International Journal of Scientific Research in Computer Science and Engineering
Vol.8, Issue.4, pp.105-115, August (2020)

Research Paper

E-ISSN: 2320-7639

Predicting Factors of Vehicle Traffic Accidents Using Machine Learning Algorithms: In the Case of Wolaita Zone

Aklilu Elias Kurika[1*], Tigist Simon Sundado[2]

[1,2]Department of IT, School of Informatics, Wolaita Sodo University, Sodo, Ethiopia

**Corresponding Author: akliluelias123@gmail.com/akliluelias@gmail.com, Tel +251 91 648 5472, +251 94 903 2208*

Available online at: www.isroset.org

Received: 28/Apr/2020, Accepted: 06/Jun/2020, Online: 31/Aug/2020

Abstract- Vehicle traffic accident is the ultimate and major agenda for government in which special attention has been given to continuously reduce its occurrence and related risks. Wolaita zone is one of the major areas in which increased vehicle traffic accident occurs. Government and concerned bodies have given special attention to reduce accident rate in the country. By having this point as the motivating factor for study, this work tried to predict factors of vehicle accidents by using machine learning algorithms. We used unbalanced datasets with 1611 instances which was seven years data from 2005-2011 E.C. In order to analyze data and evaluate patters of datasets, KDD process model was applied. The learning algorithms applied for experiments were decision tree classifiers (J48, Random forest and Rep tree, Bayesian classifiers (Naïve Bayes and Bayesian network). The experimental results, model evaluation and performance measurement shows that F-measure of J48 and Rep tree classifiers are comparatively similar i.e. 97.87% and 97.80% respectively and Random Forest tree performed less i.e. 90.9%. We identified the 1st experiment of J48 tree as the best model by performance and 23 best rules were generated from this experiment; best features were also identified. The most common victims, most commonly participated vehicles in accident and black spot areas for frequent accidents occurrences were identified. The findings of this study are significant for road and traffic authority and police commission for the revision and endorsement of the rules, regulations and standards related to traffic accidents; and therefore vehicle traffic accidents and related risks can be reduced generally in our country Ethiopia and specially at Wolaita Zone. We made accident data ready for further analysis in order to get most important patterns of datasets for any future researchers.

Keywords— Vehicle traffic accident, Decision Tree, Bayesian Classifiers, Machine Learning Algorithms, Performance measurement

I. INTRODUCTION

Road or vehicle traffic accident is a universal problem and worldwide reports show that on average, more than four million peoples die because of many reasons in one year Micheale **[**1**]**. Among this numbers, HIV AIDS and tuberculosis are the first and second cases for the deaths and vehicle traffic accident is the third known case for those dying on every day. More than half the people killed by vehicle crashes were young adults aged between 15 and 44 years often the breadwinners in a family. Furthermore, road vehicle accident injuries low cost income and middle-income countries between 1% and 2% of their gross national product; which is more than the total development aid received by these countries according to WHO and World Bank **[**2**]**. This study also shows that in worldwide, an estimated 1.2 million people were killed by road vehicle accidents each year and as many as 50 million were injured. Statistics shows that every year, 1.2 million people were known to die by road accidents worldwide. The study shows that in the 2020, vehicle traffic accident will be the first factor to cause death of human beings in the world as stated by Guardian **[**3**]**. A lot researches were conducted on accidents in every parts of the world to reduce the accident rate and they used their own view on accident data according to their respective areas and country perspectives.

Even though plenty of researches were conducted, vehicle traffic accident increases rapidly and results in massive loss of humans' life, materials damage and other equivalent losses. Projections indicate that these figures will increase by 65% over the next 20 years unless there is new commitment to prevention.

Increased loses and related injuries caused various problems to the economic development of respective countries. According to different countries perspectives, there are diverse kinds of attributes and contributing factors of vehicle traffic accidents. Accident risk factors were more over determined in the developed countries and some preventive measures has been taken to reduce it. But traffic accident risks, related material damages and life loses increase from time to time in developing countries. These points are the motivating factors for this study to be conducted. In case of Ethiopia, some researches have been conducted, but the risk factors couldn't be reduced from time to time. In the case of Wolaita Zone, timely recorded data reality on the ground shows that traffic accident is the major issue ought to be given special attention. The reason

is that risks of traffic accidents and related material and live loses show enormous increase from time to time. But the reasons for increased traffic accident factors are not well known. Additional deep analysis on accident data indeed expected and this is also a motivating factor to conduct study by machine learning algorithms.

Therefore the purpose of this study is to predict factors of vehicle traffic accidents using machine learning algorithms in order to determine most determinant attributes to the occurrence of increased accident rate, the most common victims of the accident, the most commonly participated vehicles in accident, the black spot areas for frequent accident occurrences, the best machine learning algorithms for analysis, generate important rules for the occurrence of accidents, build the predictive model and finally to evaluate performance of the model. All these objectives were attained finally as we can see from the experimental results.

II. RELATED WORK

The road features are one of contributing factors of traffic accidents and they are related to locations of accident related factors; Accident data is basic to identify these features **[**7**]** used small amount of secondary data, but types of road features were not clearly specified [8], [9] and [20]; the two wheeler vehicles involvement rates determined accident accidents prone locations, other types of vehicles were not considered yet to determine the most common accident occurring areas according to the researcher **[**9**]**.

Amount and type of data (primary or secondary) data used for study also matters the researchers to build model with better performance **[**9**]**, **[**10**]** and **[**13**]**. This data was small, and it was both primary and secondary (social media data) data which is collected in questionnaire. Secondary data is not feasible for analysis as all of data scientists know. The problem of these studies was that researchers used secondary data; another limitation is that the method used was not scientific and finally there is no evaluation parameter for performance and accuracy of his work. Only decision tree algorithms were used by **[**8**]**; Studies performed by authors **[**11**]** & **[**16**]** were on the selected features of data sets to determine symbolic descriptions. Here; author used only one algorithm; another issues related to accidents were not considered yet.

A comparative analysis in the performance measurement and accuracy of algorithms were studied in detail by authors [10] and [11]. The first author compared six algorithms (classification and regression tree, random forest, ID3, functional trees, naïve bayes and J48) algorithms to determine accidents severity level. It revealed that naive bayes value and J48 techniques value were approximately same in accuracy. The second one is comparative study on machine learning algorithms; the comparison has been made for decision tree and neural networks to determine factors of increased traffic injury. It explored decision trees are better than neural networks in performance.

The definite factors of traffic accident were conducted and identified by different researchers and their findings show that causality factors were un-adopted speech, in-attention, behavior of passengers, roadway features, demographic features, environmental characters, technical characters, speed, age, gender, younger aged drivers, alcohol, less control, wrong over-taking and tire blow [10], [11], [12], [20], [21] and [24]. These factors were identified in various areas as the contributing factors for accidents. But it is impossible to blindly take control measures over all these characteristics to be considered in particular area.

