Research Article

Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach

Salahadin Seid Yassin[1] · **Pooja**[1]

Received: 14 February 2020 / Accepted: 22 June 2020 / Published online: 28 August 2020

Abstract
Road accident severity is a major concern of the world, particularly in underdeveloped countries. Understanding the primary and contributing factors may combat road traffic accident severity. This study identified insights and the most significant target specific contributing factors for road accident severity. To get the most determinant road accident variables, a hybrid K-means and random forest (RF) approaches developed. K-means extract hidden information from road accident data and creates a new feature in the training set. The distance between each cluster and the joining line of k1 and k9 calculated and selected maximum value as k. k is an optimal value for the partition of the training set. RF employed to classify severity prediction. After comparing with other classification techniques, the result revealed that among classification techniques, the proposed approach disclosed an accuracy of 99.86%. The target-specific model interpretation result showed that driver experience and day, light condition, driver age, and service year of the vehicle were the strong contributing factors for serious injury, light injury, and fatal severity, respectively. The outcome demonstrates the predictive supremacy of the approach in road accident prediction. Road transport and insurance agencies will be benefited from the study to develop road safety strategies.

Keywords Clustering · Classification · Model interpretation · Hybrid model · Road safety

1 Introduction

Road traffic accident (RTA) is churning the world with killing thousands and bringing demolition of property in a day without discrimination but did not give much attention to mitigate the severity. However, it is one of the life-threatening incidents in the world cause of death and property damage. Identifying the primary road traffic accident factors will help to provide an appropriate solution to minimize the adverse effect of severity on human and property loss. Road Severity does not occur by chance: It has patterns and can be predicted and avoided. So, accidents are "events which can be examined, analyzed, and prevented" [20]. According to workers' health organization, accidents defined as "Fatalities are not fated; accident does not just happen; illness is not random; they are caused [33]. Traffic accidents occurred daily in the capital city of Addis Ababa—Ethiopia. Human beings' life and property damage with a fraction of seconds. It is one of the leading terrifying causes of death in the country.

RTA severity is one of the research areas in these two decades in road safety. Researchers were using interesting methods on the road accident severity classification based models. The authors were studying using a traditional statistical-based approach for model building. These techniques help to get insights and identify the underlying cause of vehicle accidents and related factors on road safety. These days, due to the presence of a massive volume of datasets, machine learning surpasses conventional statistical-based in predicting the model [41].

✉ Pooja, pooja.1@sharda.ac.in; Salahadin Seid Yassin, salahadincs@yahoo.com | [1]Department of Computer Science and Engineering, Sharda University, Greater Noida 201310, India.

Many pieces of literature explained in different countries the causes of road traffic accident severity [7, 9, 36, 37, 40, 43, 45]. However, the road traffic accident severity prediction research is still in development. In the previous study, we have seen a room using a hybrid machine learning approach to improve classification accuracy. To fill the stated space, we work on a hybrid machine learning approach for road accident classification to improve the effectiveness of prediction accuracy. The previous study mainly works on the performance of the Machine learning-based classification approach. However, there is a dearth of comparing the state-of-the-art algorithms, Hybrid Machine Learning, and deep learning algorithms. Sometimes, obtaining a suitable approach will make prediction accuracy more informative. Hence getting of best paradigm helps to identify the most determinant road accident factors. Furthermore, target specific contributing factors were not concerned and identified previously. The study used hybrid clustering and classification algorithms to predict road accident severity prediction. In this work, a new hybrid K-means and random Forest algorithm proposed to predict target specific road accident severity. The proposed approach compared with individual classifiers to measure the performance of the developed model. Accuracy, precision, specificity, and recall used to compare the new approach and conventional techniques (SVM, KNN, LR, and RF). The new approach composed of the following phases: (I) removing disturbing noise and filling missing data using mean for numeric variables and mode for the categorical variable, (II) splitting the dataset into training and test dataset, (III) creating new feature using clustering, (IV) training classifiers, (V) finally evaluating the performance of individual classifiers. Moreover, the proposed approach compared with a deep neural network to evaluate further with another state of the art classifier techniques. The evaluation outcome showed the proposed better performs than other classifiers based on classification and performance metrics.

The rest of the study prepared as follows: In Sect. 2, existing research in road accident classification concerning the Machine Learning approach discussed. In Sect. 3 new Hybrid Based Machine Learning method using k-means and random forest is presented. Experiment, Evaluation, and Discussion are summarized in Sect. 4. In Sect. 5, results and analysis driven from the experiment are explained at the last conclusion presented in Sect. 6.

2 Litrature review

In the area of road safety traditional statistical model-based techniques were used to predict accident fatal and severity. Mixed logit modeling approach [23, 26], ordered Probit model [54], logit model [11] are few of adopted conventional statistical-based studies. some studies believed the conventional statistical model better identify dependent and independent accident factors [31]. But conventional statistical-based approach lacks the capability to deal with multidimensional datasets [16]. In order to combat traditional statistical models limitations; Nowadays many studies used ML approach due to its predictive supremacy, time consuming and informative dimension. In these decade ML approach employed in construction industry [48], occupational accident [41], agriculture [22], educational classification [53], sentiment classification [50] and in banking and insurance [46].

On the other hand, in road accident prediction, many studies performed using Data mining, machine learning, and deep learning algorithms.Among clustering and classification algorithms: K-means, Support Vector Machines, K-Nearest Neighbors (KNN) Decision Tree (DT), Artificial Neural Network (ANN), Convolution Neural Network (CNN) and Logistic Regression (LR) are in front to build accident severity model. Kwon et al. [28] adopted Nave Bayes (NB) and Decision Tree (DT) on California dataset collected from 2004 to 2010. Authors used binary regression to compare the performance of the developed model but Nave Bayes were more sensitive to risk factors than the Decision Tree model.

Sharma et al. [44] analyzed road accident data using SVM and MLP on a limited number of datasets (300 datasets). Besides authors used only two independent variables (alcohol and speed) as considering key factors. Eventually, SVM with RBF kernel gave better accuracy (94%) than MLP (64%). The study showed driving with high speed after drunk was the main reason for accident occurrence.

Wahab and Jiang [51]carried out crash accidents on Ghana dataset using MLP, PART, and SimpleCART intending to evaluate classifiers and to identify the major factors for motorcycle crash. Auhors used Weka tools to compare and analyze datasets and InfoGainAttributeEval applied to see the most influential variable for motorcycle crash in Ghana. As a result simpleCART model showed better accuracy than other classification models.

