

# Knowledge Discovery from Road Traffic Accident Data in Ethiopia: Data Quality, Ensembling and Trend Analysis for Improving Road Safety

Tibebe Beshah, Dejene Ejigu  
IT Doctoral Program, Addis Ababa University  
Addis Ababa, Ethiopia.  
tibebe.beshah@gmail.com; ejigud@yahoo.com

Ajith Abraham, Václav Snášel, Pavel Krömer  
VSB-Technical University of Ostrava  
Department Of Computer Science  
Ostrava, Czech Republic  
ajith.abraham@ieee.org, vaclav.snasel@vsb.cz  
pavel.kromer@vsb.cz

**Abstract** — Descriptive analyses of the magnitude and situation of road safety in general and road accidents in particular is important, but understanding of data quality, factors related with dangerous situations and different interesting patterns in a data is of even greater importance. Under the umbrella of an information architecture research for road safety in developing countries, the objective of this machine learning experimental research is to explore data quality issues, analyze trends and predict the role of road users on possible injury risks. The research employed TreeNet, Classification and Adaptive Regression Trees (CART), Random Forest (RF) and hybrid ensemble approach. To identify relevant patterns and illustrate the performance of the techniques for the road safety domain, road accident data collected from Addis Ababa Traffic Office is exposed to several analyses. Empirical results illustrate that data quality is a major problem that needs architectural guideline and the prototype models could classify accidents with promising accuracy. In addition an ensemble technique proves to be better in terms of predictive accuracy in the domain under study.

**Keywords-** Road Safety, Road Accident, CART, RandomForest, TreeNet, Data Quality

## 1. INTRODUCTION

Road safety, which is mainly affected by road accident is said to be one of the major national health concern. The burden of road accidents causalities and damage is much higher in developing countries than developed nations. Ethiopia is one of the developing countries where road accidents are major problems of Road safety. Road safety improvements can be achieved with in the three components of the road safety system through changes in infrastructure design (which includes road and road signs), vehicle safety, and road user behavior (driver, pedestrian, passengers) [1]. This paper employed different machine learning methods and algorithms in road safety analysis which permits identify patterns and factors to the three components of a road safety system in general and road user behaviors in particular. The work tried to address issues like data quality and trend analysis in addition to identifying interesting patterns. It is also worth mentioning that hybrid architecture approach was used to combine classifiers in order to improve accuracy of the models.

Insight into the effectiveness of injury-reduction technologies, policies, and regulations require a more detailed empirical assessment of the complex interactions that vehicle, roadway, and human factors have on resulting crash-injury severities [2]. Descriptive analysis of the magnitude and situation of road safety in general and road accidents in particular is important, but understanding of data quality, factors related with dangerous situations and different interesting patterns in a data, is of even greater importance. This gives a big picture of the scenario instead of a fragmented effort to

address an aspect of it. Our research is inspired by previous works on the domain and the versatile applicability of machine learning paradigm, which will add on the ongoing effort of improving road safety.

Under the umbrella of information architecture research for road safety improvement in developing countries, the result of a situational analysis made in a three selected regional administration in Ethiopia, exhibited that accident analysis in general is at its immature level, which prohibits the effort of discovering relevant knowledge for decision making, from the accumulated data. This paper reports on a part of a grand research project that aims to better understand data quality issues in general, road users' factors, accident factors, vulnerable groups and vehicles role on accident and injury risk. More specifically the objectives of this specific machine learning experimental research are;

- To explore the magnitude of data quality issues and mitigations.
- To explain and predict the role of road users' related factors on possible injury risks.
- To perform trend analysis on factors affecting accident severity

To the authors' knowledge, this work is unique in the machine learning approaches used, comprehensiveness, time coverage of the analyzed data set used and the actual observation of the road safety related issues. We believe that identifying and describing hidden patterns in accident data in form of innovative classification, visualization and association rules is very understandable for road safety experts to be able make sound decisions.

The remaining part of the paper is organized as follows. In section two, reviews of literatures pertinent to the central theme of the research are presented while the third section is dedicated to explaining details on the research process, approach and data set description. The experiment and the resulting output of the research are presented in the fourth section, which is followed by a conclusion and indications of the future remaining work.

## **2. BACKGROUND AND RELATED WORKS**

In line with the scope of the grand research, attempt has been made to assess the existing accident analysis practice at the three selected regional administrations in Ethiopia. Generally the result revealed that, currently no such analysis is actually being done at the Gambela region (South West part of Ethiopia) while limited descriptive analysis is practiced at Amhara (North West part of Ethiopia) and Addis Ababa (Central Part of Ethiopia) regions. In addition the data quality issues are not yet addressed. However, though they lack systemic approach, there were some fragmented efforts to show the application of data mining techniques on road safety analysis domain. Tibebe, Abraham and Grosan [3] used adaptive regression trees in their rule mining and classification of road traffic accidents, which provides a foundational work on severity analysis in Ethiopian context. The results, according to the authors, showed that the developed models could classify accidents severity within reasonable accuracy.

Zelalem [4] explored classification algorithms for the study of accident severity and driver characteristics. The study focused on predicting the degree of drivers' responsibility for car accidents. The research used WEKA, data mining tool, to build the decision tree (using the ID3 and J48 algorithms) and MLP (the back propagation algorithm) predictive models and identify important relationships between variables that influence driver's degree of responsibility such as; age, license grade, level of education, driving experience, and other environmental factors. Accuracies of the models were 88.24% and 91.84% respectively. In addition, the research reveals that, the decision tree model is found to be more appropriate for the problem type under consideration. With a different approach, [5] explores the application of data mining to identify dangerous locations in Addis Ababa.

In another study, Getnet [6] demonstrates data mining models for accident severity analysis in support of reducing road traffic accidents by identifying and predicting the major vehicles and driver's determinant risk factors (attributes) that cause road traffic accidents. The research uses WEKA, version 3-5-8 tool to build decision tree (using J48 algorithm) and rule induction (using PART algorithm) techniques. The result of the research proves that the performance of J48 algorithm is slightly better than PART algorithm and it identified that LicenseGrade, VehicleServiceyear, Typeofvehicle and Eexperience as most important variables to predict accident severity pattern. Bayesian Network power predictor and constructor was employed by Alemayahu [7] for prediction and model construction purpose respectively in the process of, two experiments which were made before and after the elicitation of the domain experiments. According to the first experiment, type of accident is directly influenced by four factors namely; license grade, time and cause of accident and driver experience with the accuracy of 87.96%. In the second experiment (after elicitation of domain experts) the best accuracy was 80.28% and type of accident is highly influenced by weather condition, road joint and type of vehicles. Tibebe and Shawndra [8] utilize Decision Tree (J48), Naive Bayes and K-Nearest Neighbors algorithms to explain the role of road related factors for severity. The result shows that, all the three classifiers perform well similarly with respect to correctly classified cases. A PART algorithm was also used to generate user understandable rule, with the accuracy of 79.94%. The authors propose further investigation by combining different factors like road and driver related variables.