Akinbola et al., used machine learning algorithms to predict the factors of traffic accidents [14] and [15]. Classification and machine learning algorithms were used to determine traffic injury occurrences by Gupta and Baluni [11]. Both of these authors used only decision tree; and Tibebe et al., was all about machine learning algorithm but it was not for determining causes of traffic accidents [16]. Experimental findings show that majority of participants in vehicle traffic accidents were females aged between 30 to 59 years, with primary or secondary education levels. By using multivariate logistic regression models, the researchers revealed that white people accounted for 48.1 % of participants and 61.2 % were those living with partners [22].

Works in classification algorithms and artificial intelligence has also comparatively similar findings as represented in [25] and [26].

Generally amount of data used by some of researchers was small and not suitable for analysis like social media data; secondary data which is collected by questionnaire. Using such kind of data for predicting factors of vehicle traffic accident is not feasible. Most of studies were conducted only by J48 decision tree algorithms. Performance comparisons have been made for only two algorithms; only three types of vehicles being participated in traffic accident occurrences were identified. In the case of Wolaita, there are various vehicles from smallest to the heavy ones (vans and trucks) flow on the road day to day. Most of researchers used only decision tree algorithm; Bayesian networks and Naïve Bayes and decision tree algorithms were not widely used. Accuracy of predictive model for accident occurrence was also not good i.e. 85% and recommended to be tried again with large amount of data [26].

So predicting factors of vehicle accident is expected to identify the most commonly contributing factors that hold a lion share. In Ethiopia, Wolaita zone is one of the most commonly known areas in which traffic accidents and related injuries take place. By predicting factors with machine learning algorithms, the most contributing factors was determined from traffic accident data which is obtained from Wolaita police commission.

III. METHODOLOGY

To address the problems the researcher proposed knowledge discovery in datasets (KDD) process modeling as study design and its possible steps are given as follows diagrammatically.

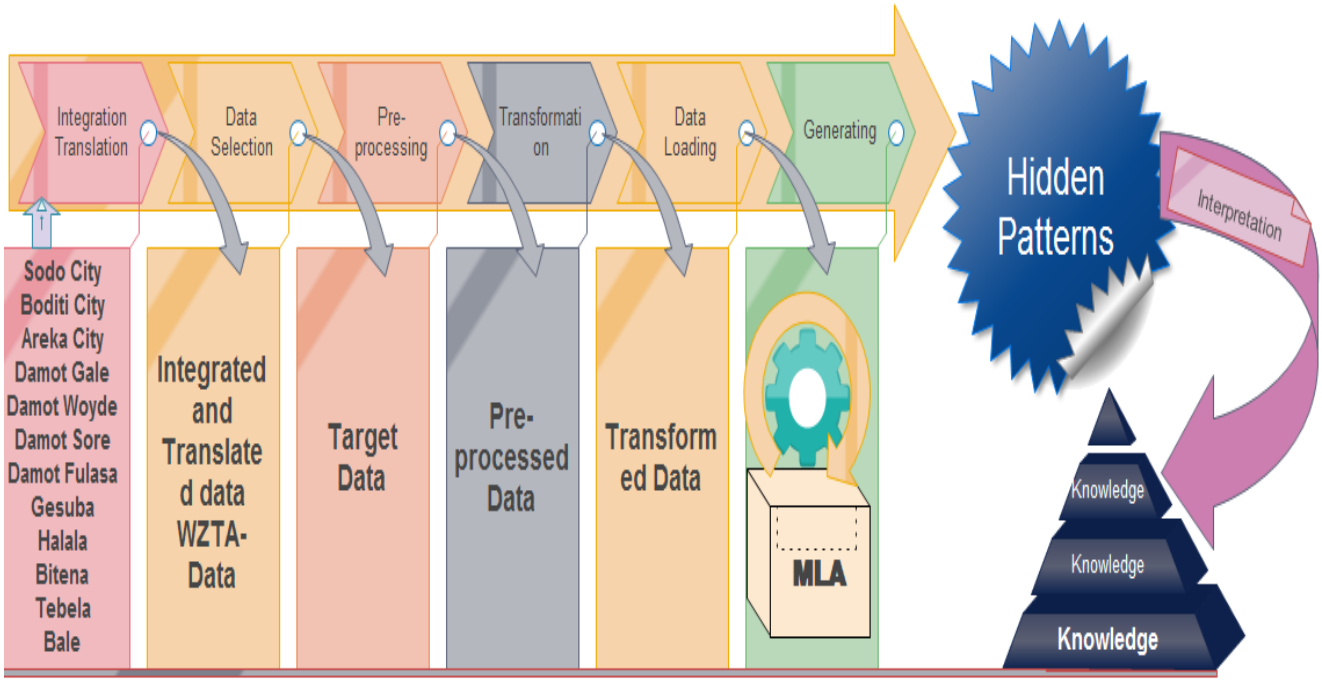


Figure 1 Study design

A. Data Integration: To keep normal compliance of data, we integrated data to common format according our objectives and identified most important attributes to our study. Some of attributes were ignored from original data because they are less meaningful to our study. Accordingly, 36 important attributes were identified and 1611datasets were prepared for analysis, which is 7 years data from 2005-2011 E.C. The amount of data was limited to 1611; because five years (2000-2004) data was burned before it was being transformed to police commission from road and transport authority.

B. Data Selection

In order to get data for prediction, applicable data was selected from 12 districts and three city administrations of Wolaita Zone. The case study is limited to Wolaita zone only. This is because we wanted to define the scope of our study only Wolaita.

C. Data Preprocessing: In this step, data cleaning, data reduction and data transformation has been made to prepare the best quality datasets for further analysis. Original data was obtained from Wolaita Zone police commission (PC) but, it has a lot of drawbacks such as spelling errors, unreadable data, misspelt attributes names, unknown values for some attributes and irrelevant personal representations of some terms. Some terms were inconsistent and considered to be outliers. We removed irrelevant attributes from original Data. In this step we made cleaning process of data before loading it to WEKA.

D. Data Transformation: The original data was in word processor. Some data were in spread sheet or excel document. We transformed it to the .CSV format which the WEKA work bench supports. Loading data to WEKA is the next step after data transformation.

E. Algorithm Selection: Classification algorithm has been identified as the best technique to attain our objectives in accordance with predetermined datasets we had. From various classification algorithms, decision tree classifiers (J48, Random Forest and Rep Tree) classifiers and from Bayesian classifiers (Naïve Bayes and Bayesian Network) classifiers were selected to conduct our experiments. We have computed 15 experiments, (three for each classifiers i.e. by 10 fold cross validation, by 66% split and by 90% split for each of them respectively.) We have identified 14 best features among 36 attributes with wrapper method.

F. Knowledge Generation: Finally the researcher generated hidden knowledge with proposed algorithms for the prepared datasets; and reported findings.