Kumar et al. [27] implemented kmeans and Association Rule data mining approaches to identify the frequency of accident severity locations and to extract hidden information. From the total 158 locations; 87 of them were selected after removing accident location frequency count less than 20. Then k-means were applied to cluster into three groups, Number of clusters are determined by gap statistics. To get rules, they used minimum support of 5 percent. As a result, curved and slop on the hilly surface were revealed as accident prone locations. Authors worked on the FARS data-set using data mining techniques to combat death and injury severity during 2007.

After prepossessing the study applied clustering Association rule and Nave Bayes to get trends of fatal accidents in the USA. The study explained and identify human and collusion types were the main cause of the fatality rate [34]. Other studies conducted using clustering and classification techniques to predict an accurate model in Iran. The research mainly focused on combining k means clustering with self-organizing maps to get better classification accuracy than ANN and ANFIS. The author's preference model better performs than the single classifiers [5]. AlMamlook et al. [6] used AdaBoost, Nave Bayes, Logistic Regression and random forest to get determinant factors and to identify high risky highways for Michigan traffic Agencies. Performance measurement ROC, AUC, Precision and recall and F1-score were applied to evaluate models. The Study showed random forest outperforms other classifiers with an accuracy of 75.5%. Tiwari et al. [47] conducted a data mining approach to analyze causality class traffic accidents. The authors implemented clustering like K-modes and SOM and classification techniques like NB, DT, and SVM. As a result, better accuracy was presented on cluster dataset over classification.

The existing study on road accident severity in Ethiopia see [2, 8, 10, 17, 25, 49]. These stated works concerned mostly on road accident analysis and pedestrian severity in Ethiopia using Statistical methods. on the other hand some studies employed a data mining techniques (Decision Tree and MLP) on Weka tool focusing mainly on driver responsibility [39]. Another study employed J48 and PART a data mining algorithm on driver and vehicle information considering as a major risk in accident severity on Weka tool [19]. Other related work in the country, Beshah [12] studied to identify the key road way related variables for accident severity in Ethiopia. Authors used a data mining approach (Decision Tree, Naive Bayes and KNN) to develop a decision rule to improve road safety. Their focus has been analyzing driver and pedestrian crashes without giving more attention to the influence of machine learning accuracy for better identification of major risks influencing in road accident in Ethiopia. At this time there is a great need for increasing road safety prevention study due to the growth of crashes. There is still room for improvement in the prediction accuracy of RTA in the case of Ethiopia to improve prediction accuracy in road safety. Therefore, we tried to develop a new hybrid approach to classify road accident severity by combining or collaborating clustering with classification, which will give remarkable classification results in road accident prediction. Clustering minimizes the sample dataset in the cooperation. Classification predicts road traffic severity. In this vein, clustering provides indirect cooperation for classification to extract hidden information from the training set to improve classifier performance.

3 Methodology

The study concerns mainly variable-based classification on road accident severity. It combines K-means clustering and classification technique to get better result than individual classifier. K-means employed to create new features and random forest used for classification prediction. The proposed approach workflow in the study is shown in Fig. 1. Major components of the flow chart are as follows:

3.1 Road accident dataset manipulation

3.1.1 Raw traffic accident data-set

Dataset in this study comprises 5000 road traffic accidents collected from federal traffic police agency reports from 2011 to 2018 in Addis Ababa, Ethiopia. One of the challenging parts of this research is collecting sample datasets from the organization. The original dataset collected from the authority handled manually. Most of the fields are incomplete. Some of the fields are useless, and the main vital fields did not include as a field on the manual document. Documents were invisible to read. They are written either heedlessly or in a rush. We forced to record in excel format to ease analysis and prediction. In each instant accident, 14 variables (ten categorical variables and four numeric variables were recorded. Among these variables, Severity class is a target variable with three values (Fatal, Severe Injury and Light Injury). Full Dataset description described in the following manner:

Accident time This variable implies the time on which road traffic severity occurred throughout the day (24 hours).

Driver age This variable shows the age of the driver. Drivers age mainly in the range of 18-80 years of age.

Sex This variable indicates the driver's sex. The driver's sex (observed from the data collection) is either male or female.

Drivers experience This variable indicates the driver's experience. It mainly represents the duration of how much time the driver drives a car.

Type of vehicle This variable implies the different types of vehicles. Namely: ambulance, car, automobile, Isuzu, taxi, truck, motorcycle, pick up, bus and minibus.

Service year This variable indicates the year of service the vehicle gives to the owner.

Location This variable indicates where the accident occurred: namely: canteen area, pubic area, organization, government office, hospital, college, vehicle station, market living area, and hospital.

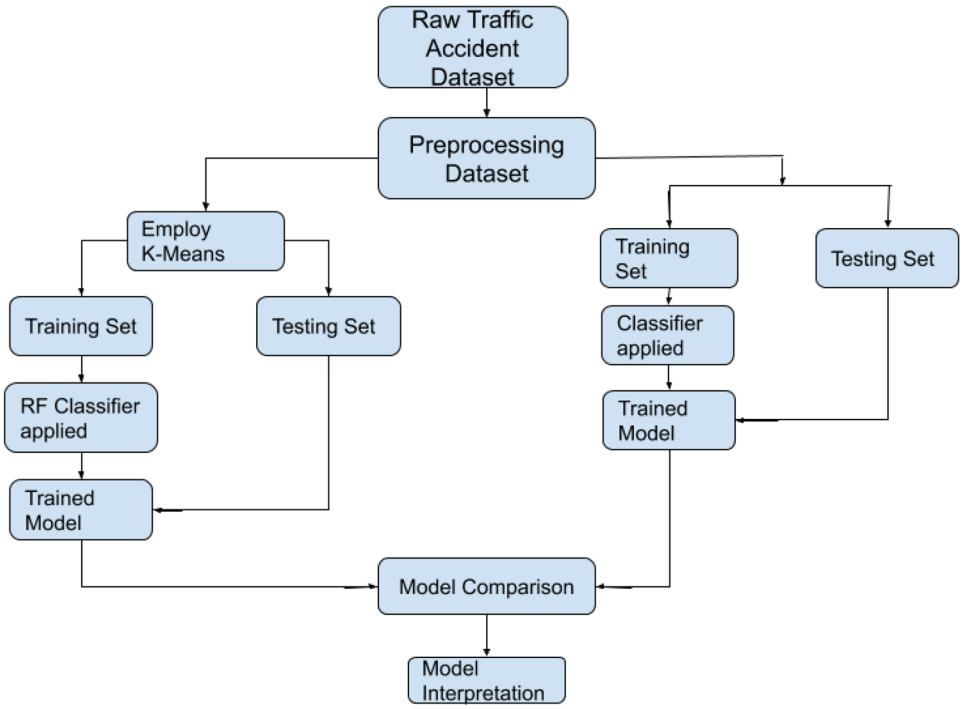

Fig. 1 Flowchart of proposed model framework for predicting road traffic accident—case of Ethiopia

Road condition This variable shows the situation of the road during the accident. Variable represents Namely: dry, muddy and wet.