While the above were specifically targeted attempts in applying machine learning in a road safety domain in a local context, it is also worth mentioning other efforts in employing different methods and tools for better understanding of the domain and accuracy worldwide. Janecka and Hulova [9] conducted an experiment using spatial data mining to discover the hidden rules in the crime data happened in Czech Republic in the year 2008. Oracle data miner along with Apriori algorithm was used for identifying hidden relationship and association rules in the crime data in the form IF A AND B THEN C. The result shows that the situation about the crime perpetrated by youth differs from region to region. Daigavane and Bajaj [10] analyzed road traffic accident data and identified that the causes of accidents stem from different elements namely; the vehicle operator, weather, poor road conditions, age of vehicle, time duration and mechanical failure. The introduction of driver and traffic safety education into the school system was suggested as a major measure to be taken. Highway patrol with a chain of Traffic Aid Centers at intervals of 30-50 Km on highways equipped with ambulance, crane, patrol vehicle and enforcement staff with their equipments to regulate traffic and provide medical assistance to victims of accidents within the first hour of accident was also another recommendation made. Moreover a suggestion in the design of vehicles so that they include inbuilt warning system for minimum distance between two vehicles to avoid collision is also worth mentioning proposal made by the researchers.

Hongguo, Huiyong and Fang [11] explored the applicability of Bayesian Network in traffic accident causality analysis. In undertaking the research the structure and parameter of the Bayesian network was learnt with K2 algorithm and Bayesian parameter estimation respectively. According to the authors, the results show that the Bayesian Network can express the complicated relationship between the traffic accident and their causes, as well as the correlations among the factors of causes. And it is reported that the results of the analysis provided valuable information on how to reveal the traffic accident causality mechanisms and how to take effective measures to improve the traffic safety situations.

Krishnaveni and Hemalatha [12] also conducted a perspective analysis of traffic accident data using data mining techniques. The study deals with some of classification models to predict the severity of injury that occurred during traffic accidents. Naive Bayes Bayesian classifier, AdaBoostM1, Meta classifier, PART Rule classifier, J48 Decision Tree classifier and Random

Forest Tree classifier were compared for classifying the type of injury severity of various traffic accidents. According to the authors, the final result shows that the Random Forest outperforms than other four algorithms. An application of Factor Analysis on Road Traffic Accident was explored by Haixia and Zhihong [13]. The paper analyzes the causes of 372 traffic accidents that occurred in China by factor analysis. According to the authors, five main factors are extracted through the research process and corresponding explanations are given, which can provide not only strategic support for traffic control department, but also some warnings to perpetrators. Li et al. [14] analyzed road accident data to partition highway roads to avoid the occurrence of accidents. They employed fuzzy k-means clustering to classify numerical data of accidents for producing numerical clustering membership, and produce categorical memberships using values of corresponding categorical attributes which was followed by using clustering ensemble to merge all clustering memberships to solve the sole clustering. According to the authors the results showed that cluster ensemble is effective and could be used to avoid occurrence of traffic accidents.

Jinlin et al. [15] proposed a three-layer analysis system based on spatial data mining of GIS. Through the paper, the authors introduced the method of developing traffic accident analysis system by using ArcGIS Engine and C#.NET and give the class realization of system main functions. Saunier, Mourji and Agard [1] investigated collision factors by mining microscopic data (road user's trajectories) about all traffic events with and without a collision. A free and open source tool, TANAGRA, was used to conduct the experiment on video recordings of traffic conflicts and collisions collected at one signalized intersection. Decision trees, the K-means algorithms and hierarchical agglomerative clustering methods were employed to analyze the data. The research revealed that decision tree confirms the importance of the evasive action in interaction outcomes.

Another study by Nayak et al [16] presents a data mining methodology using decision trees for modeling the crash proneness of road segments using available road crash data. The models quantify the concept of crash proneness and demonstrate that road segments with only a few crashes have more in common with non-crash roads than roads with higher crash counts. They also examine ways of dealing with highly unbalanced data sets encountered in the study. Pakgohar et al [17] explored the role of human factors on incidence and severity of road crashes in Iran. The study explains driver's responsibility on an occurrence of an accident. Accordingly the result of the study indicates the important role of human factor such as "Driving License" and "Safety Belt" in severity of accidents in Iran. The study employed descriptive analysis; Logistic Regression, Classification and Regression Tree were employed. Chang and Wang [18] used classification and regression tree (CART), to analyze the 2001 accident data for Taipei, Taiwan. More specifically a CART model was developed to establish the relationship between injury severity and driver/vehicle characteristics, highway/environmental variables and accident variables. It is reported that the most important variable associated with crash severity is the vehicle type. Pedestrians, motorcycle and bicycle riders are identified to have higher risks of being injured than other types of vehicle drivers in traffic accidents.

Computational intelligence methods for information understanding and management were presented by Duch, Jankowski, and Grabczewski [19]. The major software tool used was DataMiner. In addition to that, a large library written in C++, called InfoSel++, implementing different methods for feature selection, have been developed. As reported by the authors, the methods are based on information theory, distance between probability distribution, and statistical approaches. The authors also indicated that, dimensionality reduction based on Multidimensional scaling (MDS) is another unexplored technique. It is

an algorithm basically for data visualization. Besides feature selection, the authors also experimented different algorithms like support vector for clustering the breast cancer data and Principal component analysis (PCA) for visualization. Ona, Mujalli, and Calvo [20] showed the possibility of using Bayesian Networks (BNs) to classify traffic accidents according to their injury severity. Accordingly they presented an analysis of 1536 accidents on rural highways in Spain, where 18 variables representing contributing factors used to build 3 different BNs that classified the severity of accidents into slightly injured and killed or severely injured. Finally the variables that best identify the factors that are associated with a killed or seriously injured accident namely accident type, driver age, lighting and number of injuries, were identified by inference.