IV. EXPERIMENTS AND RESULT DISCUSSION

A. Most Commonly Participated Vehicles

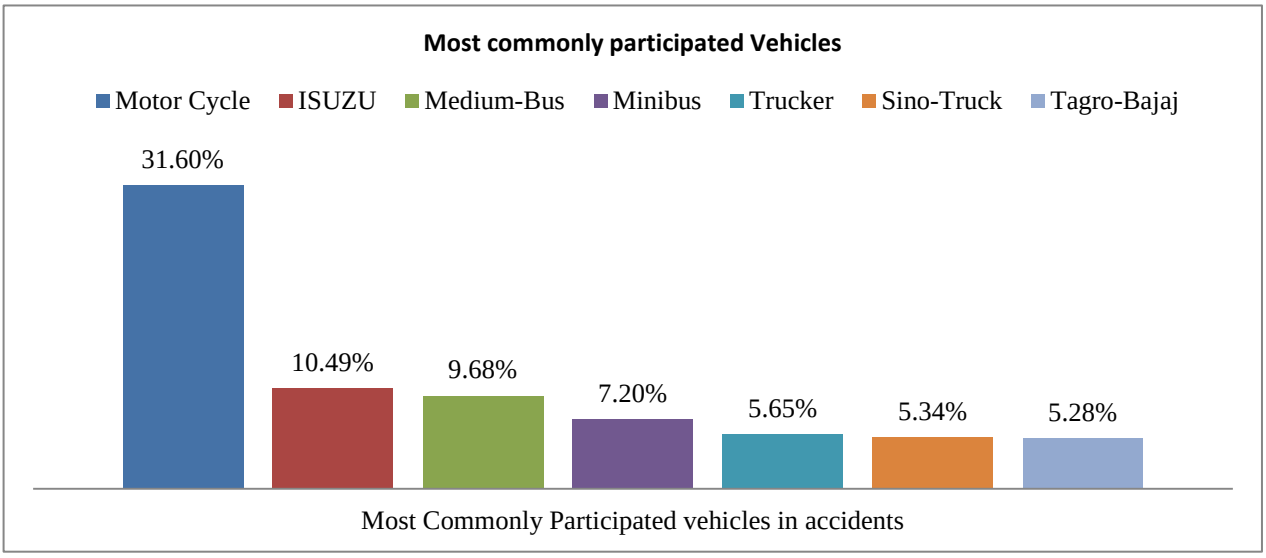
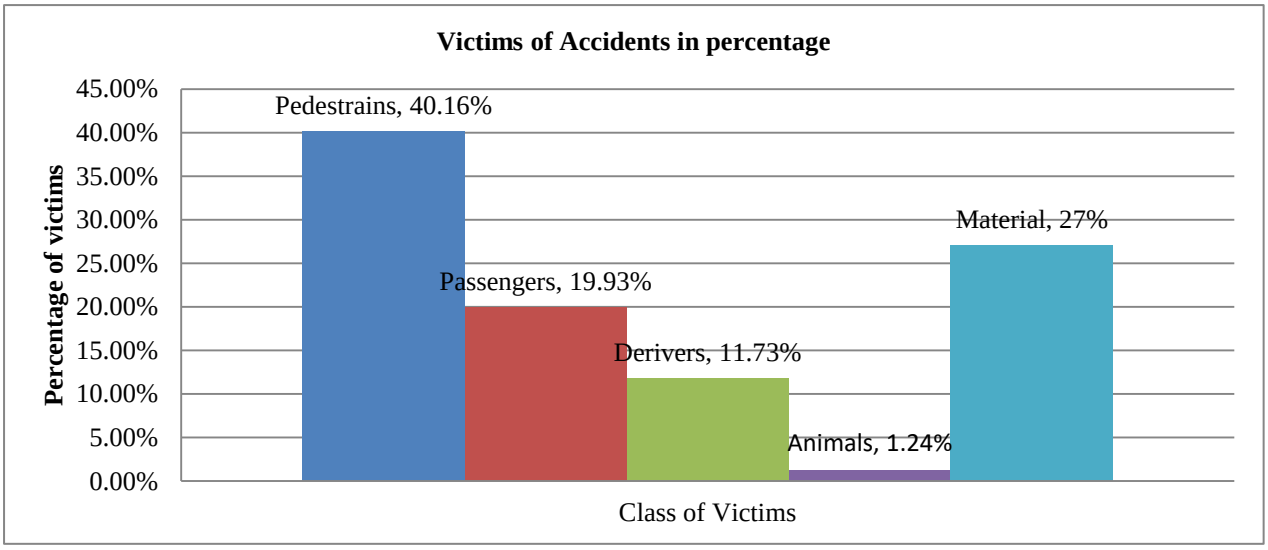
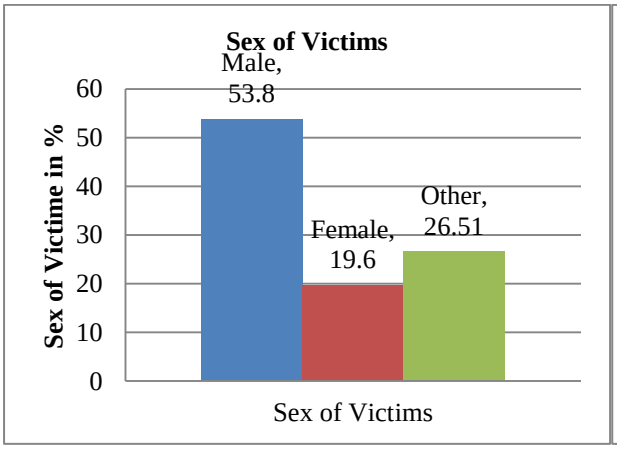


Figure 2 most commonly participated vehicles.

From the total 31 kinds of vehicles participated in accidents, we have identified 7 kinds vehicles as the most commonly participated. But only Vans and Trucks as most commonly participated vehicles were identified by [21]. They account 75.34% and remaining 24 vehicles participation is only 24.66%. So we can conclude that if these vehicles were given separate road in cities specially Sodo-City (>25%) traffic accident can be possibly reduced.

B. Most Common Victims of Accidents





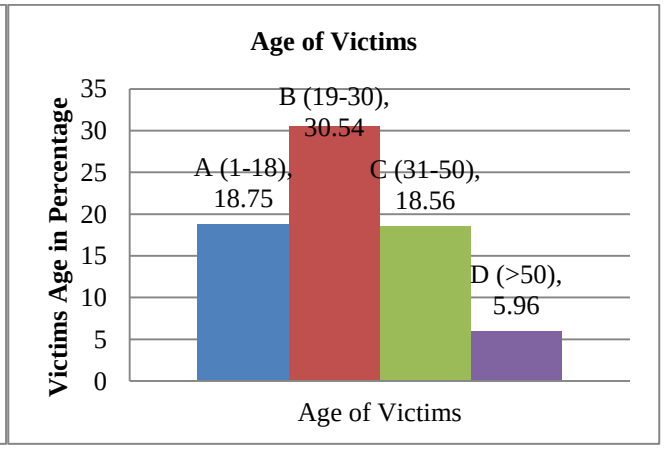


Figure 3 Most common victims

The above diagram shows that the most common victims of accidents are pedestrians (40.16%) and passengers (19.93%). Derives are less victims. So we can conclude that car traffic accident most commonly affects pedestrians and passengers in our case study.