Light condition This variable connotes the situation of the road during the accident. Variable represents Namely: dry, muddy and wet.

Weather condition This variable indicates the climate condition during the accident. Variable represents namely: rainy, sunny, cold and windy.

Causality class This variable indicates the severity of the class. The variable represents namely: driver, passenger, pedestrian, cycle driver, and resident.

Causality age This variable indicates the severity of class age.

Causality sex This variable indicates the severed class sex male or female.

Severity This variable is the target variable represents three classes, namely: fatal, serious injury and light injury.

3.1.2 Preprocessing

Raw datasets were sadly dirty, not in a proper format to be understood by computing machines and give incomplete information to use as it is. Using Such datasets will reduce the efficiency of the accident severity prediction model. Therefore, irrelevant datasets need to remove to obtain quality data. In the study before building a model intensive data preprocessing technique employed to get meaningful and determinant risk factors Like Data cleaning, missing value handling, outlier treatment, dealing with absolute value—encoding and normalization are carefully purify before using it.

3.1.3 Splitting dataset

Raw datasets and k-means created features split into training set and testing sets. The aforementioned training set helps to learn the newly proposed method. On the other hand, the testing set used to measure the performances of the new proposed model. In the study, a 70:30 ratio is used to split the raw dataset. Then 70% is used to train the prediction model, whereas, 30% of the dataset used to evaluate the performance of the prediction classification accuracy.

3.1.4 Prediction model

A prediction model is mainly used in machine learning techniques to forecast future behavior by analyzing current and historical data.

3.2 K-means techniques

K-means Technique [35] is unsupervised Machine learning technique mainly used in statistical data analysis, image processing, signal processing, information retrieval. The presence of heterogeneity in a road accident may lead to wrong model building and prediction. Unobserved heterogeneity defined as the presence of critical unseen features correlated with the observed

feature in a model building.To overcome this problem, we are engaged in using clustering in our accident dataset. The split of datasets based on its similarity makes homogenous within clusters and heterogenous between clusters. Besides, clustering in collaboration with classification makes the classifier to train a model with a short time, more accurate and needs less computational memory when dealing with a massive amount of dataset [29]. K-means technique works on M data points as input in the N dimension in initial k cluster centrists, k is user defined to determine the total number of clusters; as a result after claculating there distance from each cluster data points assigned to each nearest cluster. Hartigan and Wong [24]. All points within a cluster are closer in the distance to their centroid than they are to any other centroid. The primary goal of the K-means technique is to reduce the Euclidean distance D(Xi, Cj) between each point from the centroid. as a result intra-cluster variance can be reduced and inter cluster similarity increases. Squared error function represented in Eq. 1.

$$f(x) = \sum_{n=1}^{k} \sum_{n=1}^{n} |X_i - C_j|^2 \quad (1)$$

where k is number of clusters, n-number of cases and Cj-number of centroids and X is data points of which Euclidean distance from the centroid is calculated. K means algorithm has initialization and iteration phases. In the first phase data points assign randomly in to k clusters, then in iteration phase the algorithm calculate the distance between each data points to each cluster centers, finally the algorithm converges when each road accident data points assigned to the nearest cluster [24]. Let us see how K-means algorithm works as follow:

1. Randomly initialize and select the Cj-centroids.
2. Calculate the distance between each instance to the Cj-centroid.
3. Compute mean of each data points in each cluster to find their centroid.
4. Then repeat the aforementioned steps until each points assigned to their nearest cluster.

3.3 Random forest

Random forest is ensembled classification technique proposed by Breiman and Adele Cutler mainly works building multiple tress to make uncorrelated decision trees [13]. It is one of the robust algorithms to predict a large number of datasets. Mainly decision tree prone to overfitting but random forest uses multiple tresses to reduce overfitting [13]. The random forest creates many shallow, random subset trees and then combine or aggregate subtrees to avoid overfitting. Also, when it employed in large datasets gives more accurate predictions and cannot relinquish its accuracy when it faces several missing data. Random forest combines multiple Decision Tress during training then takes the aggregate of it to build model. Therefore, weak estimators improve when they are combined. Even if some of the decision trees become weak, there overall desired output results tend to be accurate. Figure 2 illustrates sample random forest implementation.

3.4 Proposed approach

These days road accident datasets are stored in a vast database repository. A large number of datasets make the training and testing phase more complicated and reduces predicting efficiency. Therefore, it needs a powerful model to overcome or minimize the complexity of a huge amount of dataset. We developed a hybrid K-Means and Random forest model to get a better efficient predictive model to

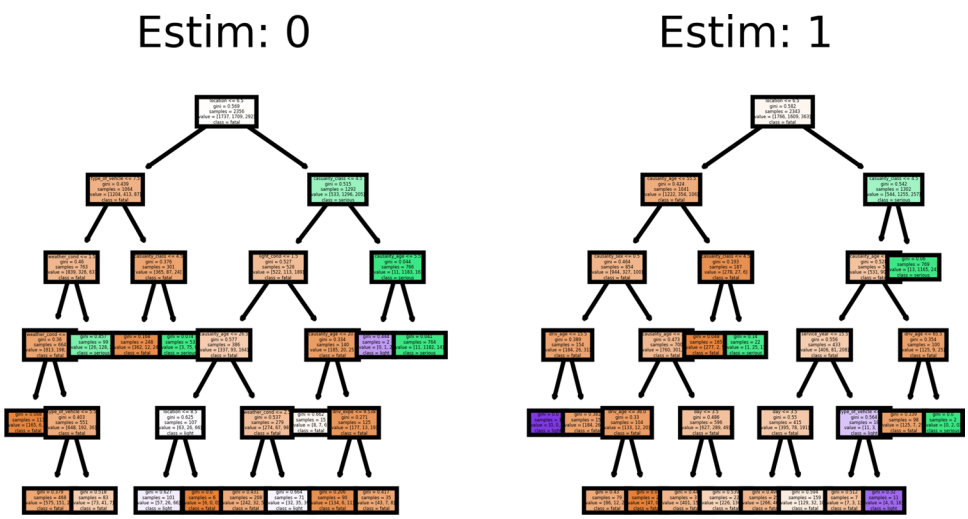

Fig. 2 Sample random forest (n-estimator= 5)

enhance the efficiency and accuracy of prediction model. K-means normally, which is an unsupervised machine learning algorithm mainly used to find similar groups within the dataset. Even though this is an unsupervised technique, k-means can create new features for the training set to improve the performance of classifier. Clustering creates cluster feature and adds to the training set. Then random forest employed on clustered training data to classify severity of RTA. There combination will produce a powerful prediction model in terms of generalization performance and predictive accuracy.