Anastasopoulos and Mannering [21], using 5-year data from interstate highways in Indiana, explored fixed and random parameter statistical models. The study used detailed crash specific data and data that include the injury outcome of the crash but not other detailed crash-specific data (only more general data are used such as roadway geometrics, pavement condition and general weather and traffic characteristics). The analysis showed that, while models that do not use detailed crash-specific data do not perform as well as those that do, random parameter models using less detailed data still can provide a reasonable level of accuracy. Another worth mentioning work in the area was conducted by Pei Liu [22]. He studied a self-organizing feature maps and data mining based decision support system for liability authentications of traffic crashes in Taiwan. Through the study the author, develops a decision support tool for liability authentications of two-vehicle crashes based on generated self-organizing feature maps (SOM) and data mining (DM) models. According to the author, although with small data size, the decision support system was considered capable of giving reasonably good liability attributions and references on given cases.

Delen, Sharda, and Bessonov [23] used a series of artificial neural networks to model potentially non-linear relationships between the injury severity levels and crash-related factors. In the process, the authors conducted sensitivity analysis on the trained neural network models to identify the prioritized importance of crash-related factors as they apply to different injury severity levels. According to the authors, the results, mostly validated by the findings of previous studies, provide insight into the changing importance of crash factors with the changing injury severity levels. Savolainen et al. [2] assessed and summarized the evolution of research and current thinking as it relates to the statistical analysis of motor-vehicle injury severities, and provides a discussion of future methodological directions.

Morgan and Mannering [24] used a mixed logit analysis, to assess the effects that age, gender, and other factors have on crash severities by considering single-vehicle crashes that occurred on dry, wet, and snow/ice-covered roadway surfaces. The results showed that there were substantial differences across age/gender groups under different roadway-surface conditions. For example, for all females and older males, the likelihood of severe injuries increased when crashes occurred on wet or snow/ice surfaces but for male drivers under 45 years of age, the probability of severe injuries decreased on wet and snow/ice surfaces relative to dry-surface crashes, as reported by the authors. The authors argue that, this and many other significant differences among age and gender groups suggest that drivers perceive and react to pavement surface conditions in very different ways, and this has important safety implications. Furthermore, the empirical findings of the study highlighted the value of considering subsets of data to unravel the complex relationships within crash-injury severity analysis. Another worth mentioning work is a study by Tibebe et.al [25] on accident severity. The research used CART and RandomForest to analyze the effect of 12 heuristically selected road user related variables on accident severity. The result revealed that pedestrian and victim attributes are more important than drivers'. The authors also recommend more investigation on data

quality issues and road user related factors so as to guide proactive methods in reducing road accident and improving road safety in general.

With respect to data quality, Januzaj [26] presented an application of data mining technologies based on clustering, subspace clustering and classification in identifying data quality problems. The authors claimed that the proposed approach was efficient in data quality problems in a case study of a financial data. The major quality problems identified were wrong entries, zero and empty fields and doublets [26]. In another study, Chen et al [27] studied the data quality of Chinese Materia Medica (Cmm) data warehouse by focusing on the problems of data integrity and accuracy and proposed the method of workflow control. As per the authors, data quality control should be carried out from three aspects such as management, workflow and technology. Farzi and Dastjerdi [28] examined the use of data mining for measuring the quality of data. The authors introduced a method, which uses data mining to extract some knowledge from database, and then use it to measure the quality of input transaction. Accordingly, an algorithm with three steps was proposed, which calculates the data quality of transaction; Extract association rules, which depend on input transaction (T) and are adapted by the functional dependency, Separate compatible and incompatible association rules and finally calculate the quality of input transaction.

Xiong et al. [29] studied noise removal techniques to enhance data analysis in the presence of high noise levels. Accordingly they explored four techniques, three of which are based on traditional outlier detection techniques; distance-based, clustering-based, and an approach based on the Local Outlier Factor (LOF) of an object. The fourth technique was hyperclique-based data cleaner (HCleaner). The techniques were evaluated in terms of their impact on the subsequent data analysis, specifically, clustering and association analysis. Through the experiment, the authors reported that all of these methods provide better clustering performance and higher quality association patterns as the amount of noise being removed increases, although HCleaner generally leads to better clustering performance and higher quality associations than the other three methods for binary data.

To summarize given the magnitude of the road safety problem, researches on accident data analysis are limited at least in a local context. This is true especially in case of researches related with data quality and combining models for better result. Empirical studies considering data quality and understanding are still insufficient. On the other hand there is an understanding that all counter measures should follow from data analysis. In connection to this, again even the data collected is not complete enough to explain all necessary patterns. Thus this implies that more works are expected on how better to collect accident data and analyze it, which is a research problem that the grand research, of which this work is part, tries to address.

### **3. TOOLS, METHODS AND MATERIALS**

This part of the paper describe the data set and methods used in addition to the explanation of the software tool employed to apply different algorithms for data quality exploration, attribute selection, dimensionality reduction, and classification.

#### ***3.1 The Data Set and Tools used***

Though the grand research covers three administrative regions in Ethiopia, this particular experimental study used data obtained from Road Traffic Office at Addis Ababa, Ethiopia. This is mainly because accident data is in long hand written format in Gambela region and it is still in a process of transferring accident data to a computer system in Amhara region. The total dataset for the study contains traffic accident records from 2004/5-2008/9. Based on the availability of the data, for this specific study a total number of 14,254 accident cases described with 48 attributes were used. According to the variable

definitions for dataset, this dataset has information related to road users (drivers, pedestrians and passengers), vehicles and road environment. The tool used to perform machine learning and apply data mining algorithms is Salford Predictive Miner v.6.6 (SPM) a newly developed software suite by Salford Systems, which includes four major predictive model building methods called CART, RandomForest, MARS and TreeNet. The motivation for using this tool includes its features related to faster training time, its ability to use raw data (no need to transform or prepare the data), automatic handling of missing values, automatic handling of categorical (nominal) predictors, handling very large numbers of predictors, and ability to handle very large training data files.

To conform to the industry-standard process, the machine learning methodology used was guided by the CRISP-DM (Cross-Industry Standard Process for Data Mining) process framework. Accordingly based on the situational analysis on the case study, business and data understanding were the first tasks. Then follows exploration of data quality issues, preprocessing and feature/attribute selection tasks relevant to the data mining goal identified. Model building and evaluation along with a possible recommendation to integrate the resulted pattern or knowledge with the existing one was the last stage.

As this is a report of an ongoing research project, attempt has been made to use three of the available predictive modeling methods, CART, TreeNet, and RandomForest, in the SPM suite. The fourth techniques, MARS, is designed in such a way that it works on a binary target class and thus was not feasible for the data mining goal of this specific research. In addition a parallel configuration of combining models with a majority vote approach is used as an ensemble technique. The experiment will continue to uncover other aspect of road safety and using different predictive and clustering techniques so as to get good understanding of the data in identifying patterns. A brief description of the three predictive methods and the model combination techniques is presented in the next subsections.

