Empirical Analysis of Deep Learning Approach for Driver behavior and Faulty Determination

P. B. S. N. Ariyathilake[1] and R. M. K. T. Rathnayake[2]

[1] Department of Computing & Information Systems, Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka
[2] Department of Physical Science & Technology, Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka
pbsnariyathilake@std.appsc.sab.ac.lk

Abstract. The stir reality of road accidents are deaths and causalities made by road accidents are increasing day by day despite the techniques which are applied with road safety. By causing deaths, economical losses, emotional effects, and social losses road accidents have become 10th most common cause for death. This study focuses on analyzing driver behaviors which cause for occurring road accidents and determining driver faulty of an accident by using machine learning and deep learning techniques. The study has uncovered as the main driver behavior which causes for accidents as the "Speeding" while researcher has proposed Multi-Layer Perceptron approach for determining driver faulty in an accident with 97.93% accuracy. The empirical analysis was conducted to find the best model to identify driver faulty. The findings of the study will be leading to establish road safety strategies to prevent and reduce road accidents.

Keywords: Deep learning, Multi-Layer Perceptron, Empirical

1 Introduction

The bitter reality of the road accident is road accidents are the major cause for the preventable deaths within the society. The tragic instances which cause deaths and the human losses to the society are road accidents. Applying road safety remedies has become a compulsory requirement with increasing tragic accidents. The responsibility not always in the hands of driver when an accident happens. Road conditions, vehicle conditions, weather, and pedestrian behaviors are root cause for preventable road accidents. As per reports of the WHO, 1.3 million people per year died because of the road accidents while disabling 20- 50 million people, becoming 10th most common cause of death in the world[1]. The driver of the vehicle can be considered as the main entity which highly interacts with the road accident. Therefore identifying the major driver behaviors which cause for occurring road accidents is very important when applying remedies to prevent and reduce road accidents. The importance of conducting a study aiming drivers of the road accidents is crucial because in some cases, driver is not main responsible party for happening road accident. Therefore determine driver fault on an accident is the main focus of this study. Currently road safety authorities, like Police departments are using eye witness taxonomy, camera footages and tire marks for determine driver faulty of an accident, that method may has errors and it is difficult to expect right judgments of an accident. Therefore in order to provide a right judgment for driver faulty, study has focused on proposing new machine learning approach which will provide accurate judgments which is a sensitive step in road accident post judgmental activities. In this study Drivers of the accidents are critically analyzed in order to identify their behaviors as well as driver faulty of an accident is analyzed by using empirical analysis techniques.

Due to pattern recognition and data extraction characteristics of the machine learning and deep learning techniques, studies with large data set has successfully applied these techniques. In order to identify the driver behaviors for occurring accidents and to determine driver faulty or not in an accident, this study has used both machine learning and deep learning techniques.

2 Related works

Multiple researches have focused attention on the road accident domain with machine learning and deep learning context.

Xue-Fei Zhang [2] et al. were used various techniques along with WEKA tool to identify major factors related to collisions and severity within those factors in highways. ID3 decision tree algorithm was used for the analysis of accident factors and results were then compared with Weka tool. An analysis was conducted with respect to age, season and gender. The challenge of the study was because of the huge volume of data and the format of the original data is not acceptable to Weka tool.

Gagandeep Kaur et al. [3], were used various tools and methods of data mining in order to focus on parametric analyze of various factors contributing to collisions of road accidents. Correlation analysis and exploratory visualization techniques were used to determine the frequency of accidents and determine road conditions. Regression analysis was done by the R tool.

Sachin Kumar et al. [4] were investigated time series data to find trends in road accidents. Z-score normalization used for data. Then representative time series data were analyzed using least square regression method. future work of the study will focus on developing a novel approach using data mining techniques to analyze the different factors associated

with road accidents in districts where road accidents are happening every time and providing preventive measures to reduce those accidents.

S.Shanthi et al. [5] have used data mining techniques to predict vehicle collision patterns within accident data set. First to classify data various algorithms have used: C4.5, CR-T, CS-MC4, ID3, Decision List, Naïve Bayes, and Rnd Tree. As the relevance algorithms CFS, FCBF, MIFS, MOD Tree, and Feature Ranking were used in this paper. Data were classified mainly in two stages. In the first stage, the classification was conducted using relevance algorithms and in the second stage, the study was performed without applying relevance algorithm. Results were obtained by performing both analyses.