4 Experiment, evaluation, and discussion

In this section, the preprocessing technique applied to the road accident dataset, evaluation metrics, and experimental result analysis presented.

4.1 Dataset manipulation

The Dataset collected from Addis Ababa City is not entirely clear and organized. Raw dataset recorded manually and prone to damage. However, it must be in a machine-understandable format to get meaningful information and to develop an efficient intelligent system. The road accident severity prediction model depends on the quality of the datasets. We used different types of data preprocessing techniques to clean the dataset.

- *Missing value handling* Missing value treatment is a mandatory task in data preprocessing. Before building model missing values needs to be filled using a different strategy. In the dataset, some attribute values are missing. Building an exciting and well-performing prediction model on incomplete data will not give a decisive output. It ought to handle wisely and either ignore or must be filled using different methods to get a better result [43]. Ignoring or dropping values is an approach to handle missing values, but dropping may lead to missing valuable information. In the study, missing value is not forced to drop missing attributes. Figure 3 presents a number of missing values and its percentage from the total dataset.The missing value is less than 50 percent of the total population. we employed substituting feature mean for numeric variables and most frequent (mode) value for categorical variable [3]. For further,on Preprocessing our previous work gives detail information, see Ref. [43].
- Categorical Value Encoding Raw traffic accident datasets consists of categorical and numeric values. However, many machine learning algorithms require numeric values to predict a model. Employing a machine learning algorithm on categorical values are a challenging problem. Therefore, categorical values should be either converted into numeric values or needs to be removed [32]. In the dataset, most of the variables are categorical; Among 14 variables, 10 of them are categorical values and needed to transform into a numeric format. predictive variables and target variables converted into numeric using one-hot encoding and label encoding respectively.

	Missing Values	% of Total Values
service_year	1128	22.6
driv_expe	898	18.0
type_of_vehcle	571	11.5
driv_age	538	10.8
sex	422	8.5
causality_age	320	6.4
location	315	6.3
causality_sex	166	3.3
light_cond	157	3.2
day	131	2.6
casuality_class	110	2.2
road_cond	105	2.1
severity	74	1.5
weather_cond	1	0.0

Fig. 3 The number of missing values and their percentage in the RTA dataset

4.2 Experimental system set up

The study implemented using python 3.7 on Jupyter notebook as IDM and intel core i7 1.80GHz processor speed CPU, 8Gb RAM, and 1TByte HD system. In this section, different experiments like Choosing an optimal value of k, evaluating the proposed approach, and finally comparing with conventional algorithms with the new approach presented.

4.3 Evaluation metrics

In the study, different types of evaluation, metrics are used to measure the performances of the proposed approach to predict road accident training set as indicated from Eqs. 1

to 6. Namely: accuracy, specificity, precision, recall and F1 score [38]

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$Specificity = \frac{TN}{FP + TN} \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$F1Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \quad (6)$$

TP: it shows predictive is positive and it is normally true
TN: it implies predictive is Negative and it is normally True
FP: denotes predictive is positive and it is normally false
FN: represents predictive is negative and it is false. Where TP implies true positive, TN denotes true negative, FP indicates false positive, and FN denotes false negative. in the actual study values are represented by true and false whereas predictive values denoted by positive and negative.

5 Experimental result analysis and discussion

5.1 Train-test split

Once the dataset prepared to train the model. It splits into a training set and testing set. The former used to learn the classifier, whereas the later used to test the performances of the predictive model. In the study, It is applied into 70:30 ratio, 70% of the proportion used to train model and 30% for the testing set applied to evaluate trained model.

5.2 Choosing k

There is no specific solution to find the exact value of k to partition training dataset. For each k, we can initialize k-means and use the inertia attribute to identify the sum of squared distances of the training set to the nearest cluster center. When k increases, the sum of squared distance leans towards zero and the percentage of variances increased as shown in Fig. 4a, b. If we use k to its maximum value in the M training set, each training set will form its cluster. Figure 5a. below is a plot of the sum of squared distances for k. If the plot looks like an arm, then the elbow on the arm is optimal k. However, from the graph, the elbow is not clear to determine the optimal value of k. Then we created line joining the first and last points (i.e. $K = 1 \text{and } k = 9$) (Fig 5b illustrates line creates to connect $k = 1 \text{and } k = 9$). Then we calculated the distance between each cluster and the line to find the maximum distance. Figure 6a, b shows values of a distance of each k points from the line. The maximum length is index 2 (i.e 3.63). so we could say that the exact optimal value of k is three and road accident dataset clustered into three groups based on the experimentation.

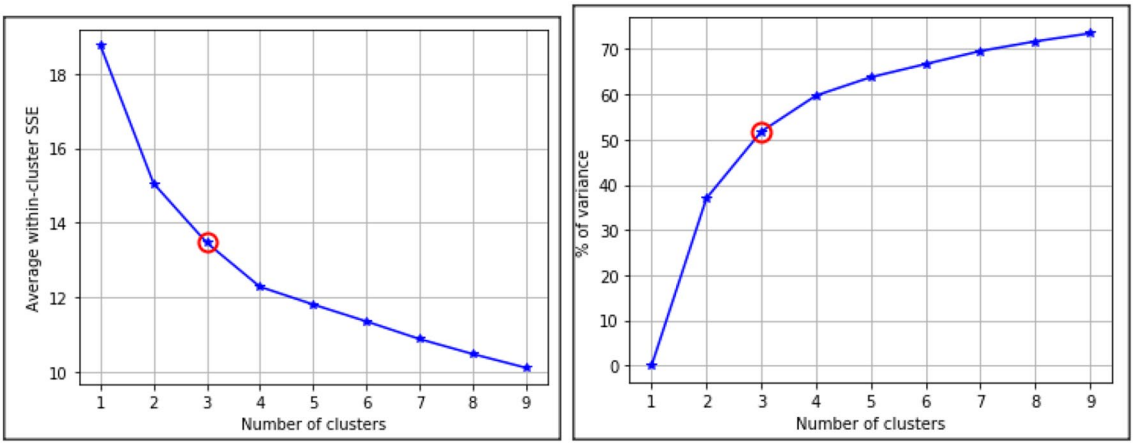

Fig. 4 **a** The average distance within clusters (SSD) and **b** the percentage of variance between clusters