### **3.2 CART Method**

As explained by Gey and Nédélec [30] Classification and Regression Trees (CART) is a robust decision-tree tool for data mining, pre-processing and predictive modeling tasks. CART can analyze complex data for patterns and relationships and uncovering hidden structures. Moreover, CART is a nonparametric technique that can select variables from a large data set and their interactions that are very important in determining the outcome variable to be analyzed. Some of the major advantages of CART as described by, Salford Systems [31] includes faster training times, its ability to use raw data (no need to transform or prepare the data), automatic handling of missing values, automatic handling of categorical (nominal) predictors, handling very large numbers of predictors, and ability to handle very large training data files.

An important feature of a CART analysis include a set of rules for splitting each node in a tree; deciding when a tree is complete; and assigning each terminal node to a class outcome. CART always base on a questions that have a 'yes' or 'no' answer to split a node into two child nodes; the *yes* answers to the left child node and the *no* answers to the right child node. CART's method is to look at all possible splits for all variables included in the analysis. Next, CART ranks the order of each splitting rule based on a quality-of-split criterion. The common criterion usually used is a measure of how well the splitting rule separates the classes contained in the parent node. Having the best split, CART repeats the search process for each child node, continuously and recursively until further splitting is impossible or stopped. The next step after having the maximal tree grown and derived set of sub-trees, CART determines the best tree by testing for error rates or costs. With sufficient data, the simplest method is to divide the sample into learning and test sub-samples. The learning sample is used to grow an overly large

tree. Then use the test sample to estimate the rate at which cases are misclassified (possibly adjusted by misclassification costs). The misclassification error rate is calculated for the largest tree and also for every sub-tree. The best sub-tree is the one with the lowest or near-lowest cost, which may be a relatively small tree [32].

### **3.3 *TreeNet Method***

Developed by Jerome Friedman, TreeNet is a robust multi-tree technology for predictive modeling and data processing [31]. TreeNet is known for its ability to offer exceptional accuracy, blazing speed, and a high degree of fault tolerance for dirty and incomplete data. More over it can handle both classification and regression problems and has been proven to be remarkably effective in traditional numeric data mining and text mining [31], [33].

Applying TreeNet model indicate improved, or at least competitive, prediction accuracy than CART [34]. TreeNet is an enhancement of the CART model using stochastic gradient boosting [35]. Boosting refers to the endeavors to “boost” the accuracy of any given learning algorithm by fitting a series of models each having a low error rate and then combining into an ensemble that may perform better [34], [36]. TreeNet can be seen as a collection of many smaller trees contributing to a final model. And a final model prediction is constructed by summing up the contributions of each tree. As explained by Salford systems [31] the key features of TreeNet models includes automatic variable subset selection; ability to handle data without pre-processing; resistance to outliers; automatic handling of missing values; robustness to dirty and partially inaccurate data; high speed; and resistance to over-training. It is also worth mentioning that, according to Salford Systems, TreeNet is resistant to overtraining and is over 100 times faster than a neural net.

### **3.4 *RandomForest Method***

As cited by Krishnaveni and Hemalatha [12], Miaou and Harry[37] described random forest consisting of a collection of tree structured classifiers ( $h(x, \_k)$ ,  $k = 1, \dots$ ) where the  $\_k$  are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input  $x$ . The algorithm works as follows:

- Choose T number of trees to grow
- Choose m number of variables used to split each node.  $m \ll M$ , where M is the number of input variables, m is hold constant while growing the forest
- Grow T trees. When growing each tree do
- Construct a bootstrap sample of size n sampled from  $S_n$  with the replacement and grow a tree from this bootstrap sample
- When growing a tree at each node select m variables at random and use them to find the best split
- Grow the tree to a maximal extent and there is no pruning
- To classify point X collect votes from every tree in the forest and then use majority voting to decide on the class label

A Decision Tree Forest (DTF) is an ensemble (collection) of decision trees, which the combination of predictions contributes to the overall prediction for the forest. A decision tree forest grows a number of independent trees in parallel, and those trees do not interact until after all of them have been built. Decision tree forest models often have a degree of accuracy that cannot be obtained using a large, single-tree model [32]. Its ability to handle thousands of input variables without variable deletion along with quick learning process and its effective method for estimating missing data and maintains accuracy are major sited attributes of this algorithm.

### 3.5 Hybrid Architecture to Combine Models-Ensemble

Literatures indicate that, combining classifiers is said to provide better result. This is mainly because patterns misclassified by different classifiers are not necessarily same [38]. In connection to this, there are different strategies and configurations of combining classifiers. Cascading, Parallel and Hierarchical are the major configurations as stated by Ranawana and Palade [39]. Similarly Wanas [40] recognized two major architectures of ensemble; Cascading and Parallel. Cascading is when the output of one is used as an input for the next in order to reach a final refined classification. Parallel architecture, as shown in fig 1, is a way of providing same input to a number of classifiers and combine their output using a given decision logic.

The decision logic could be linear which includes averaging and weighted averaging of the results or non linear which could be voting, probabilistic or rank based methods, as explained by Ranawana and Palade [39]. For this specific experiment a voted approach is selected where different classifiers provide their result for majority vote decision logic to determine the final class. A majority voting technique works very well when all the classifiers are somehow comparable or if there is no any very bad or very good classifier [40]. In case of different results from all classifiers the decision logic will consider the result of the classifier with better overall accuracy.

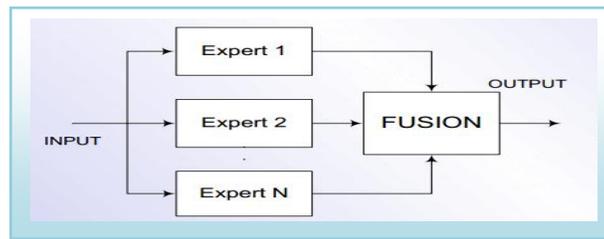


Figure 1: Parallel ensemble topology

## 4. EXPERIMENTS AND RESULT

In this part of the paper a detailed description of data exploration task and results of different experiments are presented. Accordingly, the data which was in a relational database format was first exported in to a single table format of excel sheet. This is mainly because the SPM tool supports a single table data format for processing. In addition, it was also necessary to translate the data from local language, Amharic, in to English for better readability using the filter facility of Ms-Excel application. Moreover, removal of some attributes for ethical reason and their unnecessary nature in the process of pattern identification and attribute creation through aggregation of attribute values of injury severity exposed a total of 38 features for many sided analysis.