Deep learning techniques have used by Mahler A.I e al [6]for severity prediction within Malaysia. Recurrent Neural Network, Multi-Layer perceptron has used and used deep learning techniques have proven success with high accuracy.

3 Data & Study Design

3.1 Data Set & Study Area Description

The necessary data for the study were retrieved from the Traffic police headquarters Colombo, Sri Lanka. The road accident data were stored as multiple Excel sheets by Unique Accident key. Data records from 2012 to 2016 were obtained for the study. The gathered data set has 207 648 records in that, which has Island wide accident records and each record has more than 60 attributes describing it. Every road of Sri Lanka was covered within this data set. And the study area for the research was selected by analyzing the accident counts of each road. According to the police reports as well as the exploratory analysis conducted, Colombo – Rathnapura – Batticaloa Main road (A004) was selected as the study area for the research. The selected road is 426km long and it covers major cities in Sri Lanka, such as Colombo- Rathnapura- Balangoda- Siyabalanduwa- Kalmunai- Batticaloa. The accident density of the roads was considered when selecting the study area for the research. The selected road for the study is shown in Figure 1.

Fig 23. Study Area of Research

3.2 Study Methodology

The proposed methodology for prediction model is shown in Figure 2. To identify driver behaviors which mainly cause for the road accident is discovered by using a decision tree with a visualization. In order to identify the driver faulty, several algorithms were used. The most accurate model was proposed by comparative analysis. Necessary features for both models were selected by using extra tree classifier which has feature importance features.