Males (53.8%) are most commonly affected by car traffic accidents compared to females (19.6%); which are opposite to study which revealed majority of participants as females in accidents [22] and [23]. 18.75% of victims were aged between 1-18, 30.54% of victims were aged between 19-30 and 18.56% of victims were aged between 31-50.

As it is known, the most productive human power is aged between 18 and 50. Therefore traffic accident affects the most productive classes of humans as we can conclude from the above result.

C. Most Common Black Spot Areas

We have selected 19 places with frequent accident occurrences from the above five Woredas. We selected areas with > = 15 accidents within 7 years. From the total accidents occurred, these places account 521 (32.34%) accidents. So concerned bodies must give attention to these areas.

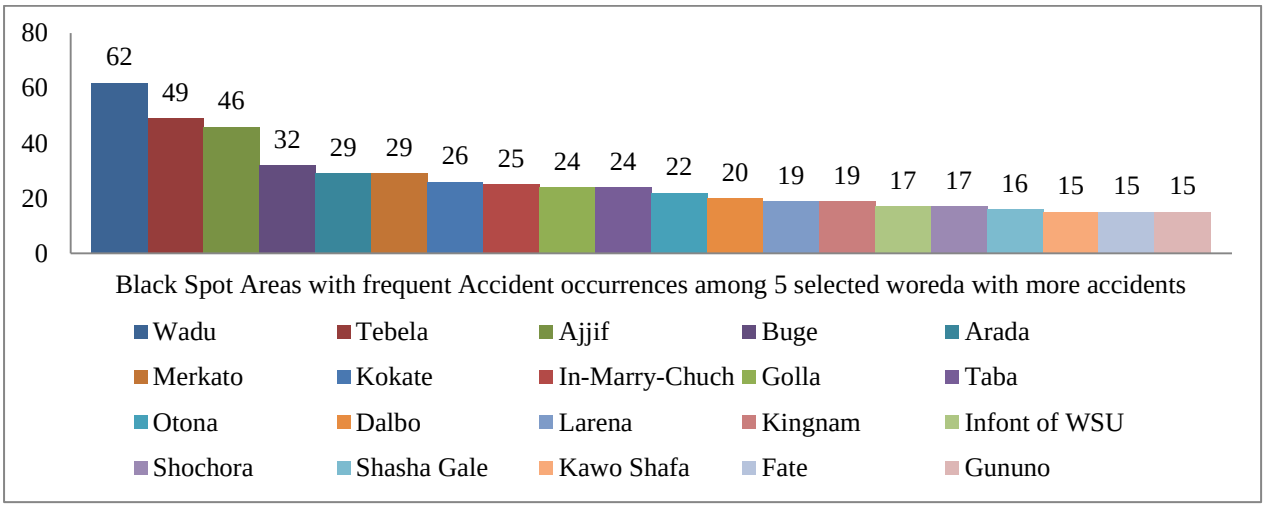


Figure 4 Accident Occurrence Places

From 15 different areas shown above, the first five (Sodo-city, Damot-Gale, Humbo, Sodo-Zuria and Boditi-City) account a lot accidents i.e. **73.37%** of total accidents. The remaining 10 districts account only **26.63%.** Each of them accounts > **5%** accident occurrences from the total one, so we selected the black spot areas for frequent accidents occurrences from these five Woredas.

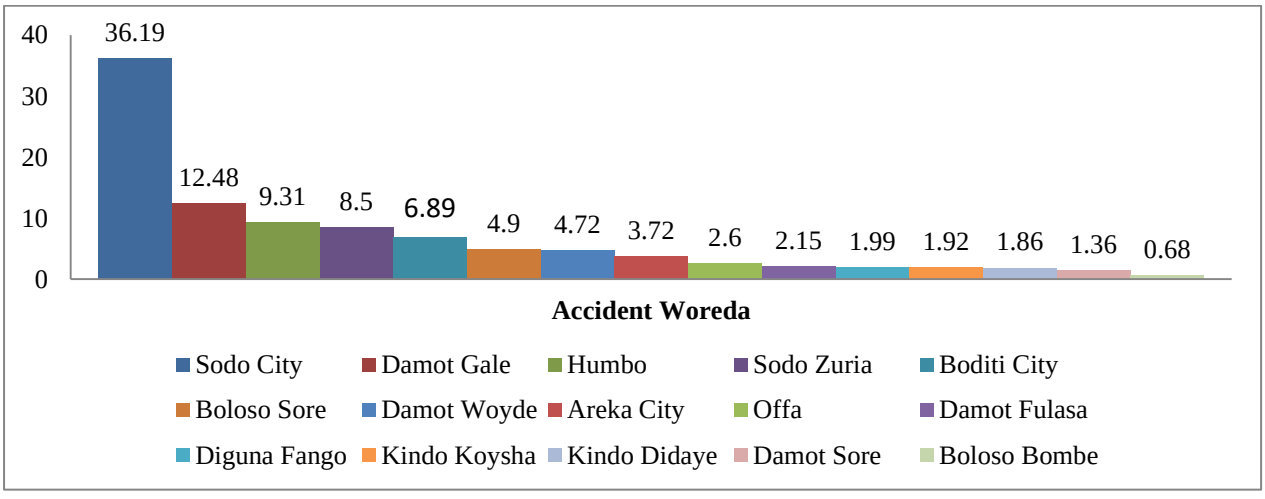


Figure 5 Most Common Black spot areas

D. Determinant Cases of Accidents- The Most Determinant Cases and causality condition of Accidents are: Lack of attention (65.49%), over speed (10.62%), Prohibiting Priority (10.37%), lack of experience (6.33%) and technic failure (3.54%)

The causality condition of accidents is mostly crossing the road (32.96%) straight crash (28.80%), roll down (16.70%), side to side crash (8.57%) and walking on the road (5.90%)**.**