### 4.1 Exploration of data quality issues

Data quality is one of the major concerns in organizational decision makings. Especially from machine learning and data mining point of view, where extracting pattern and knowledge discovery is a major task, it is an issue that needs closer attention. Though data quality can be seen from different perspectives, the focus of this paper will lay on data quality issues at the analysis level, affecting knowledge and pattern discovery. It is explained based on the road safety as a case study. Accordingly, Tan et al. [41] identified three major problems related to data quality in machine learning environment; noise and outliers, missing values and duplicate data. Noise and outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set while the major reasons for missing values are inapplicability of an

attribute for all the cases or due to inability to collect a specific value for an attribute because of different reasons. Duplicate data on the other hand may happen due to the lack of effective design of attributes for data objects.

The process of handling these data qualities in general is referred as data cleaning or preprocessing. However data cleaning or preprocessing will depend and be restrictive as per the data mining or knowledge discovery task at hand. Thus, this paper argues that data quality issues should be addressed at a different level right from the collection to the dissemination. This will be reflected on the information architecture to be proposed as a final deliverable of the grand research. However the magnitude of data quality issues at the analysis level, in a road safety data management are explored and presented. It is easy to learn from the details that the three major data quality problems are prevalent in the road accident data set. In connection to this, though there are different noises and outliers in a given data, the “unknown” value is picked as an example to show the magnitude of the problem. Accordingly, variables with their percentage of unknown values are presented in Table 2. And it easy to understand that by improving, the data quality while collecting accident data, through quality checks, it is possible to achieve better prediction and more relevant knowledge.

As to the missing values, variables exhibiting 0.5 % and above missing values are presented in Table 1. And it can be seen that variables related with road users show considerable missing which can affect the amount and quality of pattern to be discovered. And it is visible again that properly addressing these issues will add on the performance and accuracy of data analysis. The duplication issue is exhibited with variables related with accident date. There are three attributes year, month, week, which can be expressed only by proper data structure of date variable itself. In addition to missing values under existing variables, another important attribute missing is use of seatbelt/helmet. Though the use of seatbelt and/or helmet is considered to be one of the important measures in reducing accidents and fatality, it is not included in the accident data.

TABLE I: DATA QUALITY PREVALENCE TABLE (MISSING)

S.N	Variable	% Missing
1	VEHICLETECHSTATUS	1.3%
2	VICTIMAGE	0.74%
3	VICTIMCATEGORY	0.57%
4	VICTIMHEALTHST	0.66%
5	VICTIMOCCUP	0.64%
6	WEATHERCONDITIONS	1%
7	PEDSTRIANMOVEM	81%
8	ROADCONDITION	0.41%
9	VEHICLEPLATE	13%

TABLE II: DATA QUALITY PREVALENCE TABLE(UNKNOWN)

S.n	Variable	% Unknown
1	VICTIMHEALTHST	0.60%
2	VICTIMOCCUP	0.72%
3	DRIVINGLICENS	8.82%
4	PEDSTRIANMOVEM	0.27%
5	VICTIMAGE	0.01%
6	DRIVINGEXP	9.11%
7	VICTIMCATEGORY	0.01%
8	ACCUDRIVEHIRELATION	9.32%
9	DRIVERAGE	9.39%
10	VECHILEMOVEMENT	0.06%
11	DRIVERSEX	9.74%
12	ACCUDRIVEDULEVEL	9.30%

## 4.2 PreProcessing and Model Building

Data preparation or pre processing is always important in a machine learning and pattern recognition process. There are various types of preprocessing tasks like handling missing values, minimizing noises, dimensionality reductions, attribute aggregations, feature creation, descritization and binrization, attribute transformation, sampling and feature selection which

mainly are guided by the data mining goal at hand. In light of the whole objective of the experiment, the preprocessing task for this research can be consider as light weight preprocessing. The main reasons were the tool’s capability of handling data quality issues like missing data and the need to expose the actual data as it is. Preprocessing tasks undertaken for this specific experiment includes; dimensionality reduction by removing records with significant variable values missing and removing of attributes that do not contribute to the analysis like *serial number*, *date*, *year* and *month*. In addition, generalization of *serious injury* and *slight injury* in to *injury class*, and replacement of blank cells by “not applicable” (N/A) value for variables that do have such features when seen from the target variable point of view, are also done. Categorizing some variables like *age* and *hour* in to manageable categories was also done for better understandability of the pattern.

The next task of the experiment was to identify attributes or features related to the goal of the machine learning task which will obviously be evaluated by the machine learning process through attribute selection. The best explanation of the data obviously depends on the type of the problem, intention of users, as well as the type of questions and explanations that are commonly accepted in a given domain [19]. However given the data mining task mentioned above, attempt has been made to include as more attributes as possible. This is mainly to see the role of road users related factors over the others on accident severity risk and which road user related factors are more important in addition to assessing the trend of impacts of factors to severity. Accordingly 31 attributes are selected as possible predictors, *accident collision result* being a target variable. The target variable has three classes namely; fatal, injury and non-injury. Descriptions of the attributes are presented in Table 3.

TABLE III: LIST AND DESCRIPTION OF POSSIBLE PREDICTORS

S.N	Attributes	Description
1	<i>ACCCOLLISIONTYPE</i>	Accident collision type
2	<i>ACCSUBCITY</i>	Sub city where an accident occurs
3	<i>VICTIMAGE</i>	Age of victims
4	<i>VICTIMOCCUP</i>	Occupation of Victims
5	<i>VEHICLETYPE</i>	Type of Vehicle involved
6	<i>VICTIMHEALTHST</i>	Health condition of victims
7	<i>ACCIDENTCOUSE</i>	Immediate cause of an accident
8	<i>VICTIMCATEGORY</i>	Category of victims
9	<i>HOURECATEGORY</i>	Category of accident hour
10	<i>ACCAREA</i>	Specific area of an accident
11	<i>DRIVINGLICENS</i>	Driving license level of a driver
12	<i>DRIVINGEXP</i>	Driving experience of the driver
13	<i>ACCUDRIVEHIRELATION</i>	Relationship b/n a vehicle and a driver
14	<i>ACCDAY</i>	Day of accident
15	<i>LIGHTCONDITION</i>	Light condition while accident occurs
16	<i>VEHICLEPLATE</i>	Vehicle Plate Category
17	<i>ROADSEPARATION</i>	Road Separation

