January 2020.

[20] https://open.canada.ca/data/en/dataset/1eb9eba7-71d1-4b30-9fb1-30cbdab7e63a

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details