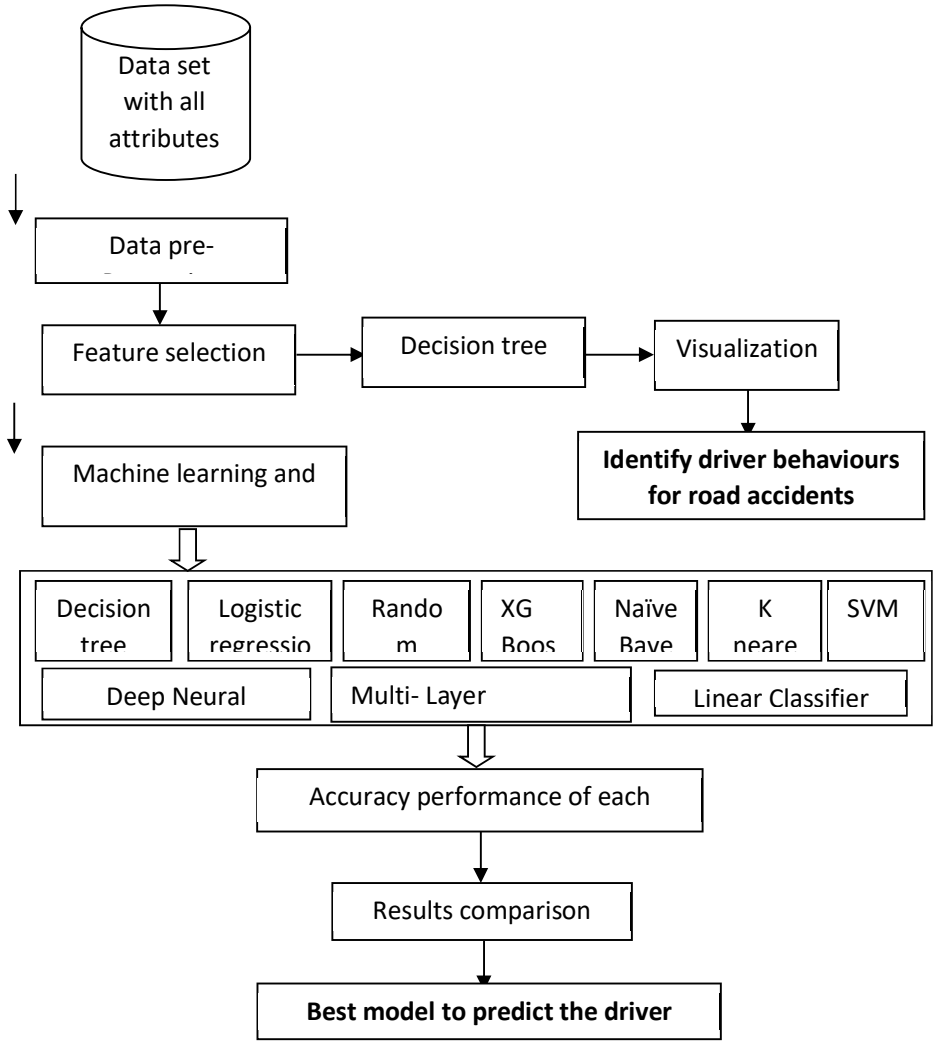

Fig 24. The methodology of the Study

3.3 Data Pre-processing

Data preprocessing is a most valuable, essential and crucial step in every machine learning-based study. The accuracy and performance of the models are highly dependent on the data set machine readability. Therefore missing value removal, data transformation, Outlier handling were conducted with python.

3.4 Feature Selection

Relevant features for the models were selected from the Extra tree classifier in the python which has feature importance characteristics. To calculate Feature importance node impurity and the probability for that link is considered. Node probability will be able to calculate by samples that reach to the node and the total samples. The higher the value gets the importance of the feature increase. The selected features for the driver fault classification is shown in Table 2.

Table 12. Details of the selected attributes

Condition	Feature	Data type
	Highest severity	Categorical
Accident characteristics	Collision type	Categorical
	Pedestrian location	Categorical
	Accident location km post	Categorical
Vehicle characteristics	Age of vehicle	Numeric
	Direction of moving	Categorical
Driver information	Driver gender	Categorical
	Driver age	Numeric
	Validity of license	Categorical
	Number of years since license issue	Numeric
	Human factor	Categorical
	Alcohol test	Categorical

3.5 Decision Tree

The decision tree is a classification or regression model which can be used according to the purpose.[7] In given training vectors,

$x_i \in R^n \quad i = 1, 2\ldots$ (1)

Label vector: $y \in R^l$ (2)

Decision tree partitioned the samples with the same labels together.

Data at m node = Q;

For each time Split: $\Theta = (j, t_m)$ (3)

(which feature j and tm as the threshold)

Partition data in to left and right side, as Qleft (Θ) and Qright (Θ) subsets,

Qleft (Θ) = (x, y)/ xj < = tm (4)

Qright (Θ) = Q/ Qleft (Θ) (5)

3.6 Logistic Regression

In order to identify the occurring of probabilities and to analyze correlation between multiple independent attributes, logistic regression is basically $l = b_0 + \sum_{t=1}^{k} b_i x_i = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$ used by several researchers. Binary and multinomial logistic regression is the basic form of logistic regression. Logit model represents the conditional mean of Y given in x. where t is taken as the likelihood [8].

$$E\left(\frac{Y}{x}\right) = f(l) = \frac{e^l}{1+e^l} \quad (6)$$

$$L = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_{3+\ldots} + b_k x_k \quad (7)$$

3.7 Naïve Bayes

The joint probability of Bayesian network has achieved through chain rule. Naïve Bayes has used Bayesian theory as the classification rule. Probabilities of the variables are considered by Naïve Bayes[9].

$$p(xj|y) = \frac{p(y|xj)\, p(xj)}{p(y)} \quad (8)$$

p(xj|y) = probability of instance y being in class xj
p(y|xj) = probability of generating instance y given class x
p(xj) = probability of occurrence of class xj
p(y) = probability of instance y occurring

3.8 XG Boost

XG boosting is used to give state-of-the-art results on classification problems.For a given data set with n Samples and f features, D = {(xi,yi)} (|D| = n,xi ∈ Rf,yi ∈ R), a tree ensemble model uses K functions to predict the output[10].

$$\hat{y}_i = \phi(X_i) = \sum_{K=1}^{K} f_k(X_i),\ f_k \in \mathcal{F}, \quad (9)$$