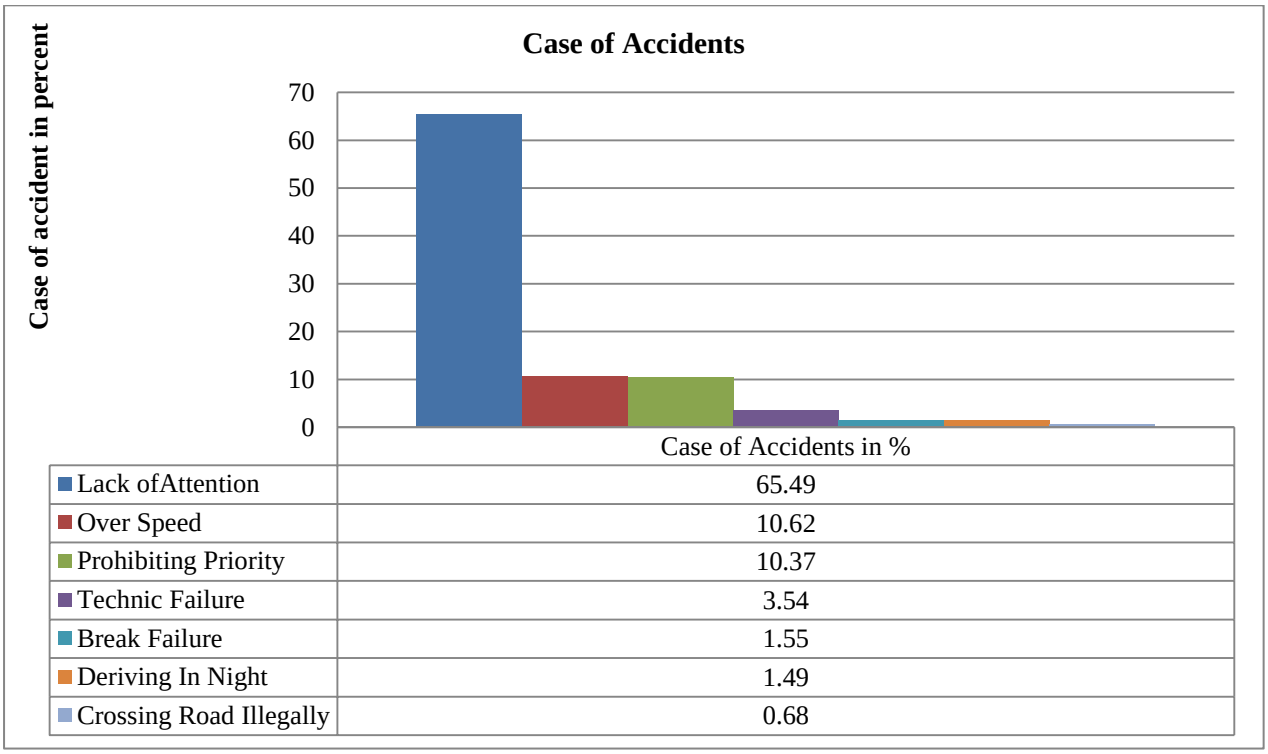


Figure 6 (a) Determinant cases of accidents

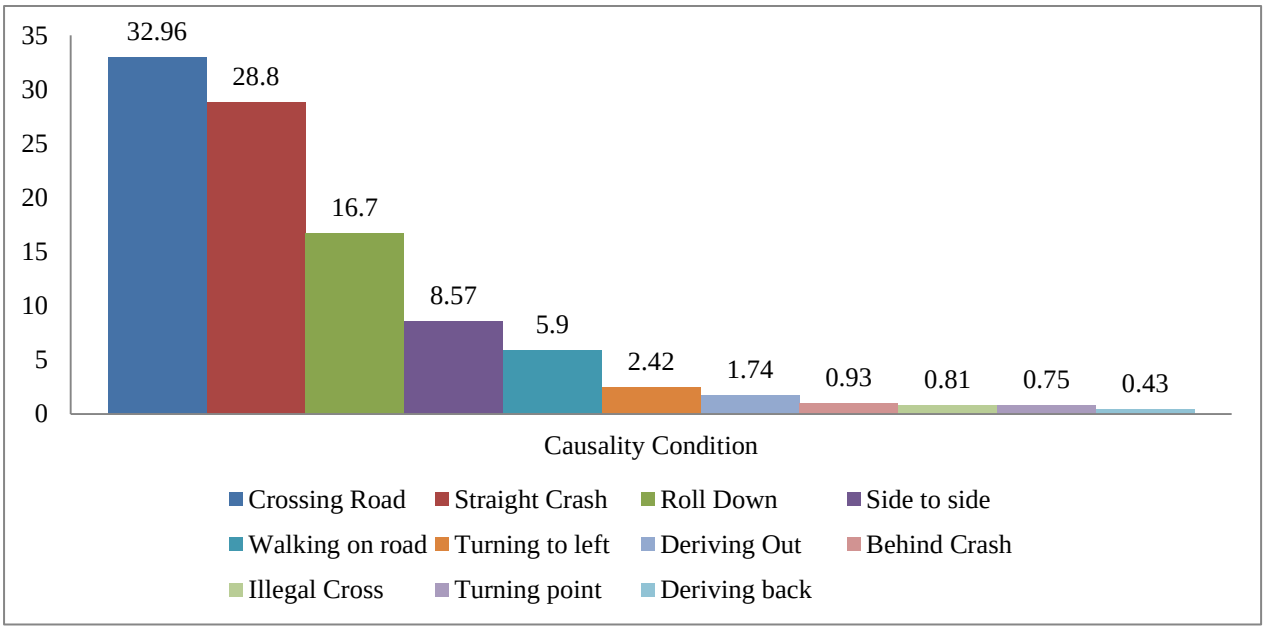


Figure 6 (b) Determinant cases of Accidents

Table 1 Summary of Experimental Results

Exp	Models	NL	ST	TP Rate	FP Rate	Precision	Accuracy
Exp.1	Trees.J48 -C 0.5-M 4 Testmode=10-fold Datasets=Unbalanced Attributes=All	141	145	0.984	0.030	0.984	98.45%
Exp.2	Trees.J48 -C 0.5-M 4 Testmode=Split=66% Datasets=Unbalanced Attributes=All	4	5	0.989	0.015	0.989	98.90 %
Exp.3	Trees.J48 -C 0.5-M 4 Testmode=Split=90%	4	5	0.989	0.0014	0.989	98.91%

	Datasets=Unbalanced Attributes=All						
Exp.4	RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1 Testmode=10-fold Datasets=Unbalanced Attributes=All	-	-	0.921	0.237	.926	92.12 %
Exp.5	RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1 Testmode=Split=66% Datasets=Unbalanced Attributes=All	-	-	0.905	0.325	0.914	90.51 %
Exp.6	RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1 Testmode=Split=90% Datasets=Unbalanced Attributes=All	-	-	0.901	0.301	0.912	90.06 %
Exp.7	trees.REPTree-M 2-V 0.001-N 3-S 1-L-1-I 0.0 Testmode=10-Fold Attributes=All Datasets=Unbalanced	4	5	0.984	0.026	.984	98.386 %
Exp.8	trees.REPTree-M 2-V 0.001-N 3-S 1-L-1-I 0.0 Testmode=Split=66% Attributes=All Datasets=Unbalanced	4	5	0.989	0.015	0.989	98.905 %
Exp.9	trees.REPTree-M 2-V 0.001-N 3-S 1-L-1-I 0.0 Testmode=Split=80% Attributes=All Datasets=Unbalanced	4	5	0.991	0.003	0.991	99.07 %
Exp.10	Bayes.NaïveBayes-output-debug-info Testmode=Split=90% Attribute= All Dataset=Unbalanced	-	-	0.946	0.068	0.948	94.60 %
Exp.11	Bayes.NaïveBayes-output-debug-info Testmode=split=66% Attribute= All Dataset=Unbalanced	-	-	0.954	0.066	0.956	95.438
Exp.12	Bayes.NaïveBayes-output-debug-info Testmode=split=90% Attribute= All Dataset=Unbalanced	-	-	0.969	0.027	0.971	96.894 %
Exp.13	Weka.Classifiers.bayes.net Testmode=10-Fold Attribute= All Dataset=Unbalanced	-	-	0.942	0.061	0.946	94.165 %
Exp.14	Weka.Classifiers.bayes.net Testmode=split=66% Attribute= All Dataset=Unbalanced	-	-	0.954	0.048	0.957	95.44 %
Exp.15	Weka.Classifiers.bayes.net Testmode=split=90% Attribute= All Dataset=Unbalanced	-	-	0.969	0.010	0.972	96.894 %