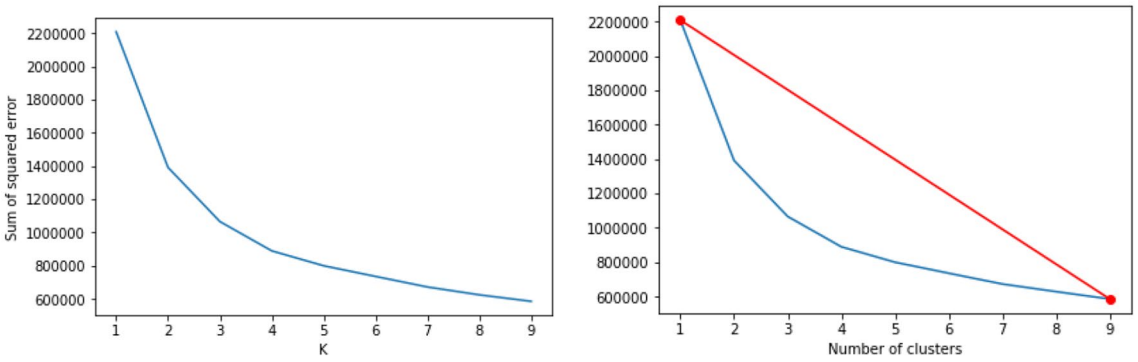

Fig. 5 **a** Elbow technique for optimal k, **b** lines created from k = 1 to k = 9

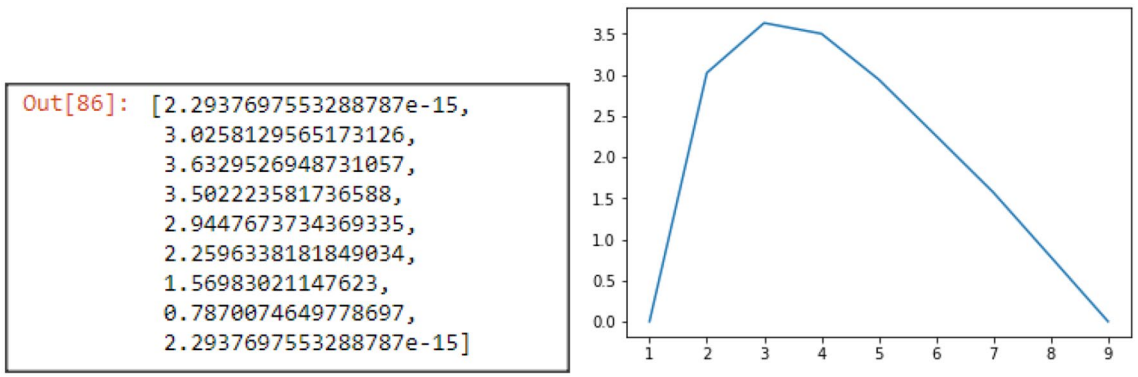

Fig. 6 **a**, **b** Calculated distance values from each k or cluster to the line (value of k = 3)

5.3 Model performance evaluation

In this section, evaluation of the performance and reliability of the model, and comparing the proposed approach with the conventional models discussed briefly.

The study employed the k-Means algorithm on a raw accident dataset to cluster into three groups based on given k value. The newly created cluster used as a new feature and added to the training set. An experiment performed on both the raw dataset and a new feature added training set. when we emplloy unsupervised K-means algorithm on raw dataset scores an accuracy of 42.25% whereas supervising machine learning algorithms like logistic regression, random forest, support vector machine, and k-Nearest Neighbors performance accuracy on a raw dataset scored 86.83%, 87.77%, 68.45%, and 64.97% respectively. While unsupervised and supervised machine learning techniques applied to a new feature added training set its performance of K-means, logistic regression, random forest, support vector machine, and k-Nearest Neighbors scored an accuracy of 35.83%, 99.13%, 99.86%, 73.13% and 68.58% respectively. The experiment revealed performances of each classifier on various classifier metric on both data set showed all classification algorithms perform well on all evaluation metrics except k means classifier. An excellent performer on both datasets is random forest. Especially on the cluster added dataset, the efficiency of the random forest algorithm performed very well. Its performance dramatically improved to 99.86% accuracy. Each supervised machine learning classifier achieved a promising result. Classification techniques performance showed a mouthwatering efficiency, especially random forest performance accuracy heightened from 87.77 to 99.86%. But unsupervised k means classifier performed somehow better on a raw dataset. In Table 1 performance evaluation of each model is presented before and after adding a new feature to the dataset. Result discovered the proposed model has better performance than other models. Table 2 presents the execution time of each model. Astonishingly KNN model had less execution time than other models.

In the study, proposed approaches compared with other related studies. Table 3 shows the previous papers that worked on road accident severity prediction using different types of methodology. Our proposed Hybrid approach used k-means from clustering and random forest

Table 1 Performance evaluation of classifiers and proposed approach

S. No	Testing set without new feature				Testing set with new feature				
	Classifier	Precision	Recall	f1 score	Accuracy	Precision	Recall	f1 score	Accuracy
1	K Means	47	42	43	42.25	36	36	35	35.83
2	LR	85	87	84	86.83	99	99	99	99.13
3	RF	86	88	87	87.77	100	100	100	99.86
4	SVM	69	68	65	68.45	76	73	70	73.13
5	KNN	64	65	62	64.97	68	69	66	68.58

Table 2 The execution time of models (ms)

Model	Training time	Testing time
K-means	191	2.57
LR	231	1.29
RF	399	38
SVM	566	134
KNN	9.7	87
K-means-RF	295	5.71

from classification to improve severity model accuracy and more importantly designed to identify target specific contributing factors for road accident severity. The proposed approach infrequently used in the related study and target specific classification was not concerned. This makes the study unique and significant in Ethiopia.

5.4 ANN experiment analysis

In this paper, we created a baseline ANN for road accident model prediction for different dense layers. Rectifier and Sigmoid activation function used as input and output layers respectively. Table 4 presents the performance of Artificial Neural network (ANN) with different dense layers. Experiment result showed the best test accuracy seen by two and three dense layers. Both model (Model 1 = 88.77% and Model 2 = 88.77%) achieved better result than the third model (model 3 = 88.03%). However as presented in Fig. 7 model 2 has low test loss value (0.3622) than model 1 (0.3819) and model 3 (0.3686) relatively. But test loss is not as such attractive in multi-class classification. On the other hand Fig. 8 presents the AUC metric values of each neural network model with different amounts of dense layers. As it has seen all models with different dense layer gives similar results.