3.9 KNN

Nearest neighbor methods are known as prototype models or non-parametric algorithms which learn from memory models. The theory for the KNN model is finding the closest distance to the new sample and values are counted from that point[11].

$$\varphi(x) = \frac{1}{k} \sum_{(x_i,y_i) \in NN(x,\mathcal{L},k)} y_i, \quad (10)$$

3.10 Support Vector Machines (SVM)

Support Vector Machines are maximum-margin linear models which is focused on optimization. SVM can be used with both classification and regression problems which learn from primary optimization[11].SVM classification,

3.11 Deep Neural Networks

$$\min_{\alpha_i} \sum_{i=1}^{1} \alpha_i - \frac{1}{2} \sum_{i=1}^{1} \sum_{j=1}^{1} \alpha_i\, \alpha_j y_i y_j K(X_i, X_j)$$

$$0 \le \alpha_i \le C, for\ all\ i;\ \ \sum_{i=1}^{l} \alpha_i y_i = 0 \quad (11)$$

DNN classifier is mostly used for the binary and the multiple output classification. All the layers are connected to each other by the neurons which are in a network layer. Each neuron sends outputs to its next layer. Each and every neuron are densely connected. Dense can be considered as a fully connected layer, which all the neurons in one layer are connected to the next layer. These layers provide learning features for all features which are learned at the previous phase[12].

3.12 Multi-Layer Perceptron

MLP is the one of neural network type which mostly used by several researchers in order to archive high accuracy with models. MLP consist of interconnected neurons which provide non-linear relationships between input and output features. Weights are the medium which connects nodes inputs of the nodes are modified according to the weights by an activation function. The neural network is described as a graph[13].

G=(V,E) (12)

Function for edges,

w : E – R (13)

Each neuron is modeled as a simple scalar function,

σ: R → R (14)

α – activation function of the neuron

Neuron input is gained by taking a weighted sum of the outputs of all the neurons which are connected to the input. Where weight is w.

X_i – Input signals, V_k – Summing output, W_{ki} – Synaptic weights

$$V_k = \sum_{j=1}^{P} W_{kj} x_j \quad (15)$$

3.13 Linear Classifier

Linear model and logistic function are the basic variations of linear classifier. In the linear model, weights are a computer with a dot product [14].

4 Results & Discussion

The study mainly focuses on identifying driver behaviors which cause occurring road accidents and identifying driver faulty of an accident by a highly accurate model. The results which were gained through the study is discussed separately.

4.1 Decision Tree to Identify Main Driver Behaviors

In order to build the decision tree; nearest lower Km, driver Rider at Fault, driver age and the human pre-crash factor were selected as attributes. The constructed

decision tree is a CART model which can classify the human reason for the accident. DecisionTreeClassiffer in Sklearn library has used to build the decision tree model which can predict the cause of the accident. Train and test data set were split with 0.75:0.25 ratio.

In this decision tree, the analysis was done to identify the factors affect the accident according to their locations. Developed decision tree showed 80.06% accuracy. Which max depth is 7 and the confusion matrix for the model was evaluated. Results of the confusion matrix and its results are shown in Table 1

Table 13. Evaluation Results of the Decision Tree

Classification	Precision	Recall	F1 - score	Support
0- Speeding	0.98	0.96	0.97	358
1-Aggressive/negligent driving	0.00	0.00	0.00	101
2- Error of judgment	0.63	0.97	0.76	232
3- Influenced by drugs or alcohol	0.00	0.00	0.00	6
4- Fatigue/ fall asleep	0.00	0.00	0.00	11
5-Distracted/ inattentiveness	0.00	0.00	0.00	2
6- Poor eyesight	0.00	0.00	0.00	1
7- Sudden illness	0.00	0.00	0.00	1