18	<i>DRIVERAGE</i>	Age of a Driver
19	<i>ACCWEEK</i>	Specific week of a month
20	<i>VEHICLESERVYEAR</i>	Service year of the vehicle
21	<i>VECHILEMOVEMENT</i>	How the driver was driving the vehicle
22	<i>VEHICLEOWNERSHIP</i>	Vehicle Ownership
23	<i>ACCUDRIVEDULEVEL</i>	Educational level of a driver
24	<i>ROADJUNCTION</i>	Type of road junction
25	<i>DRIVERSEX</i>	Sex of a driver
26	<i>ROADORIENTATION</i>	Type of road orientation
27	<i>PEDSTRIANMOVEM</i>	Pedestrian movement during the accident
28	<i>VEHICLETECHSTATUS</i>	Technical status of the vehicle
29	<i>ROADCONDITION</i>	The condition of the road
30	<i>WEATHERCONDITIONS</i>	Weather condition
31	<i>ROADSURFACE</i>	Road surface type
32	<i>AccidentResult (target class)</i>	Whether a collision ended with fatal, injury or non-injury

### 4.3 CART Analysis Result

The first experiment in classifying the class attribute *accidentresult* was using CART technique with 31 predictor variables. While running the CART analysis, the classification method used was entropy. Entropy is one of the different splitting rules in growing classification trees. Regarding dataset usage, 80/20 percent of the data is used for training and testing respectively. With the intent of finding the best prediction, a number of experiments have been done by trying different constraints and parameters. To mention some Gini and class probability were tested as a method for classification while 10 fold validations was also used as a testing mechanism. With respect to best tree selection, the CART default best tree setting which is a minimum cost tree is employed.

The best model identified indicated that, victim related features namely *VictimAge*, *VictimCategory*, *VictimOccup*, and *VictimHealSt* followed by *AccidentCollisiontype*, *AccidentCouse*, *AccidentSubcity*, *VehicleType*, *Hourscategory*, *AccidentArea* and *DrivingLicens* are the top ten important predictors of the target class injury result (risk). On the other hand road and environment related factors like *RoadSurface*, *WeatherCondition* and *RoadCondition* are from the least significant factors compared to human related factors. Accordingly given the purposeful low level of preprocessing done, using these variables with major model specification and automatic best predictor discovery, the accuracy of the predictive model is promising with a general classification error of 0.300. Road user factors are found to be determinant whether an accident ends with fatal, injury or non-injury and it can be seen from major splitters as illustrated in Fig. 2 below.



Accordingly with respect to the ROC in this specific experiment, it scored 0.9772 for training and 0.940 for test scenario in case of fatal class, 0.9887 and 0.9721 for training and test sets in case of Injury class and 0.9964 and 0.9962 for training and test sets for non-injury class. ROC charts for all the three classes containing both training and test cases are presented in Fig. 3.

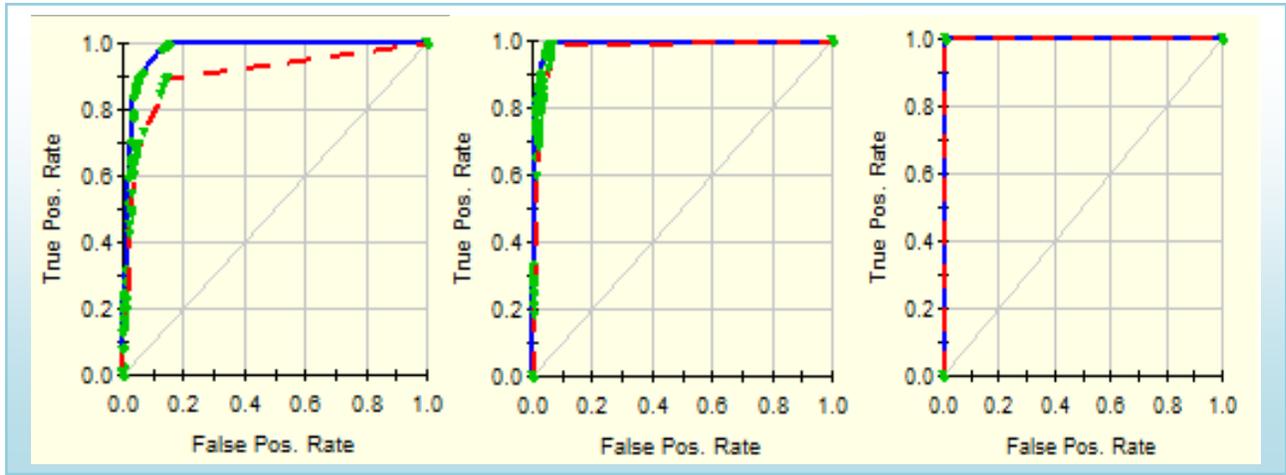


Figure 3: ROC (Fatal, Injury and Non Injury in order)

Another important concept regarding the performance of a predictive model is miss-classification rate. As it can be seen from Tables 6 and 7 below, the model is better in predicting non-injury results than injury and fatal.

TABLE VI: VIMISCLASSIFICATION FOR LEARN DATA (CART)

Class	N Cases	N Mis-Classified	Pct. Error	Cost
Fatal	518	67	12.93	0.13
Injury	2,113	433	20.49	0.20
NoInjury	8,770	0	0.00	0.00

TABLE VII: MISCLASSIFICATION FOR TEST DATA (CART)

Class	N Cases	N Mis-Classified	Pct. Error	Cost
Fatal	168	56	33.33	0.33
Injury	485	129	26.60	0.27
NoInjury	2,200	0	0.00	0.00

#### 4.4 TreeNet Analysis and Result

The second experiment was done using the TreeNet analysis method of Salford Predictive Miners suite. Accordingly 31 predictor variables were used to predict the target class, Accidentresult (risk). The analysis was done by specifying basic parameters like balanced class weights to up weight small classes to equal size of largest target classes, 80/20 percent for training and test sets respectively, and cross entropy or likelihood as a means of selecting optimal logistic model. Out of the total 31 variables, this method identified 29 of them as important predictors by excluding *roadsurface* and *roadcondition* which scored 0.00 importance. It was also interesting to see that *AccidentSubcity*, *VictimAge*, *VehicleType*, *AccidentCollisionType* and *VictimOccupation* were the top five factors for Fatal class while *VictimOccupation*, *VictimeCategory*, *VictimHealthSt*, *AccidentSubcity* and *VehicleType* were for average of all three classes.