5.5 Comparative of neural network and proposed models

In this study, the ANN model compared with the proposed Hybrid model. The Comparative performance of both models showed that the proposed model (Hybrid K means and random forest) performed better than the ANN model in terms of Precision, Recall, F1 score,

Table 4 Test accuracy, loss, and ROC curve value of ANN model with multiple dense layers

Model	Dense layer	Test accuracy (%)	Test loss	ROC curve (%)
$Model_1$	2	88.77	0.3819	96.1
$Model_2$	3	88.77	0.3622	96.1
$Model_3$	4	88.03	0.3686	96.1

Table 3 Performance comparison of related work models

References	Classifier	Dataset	Accuracy
Gu et al. [21]	PSO-SVM	China	–
Xiao et al. [52]	SVM, KNN (Ensemble)	I-880 data set	99.33%
Castro et al. [15]	BN, JR8 and MLP	DVSA—UK	72.39%, 72.02%, 71.70% Respectively
Al-Radaideh et al. [4]	RF, ANN (backpropagation), SVM	Uk	80.6%, 61.4%, 54.8% respectively
Casado et al. [14]	LCC, MNL	Spain	–
Wahab et al. [51]	MLP. SimpleCart, PART	Ghana	72.16%, 73.45%, 73.81% respectively
Sameen et al. [40]	MLP, BLR, RNN	Malaysia	65.48%, 58.30%, 71.77% respectively
Fentahun [18]	J48, ID3, PART	Ethiopia	81.21%, 81.01%, 81.18%
Seid et al. [42]	HMR	Ethiopia	NA
Abebe et al. [1]	DSA	Ethiopia	–
Lytin et al. [30]	UBA	Ethiopia	–

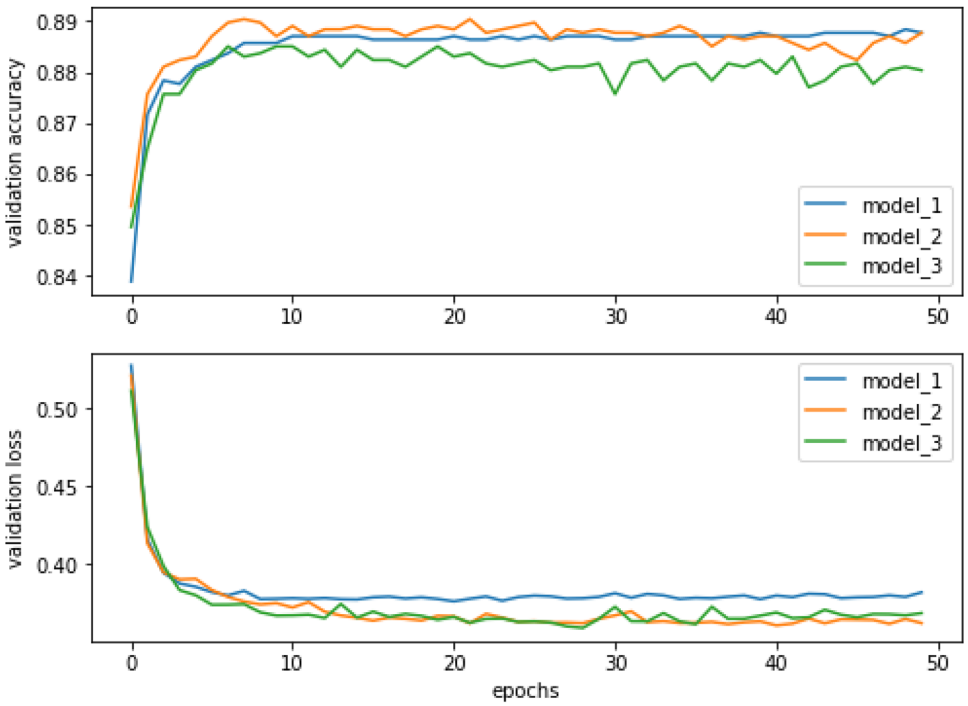

Fig. 7 The validation and loss accuracy of different ANN Model

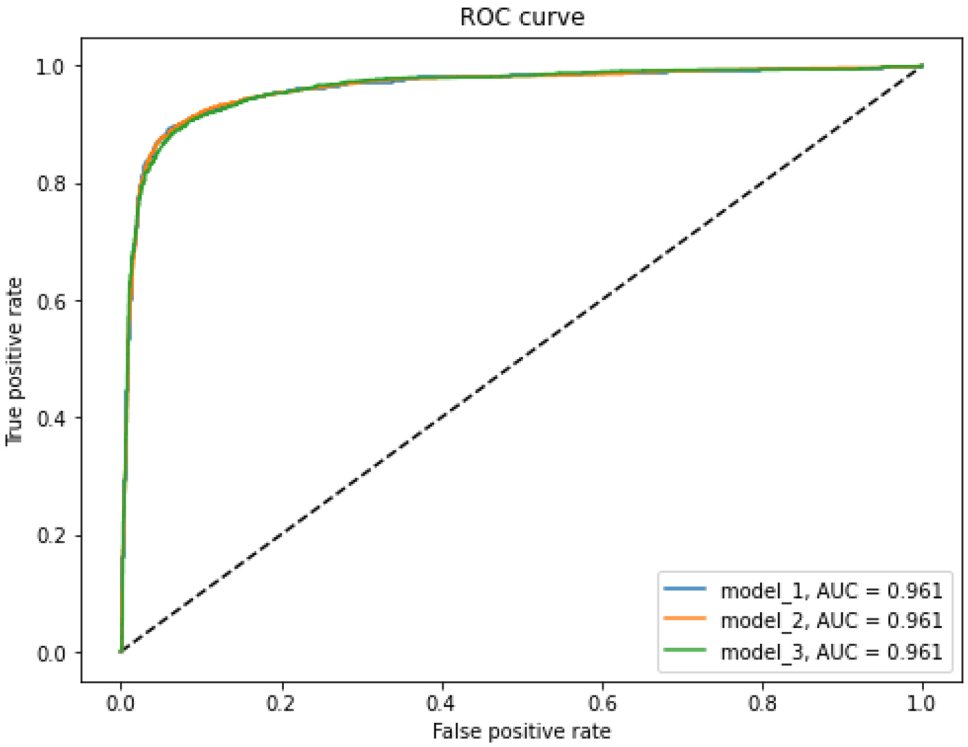

Fig. 8 ROC curve of different ANN models

and Accuracy. Table 5 presents Performance comparison of proposed and ANN models. The proposed model achieved a better result than a deep neural network.