The decision tree was visualized using GraphViz and Constructed decision tree is shown in Figure 3. It shows that the classification of the driver faulty by dividing and conquering values with entropy, samples, and values. As mentioned below root note contains all data of the data set. It shows that the root contains 2133 samples. Value illustrates the probabilities of each class. In the root node, there are true class samples 1064 and false values in 1069. Entropy means the Impurity of the data, it shows how much classes are mixed up. In the root node it is 1.644(entropy). The first condition shows that from this feature, values are greater than or similar. Satisfied ones will go to the right node and others will go to the left child. It is the first condition which the branches are dividing into nodes (x0 <= 1.5). Green color nodes illustrate the true classes and brown color once are false. When considering the left child of the root node, again it divides in to according to a certain condition(x1<2.5). From that node, 684 goes to the left child and 380 to the right child. Most samples in that node belong to the left class. Leaf nodes at the end do not further divide because no condition to meet at that time. At the end entropy has been 0.061. That means impurity is considerably less and less inconsistency in the data set and all the data has accurately divided without any mix-up. Default visualization of the decision tree gives a visualization of divide and conquer the structure of the data set. In order to conclude a statement from decision tree more specific visualization is needed. Hence the following decision visualization was retrieved from tree viz which is a more collaborative version of a decision tree which can take decisions out of it. Figure 4 shows the specific visualization of the decision tree.

.

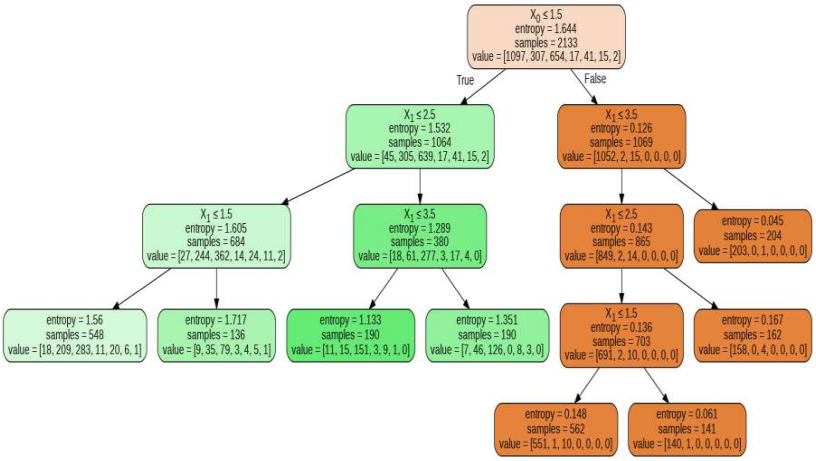

Figure 25. Default Visualization of the Python

At first, the decision tree has split at the point 1.5. It means it has divided whether the driver is faulty or not. In the decision tree if a driver is at fault for the accident it represents by 1. Therefore the left side is the drivers who have a fault on accidents. Then nearest lower km post has used to split the rest samples. After several splitting, decision tree illustrates that the several human factors by pie charts. Most of the accidents were happened because of speeding. It illustrates with the yellow colour. In three locations the main reason for accident was speeding (More than 95%). In addition to that dark green colour which indicates the error of judgment has been a second powerful reason for accidents. As the third reason, decision tree illustrates that Aggressive/negligent driving with light green color. Fatigue/fall asleep is represented by blue color. According to the pie chart for two locations it has been a considerable reason. Influenced by alcohol/drugs and Distracted/inattentiveness (handling radio, mobile) has been a considerable reason for accidents to happen at those particular locations.

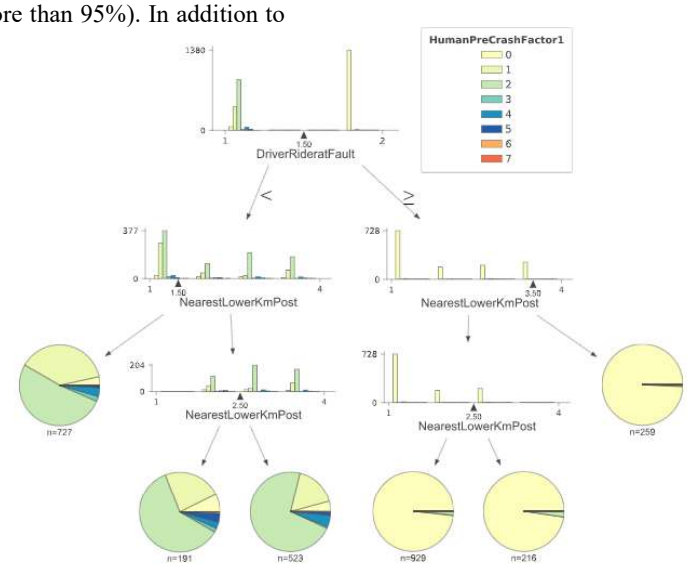

Fig 26. Advanced Visualization of Decision Tree

The major factors which contribute to the road accidents are listed according to vulnerability as follows from the visualized decision tree.