Key: Exp: Experiment, NL: Number of Leaves, ST: Size of Tree, TP: True Positive, FP: False Positive

As we can see from the above experimental results and below diagram, **J48** and **Rep** tree classifiers are comparatively similar by their accuracy.

We computed average Precision and Recall of **J48** and **Rep tree** and selected the J48 decision tree algorithm as a better than Rep tree.

1st Expt J48 tree Precision = 98% and Recall = 97.75%, (FM= 97.87%)✓

1st Expt. Rep tree Precision = 97.70% and Recall = 97.90%, (FM= 97.80%)☒

The first experimental results of J48 decision tree, includes more features than exp.2 and 3 even though the number of leaves and size of tree generated are more.

So we selected it as a working model and generated 23 best rules from this particular experiment.

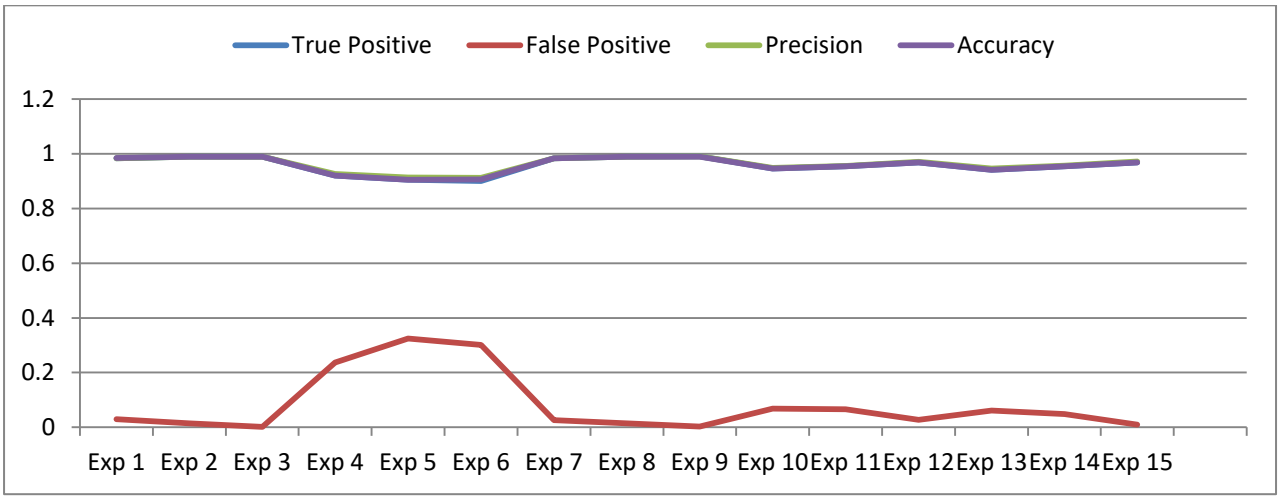


Figure 7 Diagrammatical representations of all experiments

Below are some of the best rules generated

1. **If** Severity of Accident = Material Damage and Class of Victims = Pedestrian and Time of Accident = Morning/Evening **Then Fatal in Accident: Yes.**
2. **If** Severity of Accident = Material Damage and Class of Victims = Pedestrian and Time of Accident = Night and Number of Victims > 2: **Then Fatal in Accident: Yes.**
3. **If** Severity of Accident = Material Damage and Class of Victims = Pedestrian and Time of Accident = Afternoon and Type of Crashes = Vehicle With Pedestrian: **Then Fatal in Accident: No.**
4. **If** Severity of Accident = Slight and Edu/n Level = Primary and Settlement of Road = Upward and Type of Causality Vehicle = Motor Cycle, ISUZU, ISUZU-Autobus, Minibus **Then Fatal in Accident: Yes.**
5. **If** Severity of Accident = Slight and Edu/n Level = Primary and Settlement of Road = Upward and Type of Crashed Vehicle!= Motor Cycle **Then Fatal in Accident: No.**

E. Performance Measurement of Learning Algorithms- In the experiment evaluation part, we have identified that J48 and Rep tree are comparatively similar and better that the remaining three classifiers. So we have selected the first and third experiments for each classifiers and measured performance of their classifiers accuracy as follows.

Table 2 Confusion Matrix of selected experimental results

Algorithms	Actual	Predicted		Recall✓	
J48 Tree		Non-Fatal accidents	Fatal accidents		Accuracy
Exp.1✓	Non-Fatal accidents	1213	11	99.10%	98.45%
	Fatal accidents	14	373	96.40%	
	Precision ✓	98.90%	97.10%		
Exp.3☒	Non-Fatal accidents	1217	7	99.40%	98.76%
	Fatal accidents	13	374	96.60%	
	Precision ✓	98.90%	98.20%		
Rep Tree					
Exp.7☒	Non-Fatal accidents	1211	13	98.90%	98.39%
	Fatal accidents	12	375	96.90%	
	Precision✓	99%	96.40%		
Exp.9☒	Non-Fatal accidents	1203	21	98.30%	98.76%
	Fatal accidents	0	387	100%	
	Precision ✓	100%	95.20%		

F. Model Evaluations- Since the dataset we have was unbalanced, taking accuracy of the model to decide one model as best model is misleading. In such cases, it is advisable to take precision and recall for deciding whether one model is better than the other or not. In our cases, four of the experiments listed above have comparatively similar precision and recall values. But the 1st and 7th experiments were computed by 10 fold cross validation and the rest were computed by 90% split value for training and testing the model. So model with good predictive accuracy can be obtained by experiments performed with 10 fold cross validation tests according to expert judgments. Then we ignored the rest experiments with 90% split tests and accepted experiments with cross validation tests. Experiment 1st (98%) average precision and (97.75%) average recall for two class labels and 7th experiment (97.70%) average precision and (97.90%) average recall were selected to determine the best model with good predictive accuracy for fatal and non-fatal accident occurrences.