Table 5 Comparison of ANN and proposed model performance with different metrics (%)

Model type	Precision	Recall	F1 score	Accuracy
ANN	88	88	88	88
Proposed model	100	100	100	99.86

5.6 Random forest interpretation

Random forest builds numerous decision trees for several subsets of RTA variables. It is commonly called a black box-difficult to know how it processed inside the model. Indeed, it comprises of many decision trees. Examining each deep tree decision and process is troublesome and improbable. Whereas individual tree could learn on bagged data on randomly selected features [13]. So, we can get insights from the random forest on computing feature importance. Before going to see how random forest works lets see how decision tree works. It has a series of decision paths from the node to the last leaf safeguarded by a sub-feature. Prediction is a sum of individual features and bias (mean value of top-most region covered by training set). Decision tree prediction function defined as:

$$f(x) = Cfull + \sum_{m=1}^{M} contrib(x, k) \tag{7}$$

where M—number of leaves in the tree, k—the number of features, C full—root node value, Contrib(x, k) kth feature contribution in feature vector x. Now let's move to the prediction of random forest, which is as discussed in Sect. 3.2 an average value of its tree prediction. Therefore, random forest prediction function defined as follows:

$$f(x) = \frac{1}{J} \sum_{j=1}^{J} f_j(x) \tag{8}$$

It is pretty clear that the random forest prediction is the average value of bias and the average value of each contribute feature set. Which can be defined as

$$f(x) = \frac{1}{J} \sum_{j=1}^{J} C_j full + \sum_{k=1}^{k} (\frac{1}{j} \sum_{j=1}^{J} + contrib_j(x, k) \tag{9}$$

The above expression explained how random forest black box processed by following decision routes through the tree and compute the contributions of individual features. knowing the relatedness of predictive variable to the prediction model either negatively or positively helps to understand detail information about the model. which helps to know the influence of each variable on the outcome.

In the experiment, the default parameter set up used to implement the random forest algorithm. we have seen that day, driver experience, type of vehicle, location, light condition, causality age, and casualty sex are the strong contribution for serious injury, light condition, causality class, causality age, and causality sex are a contributor for light injury whereas driver age, service year, weather condition and causality class are a strong contribution for fatal accident severity.

6 Conclusion

In the study, a hybrid-based approach developed to predict the severity of the RTA dataset. The approach is competitive and better than traditional machine learning algorithms. In the case of creating a new cluster feature and finally added to the training set, k means used and showed a convincing result when combined with classification algorithms. In the paper, K-Means used to group road accident dataset based on its similarity and random forest employed to classify road accident factors into the severity variable. The combination of K-Means with random forest outperforms other Conventional models, namely Logistic Regression, k Nearest Neighbor, and Support Vector Machine. The classification technique used in the experiment improves classification accuracy values for logistic regression, random forest, support vector machine, and k nearest neighbor are 12.3, 12.09, 4.8, and 3.61 respectively. On the contrary, k means decreased its accuracy value by 6.42. The experiment result revealed that adding a new cluster on the training set has a strong impact to improve classification accuracy. Random Forest got better accuracy (99.86%). Before clustering and classification data Preprocessing performed to purify raw datasets. Missing value treatment and conversion of categorical values done to get a better result. In the paper optimal value of k discovered after calculating the maximum distance from each cluster to the joining line from K1 to Kn. Also, to trust a prediction model interpretation made to understand how a model inside processes to predict. Prediction is a sum of bias and contribution features. In the experiment, we showed results to get insights into the contributing variables for the prediction model. Knowing the influence of individual variables on the prediction model is trustworthy. Moreover, In the study target-specific, variable contribution explained. Overall, the paper tried to show the effects of combining Clustering and Classification to improve model accuracy and identified major contributing factors class-specific wise from the collected data for road traffic accident datasets. Another dataset will straighten our model to get a better result.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Abebe Y, Dida T, Yisma E, Silvestri DM (2018) Ambulance use is not associated with patient acuity after road traffic collisions: a cross-sectional study from Addis Ababa, Ethiopia. BMC Emerg Med 18(1):7
2. Abegaz T, Gebremedhin S (2019) Magnitude of road traffic accident related injuries and fatalities in Ethiopia. PLoS one 14(1):e0202240
3. Acurna E, Rodriguez C (2004) The treatment of missing values and its effect in the classifier accuracy, classification, clustering, and data mining applications. In: Proceedings of the meeting of the International Federation of Classification Societies (IFCS), pp 639–647
4. Al-Radaideh QA, Daoud EJ (2018) Data mining methods for traffic accident severity prediction. Int J Neural Netw Adv Appl 5:1–12
5. Alikhani M, Nedaie A, Ahmadvand A (2013) Presentation of clustering-classification heuristic method for improvement accuracy in classification of severity of road accidents in Iran. Saf Sci 60:142–150
6. AlMamlook RE, Kwayu KM, Alkasisbeh MR, Frefer AA (2019) Comparison of machine learning algorithms for predicting traffic accident severity. In: 2019 IEEE Jordan international joint conference on electrical engineering and information technology (JEEIT). IEEE, pp 272–276
7. Ansari S, Akhdar F, Mandoorah M, Moutaery K (2000) Causes and effects of road traffic accidents in Saudi Arabia. Public Health 114(1):37–39
8. Asefa F, Assefa D, Tesfaye G (2014) Magnitude of, trends in, and associated factors of road traffic collision in Cntral Ethiopia. BMC Public Health 14(1):1072
9. Balogun J, Abereoje O (1992) Pattern of road traffic accident cases in a Nigerian University teaching hospital between 1987 and 1990. J Trop Med Hyg 95(1):23–9
10. Baru A, Azazh A, Beza L (2019) Injury severity levels and associated factors among road traffic collision victims referred to emergency departments of selected public hospitals in Addis Ababa, Ethiopia: the study based on the Haddon matrix. BMC Emerg Med 19(1):2
11. Bedard M, Guyatt GH, Stones MJ, Hirdes JP (2002) The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. Accid Anal Prev 34(6):717–727
12. Beshah T, Hill S (2010) Mining road traffic accident data to improve safety: role of road-related factors on accident severity in Ethiopia. In: 2010 AAAI Spring symposium series
13. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
14. Casado-Sanz N, Guirao B, Attard M (2020) Analysis of the risk factors affecting the severity of traffic accidents on spanish crosstown roads: the drivers perspective. Sustainability 12(6):2237
15. Castro Y, Kim YJ (2016) Data mining on road safety: factor assessment on vehicle accidents using classification models. Int J Crashworthiness 21(2):104–111
16. Chen WH, Jovanis PP (2000) Method for identifying factors contributing to driver-injury severity in traffic crashes. Transp Res Rec 1717(1):1–9
17. Deme D (2019) Road traffic accident in Ethiopia from 2007/08-2017/18. Am Int J Sci Eng Res 2(2):49–59
18. Fentahun A (2011) Mining road traffic accident data for predicting accident severity to improve public health-role of driver and road factors in the case of Addis Ababa. PhD thesis, Addis Ababa University
19. Getnet M (2009) Applying data mining with decision tree and rule induction techniques to identify determinant factors of drivers and vehicles in support of reducing and controlling road traffic accidents: the case of Addis Ababa city. Addis Ababa Addis Ababa University
20. Gissane W (1965) Accidentsa modern epidemic. J Inst Health Educ 3(1):16–18
21. Gu X, Li T, Wang Y, Zhang L, Wang Y, Yao J (2018) Traffic fatalities prediction using support vector machine with hybrid particle swarm optimization. J Algorithms Comput Technol 12(1):20–29
22. Habib MT, Majumder A, Jakaria A, Akter M, Uddin MS, Ahmed F (2018) Machine vision based papaya disease recognition. J King Saud Univ Comput Inf Sci 32(3):300–309
23. Haleem K, Alluri P, Gan A (2015) Analyzing pedestrian crash injury severity at signalized and non-signalized locations. Accid Anal Prev 81:14–23
24. Hartigan JA, Wong MA (1979) Algorithm as 136: a k-means clustering algorithm. J R Stat Soc Ser C (Appl Stat) 28(1):100–108
25. Hordofa GG, Assegid S, Girma A, Weldemarium TD (2018) Prevalence of fatality and associated factors of road traffic accidents among victims reported to burayu town police stations, between 2010 and 2015, Ethiopia. J Transp Health 10:186–193
26. Kim JK, Ulfarsson GF, Shankar VN, Mannering FL (2010) A note on modeling pedestrian-injury severity in motor-vehicle crashes with the mixed logit model. Accid Anal Prev 42(6):1751–1758
27. Kumar S, Toshniwal D (2016) A data mining approach to characterize road accident locations. J Mod Transp 24(1):62–72
28. Kwon OH, Rhee W, Yoon Y (2015) Application of classification algorithms for analysis of road safety risk factor dependencies. Accid Anal Prev 75:1–15
29. Kyriakopoulou A, Kalamboukis T (2008) Combining clustering with classification for spam detection in social bookmarking systems. In: ECML PKDD
30. Laytin AD, Seyoum N, Kassa S, Juillard CJ, Dicker RA (2020) Patterns of injury at an Ethiopian referral hospital: using an institutional trauma registry to inform injury prevention and systems strengthening. Afr J Emerg Med 10(2):58–63
31. Lee C, Saccomanno F, Hellinga B (2002) Analysis of crash precursors on instrumented freeways. Transp Res Rec 1784(1):1–8
32. Lee N, Kim JM (2010) Conversion of categorical variables into numerical variables via bayesian network classifiers for binary classifications. Comput Stat Data Anal 54(5):1247–1265
33. Leka S, Griffiths A, Cox T, World Health Organization et al (2003) Work organisation and stress: systematic problem approaches for employers, managers and trade union representatives. World Health Organization, Geneva
34. Li L, Shrestha S, Hu G (2017) Analysis of road traffic fatal accidents using data mining techniques. In: 2017 IEEE 15th international conference on software engineering research, management and applications (SERA). IEEE, pp 363–370
35. MacQueen J et al (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Oakland, CA, USA, vol 1, pp 281–297
36. Odero W, Khayesi M, Heda P (2003) Road traffic injuries in Kenya: magnitude, causes and status of intervention. Inj Control Saf Promot 10(1–2):53–61
37. Persson A (2008) Road traffic accidents in Ethiopia: magnitude, causes and possible interventions. Adv Transp Stud 15:5–16
38. Powers DMW (2011) Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. J Mach Learn Technol 2(1):37–63
39. Regassa Z (2009) Determining the degree of driver's responsibility for car accident: the case of Addis Ababa traffic office. Addis Ababa University, Addis Ababa
40. Sameen MI, Pradhan B (2017) Severity prediction of traffic accidents with recurrent neural networks. Appl Sci 7(6):476