1. Speeding
2. Error of judgement
3. Aggressive/negligent driving
4. Fatigue/ fall sleep
5. Influenced by drugs or alcohol
6. Distracted/inattentiveness

4.2 Empirical Analysis Results for Driver Faulty Identification

In order to determine driver faulty in an accident, an empirical analysis was done. Results for each applied technique were gathered. Results were evaluated with the accuracy and with the confusion matrix values.

The accuracies which were retrieved through the machine learning techniques are shown in table 3 and Figure 5.

Table 14. Accuracy Performance of the Machine Learning Classifiers

Machine learning classifier	Accuracy of the model
Decision tree (CART)	96.15%
Naïve Bayes	93.88%
Logistic Regression	96.47%
Support Vector Machine(SVM)	96.58%
Random forest	96.27%
XG Boost	96.58%
K Nearest Neighbor(KNN)	57.78%
Extra tree classifier	94.19%

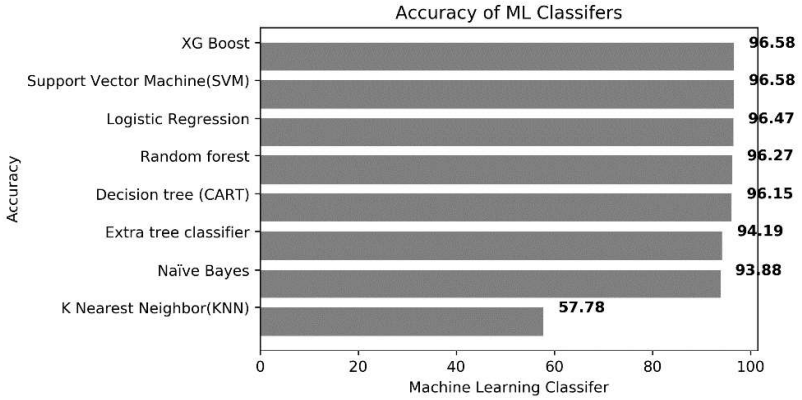

Fig 27. Accuracy of ML Models

The accuracies which were achieved through deep learning techniques are shown in Figure 6.

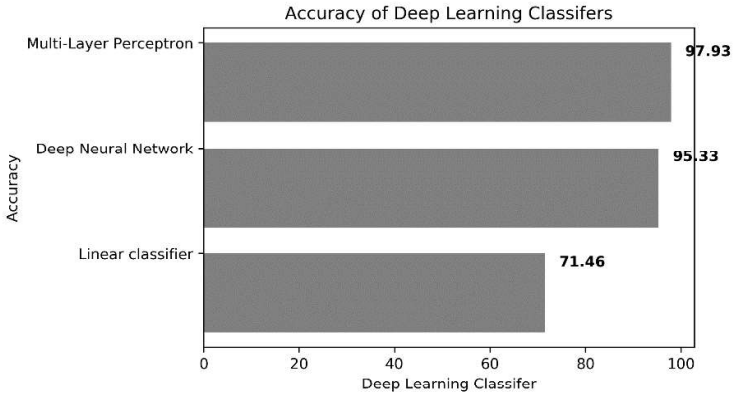

Fig 28. Accuracy of Deep Learning Classifiers

The confusion matrix results which received for the all machine learning classifiers are shown in Table 3.

Table 15. Confusion Matrix Results of the Classifiers

	Precision	Recall	F1 score	Support
Fault	0.95	0.95	0.95	513
Not fault	0.94	0.94	0.94	451

According to the accuracies of all the models which are displayed in figure 7, Multi-layer perceptron model was selected as the best model for the driver faulty determination.