On the other hand, *DriverSex*, *WeatherCondition*, *RoadSeparation*, *VehicleOwnership* and *VehicleTechStatus* were found to be least important to determine fatality, while *weatherCondition*, *DriverSex*, *VehicleTechStatus*, *RoadSeparation* and

RoadOrientation were the least important factors for average of all classes. In the process of the experiment the total trees grown were 200 and the optimal number of tree was found to be 59 with classification error of 0.204 and cross entropy of 0.399. The TreeNet result in terms of entropy and classification error is presented in fig. 4 and 5.



Figure 4: Cross Entropy (TreeNet)



Figure 5: Classification Error (TreeNet)

Entropy is a measure of dispersion in a matrix of information. Cross Entropy is a version of entropy that incorporates the modeled nature of the information content. The process of building a good model hence can be seen as initializing a model with a random parameters followed by measuring the cross entropy and then a successive adjustment and measurement of the mode until the cross entropy is low enough. Accordingly the TreeNet model exhibits cross entropy of 0.399 at 59<sup>th</sup> tree, which is referred as optimal number of tree. Similarly a classification error is a percentage of wrongly classified instances from a total number of predictions, in which is TreeNet model showed a minimum result, 0.204, which is closer to 0. As to the prediction Success, TreeNet method exhibits an overall performance of 95.40% for training and 94.15% for testing sets. The detail is presented in Tables 8 and 9.

TABLE VIII: TREENET PREDICTION SUCCESS FOR TEST SET

Actual Class	Total Class	Percent Correct	Fatal N=215	Injury N=438	NoInjury N=2200
Fatal	168	64.29	<b>108</b>	60	0
Injury	485	77.94	107	<b>378</b>	0
NoInjury	2,200	100.00	0	0	<b>2,200</b>
Total:	<b>2,853.00</b>				
Average:		80.74			
Overall % Correct:		94.15			

TABLE IX: TREENET PREDICTION SUCCESS FOR TRAINING SET

Actual Class	Total Class	Percent Correct	Fatal N=863	Injury N=1768	NoInjury N=8770
Fatal	518	82.63	<b>428</b>	90	0
Injury	2,113	79.41	435	<b>1,678</b>	0
NoInjury	8,770	100.00	0	0	<b>8,770</b>
Total:	<b>11,401.00</b>				
Average:		87.35			
Overall % Correct:		95.40			

Misclassification rate was another parameter considered to measure the performance of the model. Accordingly the misclassification rate is presented in Tables 10 and 11 for training and testing sets respectively. As it can be seen from the tables in both learning and testing scenario misclassification in case of non injury class is none.

TABLE X: LEARN MISCLASSIFICATION RESULT OF TREENET

Class	N Cases	N Mis-Classed	Pct. Error	Cost
Fatal	518	90	17.37	90.00
Injury	2,113	435	20.59	435.00
NoInjury	8,770	0	0.00	0.00

TABLE XI: TEST MISCLASSIFICATION RESULT OF TREENET

Class	N Cases	N Mis-Classed	Pct. Error	Cost
Fatal	168	60	35.71	60.00
Injury	485	107	22.06	107.00
NoInjury	2,200	0	0.00	0.00

When it comes to the ROC measure, TreeNet analysis method showed 0.96372 for training and 0.95097 for test scenario in case of fatal class, 0.98905 and 0.97823 for training and test sets in case of Injury class and 0.99374 and 0.99395 for training and test sets for non-injury class. ROC charts for all the three classes for test cases are presented in Fig. 6.

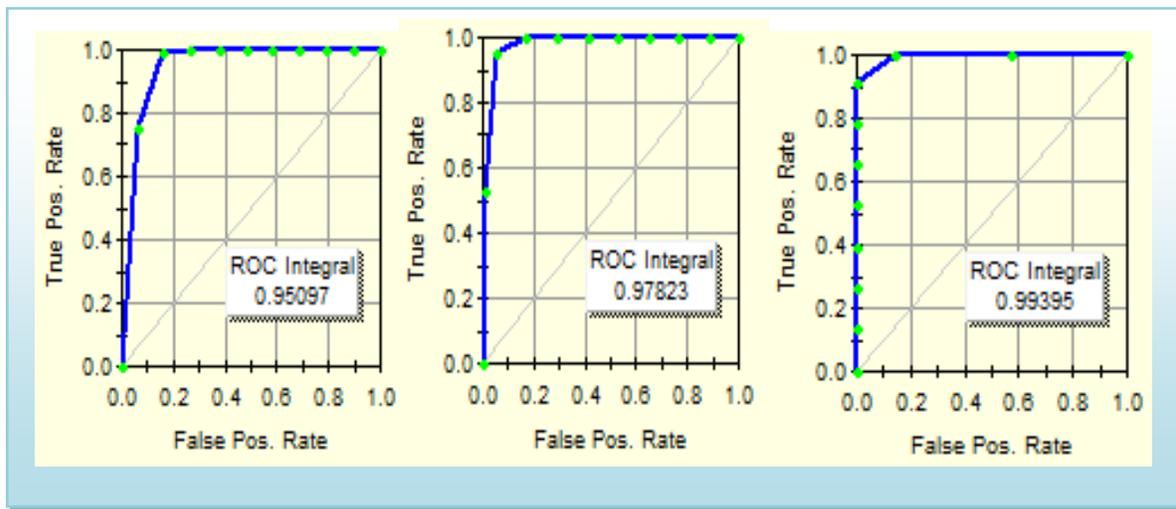


Figure 6: ROC intigral (TreeNet)

#### 4.5 RandomForest Analysis and Result

Similar to that of the other two models with 31 attributes selected, the accident data was exposed to RandomForest analysis. The analysis was done by setting basic parameters like balanced class weight to up weight small classes to equal size of largest target class and testing out of bag data technique for testing the models. Accordingly with 500 trees grown, the method exhibited over all error rate of 0.224, while the error rate for fatal, injury and non- injury are 0.226, 0.446 and 0.000 respectively. The performance of a predictive model error rate lies in between 0 and 1.

The overall error rate is presented in fig 7.