Tester: weka.experiment.PairedCorrectedTTester -G 4 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-prec 2 -stddev-prec 2 -col-name-width 0 -row-name-width 25 -mean-width 0 -stddev-width 0 -sig-width 0 -count-width 5 -print-col-names -print-row-names -enum-col-names"
Analysing: F_measure
Datasets: 1
Resultsets: 6
Confidence: 0.05 (two tailed)
Sorted by: -
Date: 9/14/19 10:59 PM

Dataset	(1) bayes.N	(2) baye	(3) tree	(4) tree	(5) tree	(6) tree
W-Z-A-Data(100)	0.96	0.96	0.99 v	0.95	0.99 v	0.88 *
(v/ /*)		(0/1/0)	(1/0/0)	(0/1/0)	(1/0/0)	(0/0/1)

Key:
(1) bayes.NaiveBayes
(2) bayes.BayesNet
(3) trees.J48
(4) trees.RandomForest
(5) trees.REPTree
(6) trees.RandomTree

(V/ /*) The symbol "V" represents victory = represents best algorithm and the symbol "*" represents Astrix = represents the poorest algorithms against the base algorithm.

The above result shows that J48 Tree and Rep tree are significantly best by performance than all other classifiers with the given dataset. Naïve Bayes and Bayesian network classifiers are significantly good by their performance and the rest two algorithms (Random forest and Random tree) classifiers are poor by performance when compared to other classifiers with the given dataset.

V. DISCUSSIONS

Feature selection experiment, the researchers identified determinant factors for the increased vehicle accident occurrences; that are *Accident Woreda, Specific Place, Month of Accident, Year of Experience, Crash Cost in Birr, Type of Crashes, Year of Accident, and Day of Accident* including the four determinant attributes identified from the decision tree rules.

From the generated decision tree rules the researchers observed ten most predictive attributes for vehicle traffic accident occurrences despite the attribute for splitting criteria/root node attribute i.e. Severity of accident and Class level attribute/leaf node attribute i.e. fatal in accident. These attributes are *Settlement of Road, Education level (for derivers), Type of Crashed vehicle, Class of victims, Time of Accident, Number of Victims, Type of Crashes, Age of Deriver, Day of Accident and Type of Causality Vehicle.*

The most common victims of accident are also identified from decision tree while traversing from the root node to the leaf node. They are *pedestrians (40.16%), passengers (19.93%) and derivers (11.73%).* The remaining 28.24% were material damages and animals. Again 53.8% of victims were males and only 19.6% are females, remaining 26.51% accounts for others.

When we see the age of victims, 18.75% of victims were aged between 1-18, 30.54% were aged between 19-30, 18.56% were aged between 31-50 and only 5.96% were aged above 50. The most productive human power was aged between 19-30 and 31-50. *Therefore we identified that the traffic accident most commonly affects the most productive classes of human populations.*

The most determinant cases of accident occurrences are *lack of attention (65.49%), over speed (10.62%), Prohibiting Priority (10.37%), lack of experience (6.33%) and technic failure (3.54%)* When we relate the case of accident to causality condition, the conditions of accidents are *mostly crossing the road (32.96%) straight crash (28.80%), roll down (16.70%), side to side crash (8.57%) and while victims are walking on the road (5.90%).*

The most commonly participated vehicles in the accidents are *Motor Cycle, ISUZU, MEDIUM-BUS, Minibus, Truck, Sino Truck and Tagro Bajaj* among 31 different kinds of vehicles. They account 31.6%, 10.49%, 9.68%, 7.20%, 5.65%, 5.34% and 5.28% respectively.

The researchers identified the black spot areas for the frequent accident occurrences broadly in Woreda levels and Specific place levels. In Woreda level, among 15 different places, 5 Woreda are selected; which are *Sodo City (36.19%), Damot Gale (12.48%), Humbo (9.31%), Sodo Zuria (8.5%) and Boditi City (6.89%)* Specific

places for frequent accident occurrences from these Woreda are *Wadu, Ajif, Arada, Merkato, In front of Marry Church, Kokate, Golla, Otona, Larena, Infront WSU, Buge, Taba, Shasha Gale, Tebela, Shochora, Dalbo, Kawo Shafa, Kingnam, Fate and Gununo.*

From experimental results and performance measurement for learning algorithms, we identified the best machine learning algorithms for vehicle traffic accident prediction. Experiment 1st (98%) average precision and (97.75%) average recall for two class labels and 7th experiment (97.70%) average precision and (97.90%) average recall were selected to determine the best model with good predictive accuracy for fatal and non-fatal accident occurrences. From these two experiments, we calculated F-Measure (harmonic mean of precision and recall) identified that the 1st experiment is comparatively better than the 7th experiment. So we *selected the first experiment (J48 Tree) as the best algorithm and classifier model as the best predictive model to predict factors of vehicle traffic accidents and generate important rules for vehicle traffic accident occurrences.*

Finally from the decision tree experiments, *J48 decision tree is identified as better algorithm that Rep tree and selected the 1st experimental model to generate the best rules*. Because it holds most of attributes that are identified in best feature selection experiment even though the number of leaves and size of tree are more than the rest two experiments.
Accordingly, from first experiment of J48 decision tree, *23 best rules were generated by using IF…Then rules.* These rules show the cases of various fatal and non-fatal accident occurrences. They also hold the most predictive attributes for vehicle traffic accident occurrences.

We also evaluated classifier models select best algorithms from 15 different experiments. We have used f-measure for model evaluation. Finally we identified that J48 and Rep tree are comparatively best algorithms by f-measure than Naïve Bayes and Bayesian Network classifiers and Random Forest tree is poorest by its f-measure than the rest four learning algorithms.

VI. CONCLUSIONS AND FUTURE WORK

In this study, machine Learning approaches have been applied for data analysis and prediction of vehicle traffic accidents. The researcher used seven years accident datasets which have been used to explore important features and pattern relationships of datasets to predict vehicle traffic accident occurrences. Dataset used for this study was unbalanced and it was collected from Wolaita Zone police commission; it was 1611 instances with 36 attributes. The researchers used F-measure for performance measurement of the model. The reality behind is; accuracy is used to measure performance of the model if and only if the dataset used for experiment is balanced. Unless, F-measure is used for performance evaluation of the model. Classification algorithms (decision tree classifiers and Bayesian classifiers) were used to address the problems as the class labels are used for datasets. KDD process modeling was used as a study design.

We addressed various statements of problems and objectives to determine determinant factors of vehicle traffic accidents. From the experimental results, 11 attributes were selected as the most determinant factors for accident occurrences. Seven most commonly participated vehicles were identified, 20 areas for frequent accident occurrences were identified, pedestrians and passengers were identified as the most common victims and J48 and Rep tree classifiers were explored as best algorithms by performance and model accuracy tan the rest. Comparatively, J48 algorithm was selected as the best working model and from this particular model, 23 best rules were generated from the selected model for accident occurrences. The limitation of this study was that the researcher used small amount of datasets, difficulty to obtain suitable datasets and existence of attributes with missing values. Another limitation for the researcher is that only decision tree and Bayesian classifiers were used for prediction.