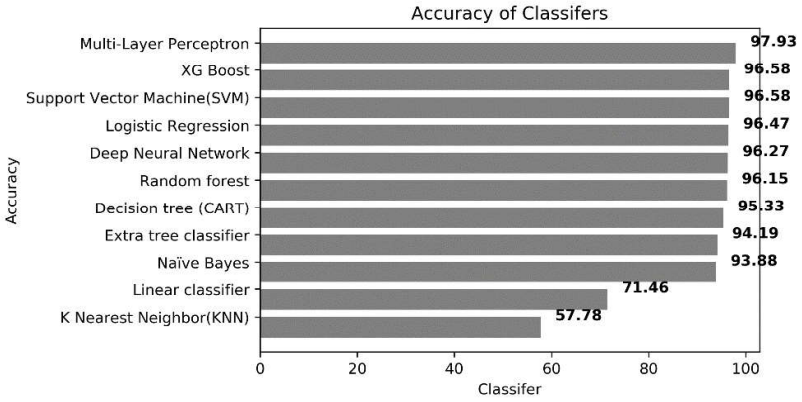

Fig 29. Accuracy Comparison of All Classifiers

4.3 Multi-Layer perceptron model approach Results (Proposed model)

The developed Multi-Layer perceptron model comprises of two hidden layers and an output layer with drop out function. The main inputs for the model is a set of thirteen features and the output is the driver fault or not. The input dimension to the first hidden layer is equal to the thirteen. Outputs of the first hidden layer are results from the activation function, RELu(Rectified Linear Unit). RELu is applied to the weighted sum of both input and the output of the layers. The second hidden layer has used TANh as the activation function. Fully connected layers were trained in order to classify the two classes in the output layer. Two dropout functions were used to each hidden layer to maximize accuracy. To reduce the mean of the error, softmax_cross_entropy_with_logits loss function has used with RMSPropOptimizer.

Multilayer feed-forward neural network used with k fold validation was trained with 80% of training data. Data set was divided into three different data sets. Training, testing and validating. The model was trained with different epochs (iterations). Starting from 10 epochs to the 100 epochs multilayer feed-forward network was trained. The testing accuracy, validation accuracy, and loss which received at each epoch are shown in figure 8 and Figure 9. In general model accuracy has been increased with each iteration and at certain point accuracy has been constant even though the number of iterations was increased.

At 50 epoch, accuracy has been increased to a certain limit with a minimum loss. Then the accuracy and the loss has come to a consistent value. At the 50 points, the model achieved 0.9793 accuracy for the validation while the loss was 0.1240. Testing accuracy for the 50 epoch was 0.9638 and the AUC value was 0.9702.

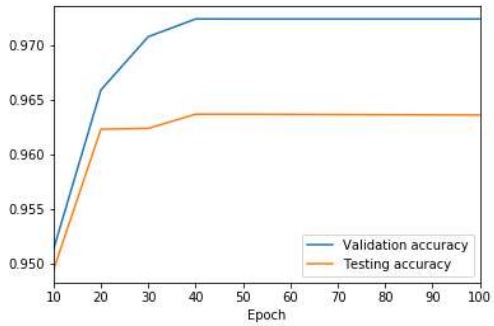

Fig 30. Validation and Testing Accuracy for Different Epochs

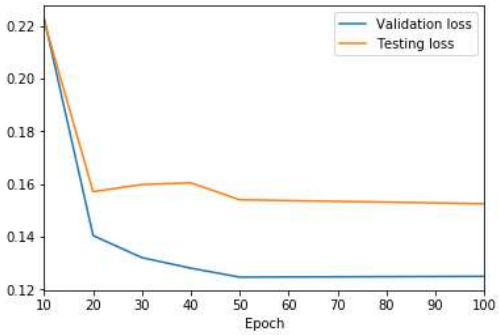

Fig 31. Loss Distribution for Different Epochs

In order to find the best optimization function for the model, several optimization functions were used with the model to determine the accuracies. At their RMSProp was selected as the best optimization function out of Adam, Adagard and SDG. And for the learning rate 0.001 was selected by performing a sensitivity analysis. Therefore at the 50 epoch, model was selected as its best performing level at a minimum epochs. The complete results of the 50 epoch are shown in Table 4.