41. Sarkar S, Vinay S, Raj R, Maiti J, Mitra P (2019) Application of optimized machine learning techniques for prediction of occupational accidents. Comput Oper Res 106:210–224
42. Seid M, Azazh A, Enquselassie F, Yisma E (2015) Injury characteristics and outcome of road traffic accident among victims at Adult Emergency Department of Tikur Anbessa specialized hospital, Addis Ababa, Ethiopia: a prospective hospital based study. BMC Emerg Med 15(1):10
43. Seid S et al (2019) Road accident data analysis: data preprocessing for better model building. J Comput Theor Nanosci 16(9):4019–4027
44. Sharma B, Katiyar VK, Kumar K (2016) Traffic accident prediction model using support vector machines with Gaussian kernel. In: Proceedings of fifth international conference on soft computing for problem solving, Springer, Berlin, pp 1–10
45. Singh SK (2017) Road traffic accidents in India: issues and challenges. Transp Res Procedia 25:4708–4719
46. Sundarkumar GG, Ravi V (2015) A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance. Eng Appl Artif Intell 37:368–377
47. Tiwari P, Kumar S, Kalitin D (2017) Road-user specific analysis of traffic accident using data mining techniques. In: International conference on computational intelligence, communications, and business analytics. Springer, Berlin, pp 398–410
48. Tixier AJP, Hallowell MR, Rajagopalan B, Bowman D (2016) Application of machine learning to construction injury prediction. Autom Constr 69:102–114
49. Tulu GS (2015) Pedestrian crashes in Ethiopia: identification of contributing factors through modelling of exposure and road environment variables. PhD thesis, Queensland University of Technology
50. Vinodhini G, Chandrasekaran R (2016) A comparative performance evaluation of neural network based approach for sentiment classification of online reviews. J King Saud Univ Comput Inf Sci 28(1):2–12
51. Wahab L, Jiang H (2019) Severity prediction of motorcycle crashes with machine learning methods. Int J Crashworthiness 24:1–8
52. Xiao J (2019) SVM and KNN ensemble learning for traffic incident detection. Phys A 517:29–35
53. Yahya AA (2017) Swarm intelligence-based approach for educational data classification. J King Saud Univ Comput Inf Sci 31(1):35–51
54. Zajac SS, Ivan JN (2003) Factors influencing injury severity of motor vehicle-crossing pedestrian crashes in rural Connecticut. Accid Anal Prev 35(3):369–379