Therefore the researchers recommend the future researchers try accident predictions with techniques like support vector machine, multilayer perceptron and artificial neural networks. The researchers also recommend future researchers to use use convolutional neural network with python programming language the get better result than the revealed results in this study. It is also recommended for them to add some unconsidered attributes to datasets and relate cases to behavior of derivers like amount of alcohol taken and mental normality of derivers to get better results. Try with deep learning with large amount of instances to get better result and integrate it with knowledge base to know cases for accident occurrences to use is as an expert system.

ACKNOWLEDGMENT

The Authors acknowledge Mrs. Tigist Simon Sundado (Wud-Mimi) for her unforgettable support during these thesis accomplishments from the proposal session to the final defense and the reviewers for their constructive comments. Special thanks deserves for my lovely mother Mrs. Zenebech Daka Daracho (Buluke) whom nursed me from baby to who I am right now.

REFERENCES

[1] Micheale Kihishen Gebru, "Road traffic accident: Human security perspective," International Journal of Peace and Development Studies, vol. 8, no. ISSN 2141–6621, pp. 16, March 2017.
[2] WHO and World Bank, "World Report on Traffic Injury Preventions," New York, 2013.
[3] Guardian, “Traffic Accident Predictions,” The Guardian Publisher, United Kingdom, pp. 23, 2012.
[4] L. Deng and D, Deep Learning: Methods and Applications.: Deep Learning Now Publishing, 2014.

[5] Yoshua Bengio, Learning Deep Architectures.: Foundations and Trends in Machine Learning, 2009.
[6] A. Courville, and P. Vincent., Y. Bengio, Representation Learning: A Review and New Perspectives.: IEEE Trans PAMI, special issue Learning Deep Architectures, 2013.
[7] Schick S. (LMU), "Accident Related Factors," Europe, September 2009.
[8] David Ian White, An Inverstigation of Factors Associated with Traffic Accidents and Causality Risk in Scotland. Scotland: Napier University, October 2002.
[9] Durga Toshniwal2 Sachin Kumar1, A data mining approach to characterize road accident locations.: Published Online: Springerlink.com, 2016.
[10] Armit Kaur Maninder Singh, "A Review on Road accidents in Traffic system Using Data Mining Techniques," International Journal of Science and Research, pp. 6, 2014.
[11] Mrs.Bhumika Gupta Pragya Baluni, "A comparative study of various Algorithms to explore factors for vehicle collision," International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 2012.
[12] Sani Salisu, Atomsa Yakubu, Yusuf Musa Malgwi, Elrufai Tijjani Abdullahi, I. A. Mohammed and Nuhu Abdul'alim Muhammad L. J. Muhammad, "Using Decision Tree Data Mining Algorithm to Predict Causes of Road Traffic Accidents, its Prone Locations and Time along Kano –Wudil Highway," International Journal of Database Theory and Applications, 2017.
[13] Claus Pastor, Manfred Pfeiffer, Jochen Schmidt Heinz Hautzinger, "Analysys for Accident and Injury Risk studies.," Heilbronn University, November 2007.
[14] *, Akinbola Olutayo2 Dipo T. Akomolafe1, "Using Data Mining Technique to Predict Cause of Accident and Accident Prone Locations on Highways," American Journal of Database Theory and Application, pp. 1-13, 2012.
[15] S. Vasavi, "Extracting Hidden Patterns Within Road Accident Data Using Machine Learning Techniques," in Information and Communication Technology Proceedings, Kanuru, AP, India, pp. 11, 2018.
[16] Dejene Ejigu, Pavel Kromer, Vaclav Snasel, Jan Platos and Ajith Abraham Tibebe Beshah, "Mining Traffic Accident Features by Evolutionary Fuzzy Rules," IEEE Symposium on Computational Intelligence in Vehicles and Transportation Systems, 2013.
[17] Micheline Kamber, Jia Pei Jiawei Han, Data Mining Concepts and Techniques, 3rd ed. Canada, USA: Simon Fraser University, 2012.
[18] J,H Kamber, Data Mining concepts and techniques, Second Edition, USA, 2010.
[19] Ajith Abraham and Marcin Paprzycki Miao Chong, "Traffic Accident Analysis Using Machine Learning Paradigms," ResearchGate, pp. 89, December 2004.
[20] Ankit Gupta Malaya Mohanty, "Factors affecting Road crash Modeling," Journal of Transport Literatures, 2015.
[21] Genc Burazeri, Bajram Hysa, Enver Roshi Gentiana Qirjako, "Factors Associated with Fatal Traffic Accidents in Tirana, Albania: Crosssectional Study," 2008.
[22] . Ana Lúcia Andrade1, Rafael Alves Guimarães2, Polyana Maria Pimenta Mandavehicle ú 3,4 and Gabriela Camargo Tobias 4,5 Otaliba Libanio Morais Neto1*, "Regional disparities in road traffic injuries and their determinants in Brazil," International Journal for Equity in Health, pp. 4, 2016.
[23] Hermant Kumar Soni, "Machine Learning AC A new paradigm of AI," International Journal of Scientific Research in Network Security and Communication, Vol.7 , Issue.3 , pp.31-32, Jun-2019.
[24] N. SelvaKumar, M. Rohini, C. Narmada, M. Yogeshprabhu, "Network Traffic Control Using AI," International Journal of Scientific Research in Network Security and Communication, Vol.8 , Issue.2 , pp.13-21, Apr-2020.
[25] Hermant Kumar Soni, "Cervical Cancer prediction based on Hybrid Feature Selection Model and Classification Algorithm," International Journal of Computer Sciences and Engineering, Vol.8 , Issue.6 , pp.101-105, Jun-2020.
[26] Ajeesh Babu, Fathima Basheer, Jayasanker M, Tintu Mariyam Paul, Sithu Ubaid, "Disease Prediction Using Machine Learning Over Big Data," International Journal of Computer Sciences and Engineering, Vol.8 , Issue.7 , pp.11-15, Jul-2020.

AUTHORS PROFILE

Mr. Aklilu Elias Kurika, MSc in IT, is working as a lecturer in the Department of Information Technology at Wolaita Sodo University, Sodo, Wolaita State Ethiopia. He has 4 years of teaching experiences in various Universities and Colleges. He has presented in 1 International Conferences & presented in 2 National Conferences at various Engineering and Informatics Colleges. His areas of specializations and interests are ML, DWDM, AI, NLP and SWE.

Mrs. Tigist Simon Sundado is Working as a lecturer in department of Information Technology, at Wolaita Sodo University, Ethiopia. Her areas of interests are NLP, MLA, and DWDM & IOT.