Table 16. Results of the Epoch 50

Epochs 50	Training 0.80 /Testing 0.20			Training 0.70 /Testing 0.30		
	Accuracy	Loss	AUC	Accuracy	Loss	AUC
Fold0	0.9659643	0.1489272	0.973129767	0.94169814	0.15775863	0.956702968
Fold1	0.97408846	0.1303627	0.972129775	0.95415094	0.14760377	0.955641344
Fold2	0.97088865	0.1426373	0.984127913	0.95798744	0.15046333	0.963506179
Fold3	0.96115726	0.1763699	0.961118890	0.9403774	0.16626956	0.954337063
Fold4	0.97938237	0.1240679	0.976619114	0.9502771	0.13649173	0.959617267
Validation	**0.97938237**	**0.1240679**	**0.976619114**	0.9502771	0.13649173	0.959617233
Testing	0.96385316	0.1540825	0.970220142	0.94852112	0.12695076	0.964283101

5 Conclusions

Road accident domain can be considered as an area which needs high attention in reducing road accidents. The driver can be considered most responsible party which directly engage with accidents. Therefore in order to identify human factors which cause for road accidents has been revealed by using CART model. Main 6 factors for the accidents happening in Colombo – Batticaloa road are Speeding, Error of judgment, Aggressive/negligent driving, Fatigue/fall asleep, Influenced by alcohol/drugs, Distracted/inattentiveness. The results showed that the most vulnerable reason is speeding of the driver. In order to reduce the accident risks speed limits must be established with the road. Checkpoints must be established with barriers in order to avoid accidents. In addition, multiple machine learning and deep learning techniques were used to determine driver faulty. Among Decision tree, naïve Bayes, XG boost, SVM, KNN, Random Forest, DNN, MLP, and linear classifier, Multi-Layer Perceptron model has outperformed with 97.93% accuracy. Due to the dropout and optimization functions of the MLP, it has achieved highest accuracy becoming the best model to determine driver faulty. As the future works, researcher will be focus on implementing proposed method in order to gain real time benefits for the society.

References

1. "Road traffic injuries." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries. [Accessed: 30-May-2019].
2. X. F. Zhang and L. Fan, "A decision tree approach for traffic accident analysis of Saskatchewan highways," *Can. Conf. Electr. Comput. Eng.*, 2013.
3. G. Kaur and E. H. Kaur, "Prediction of the cause of accident and accident prone location on roads using data mining techniques," *8th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2017*, 2017.
4. S. Kumar and D. Toshniwal, "A novel framework to analyze road accident time series data," *J. Big Data*, vol. 3, no. 1, 2016.
5. Ss. Senior Lecturer and D. Ramani Professor, "Classification of Vehicle Collision Patterns in Road Accidents using Data Mining Algorithms," *Int. J. Comput. Appl.*, vol. 35, no. 12, pp. 975–8887, 2011.
6. M. Sameen and B. Pradhan, "Severity Prediction of Traffic Accidents with Recurrent Neural Networks," *Appl. Sci.*, vol. 7, no. 6, p. 476, 2017.
7. "The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark." [Online]. Available: https://medium.com/@srnghn/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3. [Accessed: 11-Jun-2019].
8. H. A. Park, "An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain," *J. Korean Acad. Nurs.*, vol. 43, no. 2, pp. 154–164, 2013.
9. E. Keogh, "Naïve Bayes Classifier," 2006.
10. A. Mariot, S. Sgoifo, and M. Sauli, "I gozzi endotoracici: contributo casistico-clinico (20 casi)," *Friuli Med.*, vol. 19, no. 6, 1964.
11. G. Louppe, "Understanding Random Forests: From Theory to Practice," . July, 2014.
12. "Classification with TensorFlow and Dense Neural Networks." [Online]. Available: https://heartbeat.fritz.ai/classification-with-tensorflow-and-dense-neural-networks-8299327a818a. [Accessed: 20-May-2019].
13. K. Y. Lee, N. Chung, and S. Hwang, "Application of an artificial neural network (ANN) model for predicting mosquito abundances in urban areas," *Ecol. Inform.*, vol. 36, pp. 172–180, 2016.
14. "Linear Classifier in TensorFlow: Binary Classification Example."