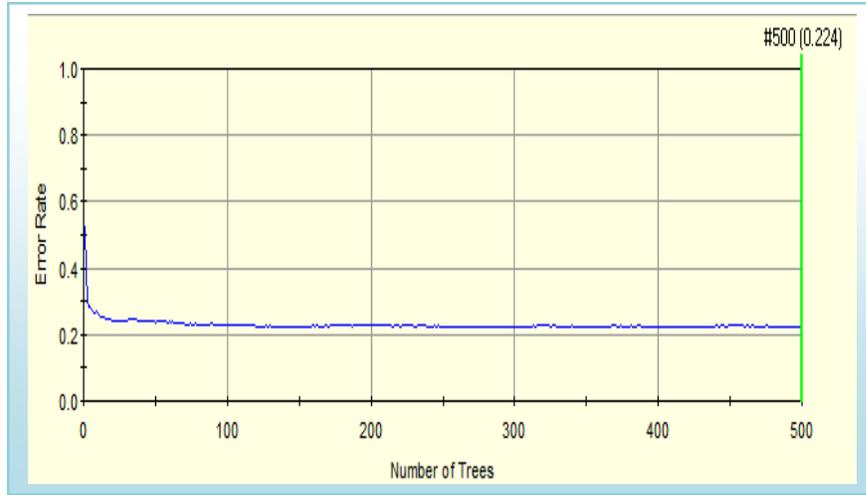


Figure 7: Error rate result (RandomForest, overall)

With respect to variable importance *VictmOccup*, *VictimHealthSt*, *VictimCategory*, *VictimAge* and *Accidentcollisiontype* were the top important factors selected based on their predicting power in descending order respectively. On the other hand factors related with road and environment namely *Roadsurface*, *WeatherCondition*, *RoadOrientation*, *VechileMovement* and *AccidentArea* are found to be least important in determining the risk of fatality.

As to miss-classification, somehow similar to that of the other two methods result, RandomForest analysis is also less accurate in classifying injury category while the miss classification rate is by far less for non-injury category. This is shown with 22.59%, 44.61 % and 0.04% classification error for fatal, injury and non injury classes respectively. The detail is presented in Table 12.

TABLE XII: MISCLASSIFICATION RESULT OF RANDOMFOREST

Class	N Cases	N Mis-Classed	Pct. Error	Cost
Fatal	686	155	22.59	155.00
Injury	2,598	1,159	44.61	1,159.00
NoInjury	10,970	4	0.04	4.00

Prediction success and ROC results are also important indicators of a given predictive model. Accordingly, percentages of correct prediction for fatal, injury and non-injury case are 77.41%, 55.39% and 99.96% respectively. The detail is presented in Table 13. In the same token as shown in fig. 8, the ROC integral indicates, 0.94260, 0.97671, and 0.98941 for fatal, injury and non injury classes respectively. As it is closer to one and indicates minimal zero positives and negatives, it entails good performance.

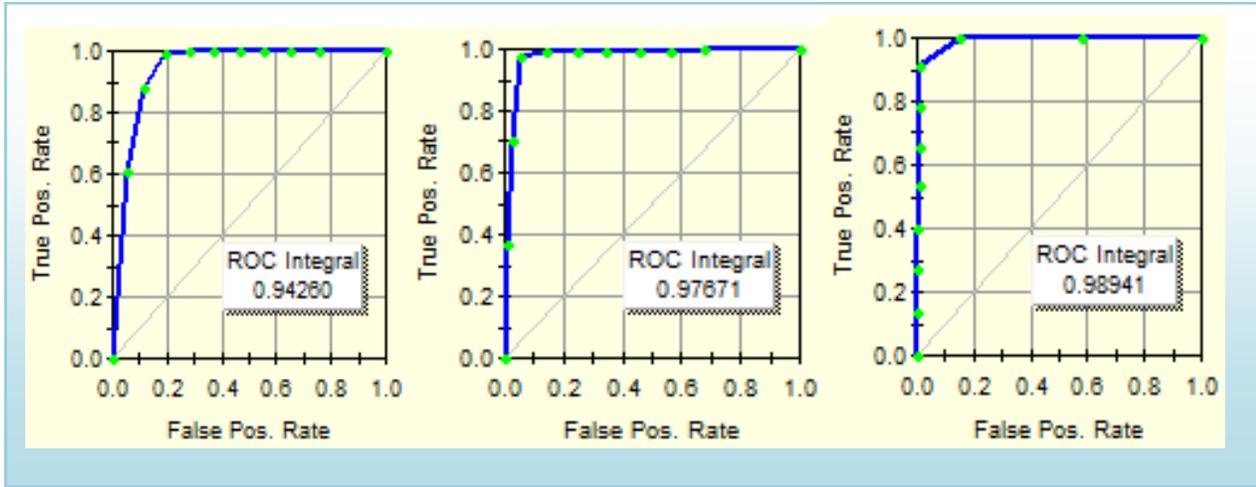


Figure 8: Error rate result (overall)

TABLE XIII: INTERNAL TEST/OUT OF BAG PREDICTION SUCCESS

Actual Class	Total Class	Percent correct	Fatal N=1634	Injury N=1591	NoInjury N=11029
Fatal	686	77.41	<b>531</b>	152	3
Injury	2,598	55.39	1,099	<b>1,439</b>	60
NoInjury	10,970	99.96	4	0	<b>10,966</b>
Average		77.58			
Overall		90.75			

#### 4.6 Scoring and Ensembling

According to Salford Systems, [31] predictive modeling process would not be complete without the ability to apply the model to a data. The data could be a new data or an existing data used for training and testing the models. This process is termed as scoring. It can be done externally by translating the model in to any other supported language or internally by a built in facility of the tool being used like SPM, in this specific case. In line with this, internal scoring is employed on the whole training and testing data to predict the target class of injury risk. And the result shows that the models work well in a new data too and its result is used for the final stage of combining classifiers. This is exhibited with the 95.19%, 94.55% and 95.15% over all prediction success of CART, RandomForest and TreeNet respectively.

The final stage of the experiment is to combine models. There are different configurations and techniques to combine classifiers. As discussed by Ranawana and Palade [39] cascading, parallel and Hierarchical are the major configurations. In this specific experiment a parallel combination of classifiers where the result of each classifier is exposed to a given decision logic, voting techniques, is employed. According to Hall et al [42] voting is an aggregation technique used to combine decisions of multiple classifiers. In its simplest form, which is based on plurality or majority voting, each individual classifier contributes a single vote. The aggregation prediction is decided by the majority of the votes, i.e., the class with the most votes is the final prediction [43].

Accordingly it was possible to exhibit an overall prediction success of 95.47%, while it is 87.61%, 78.41% and 100% for fatal, injury and non injury classes respectively. The detail is presented in Table 14. It is easy to see that, it is by far better than CART, RandomForest and TreeNet predictions independently. It also important to note that, TreeNet is more closer than other techniques to the ensemble or combined classifiers in terms of accuracy especially in case of non injury, injury and overall accuracy. This exhibits that, combining different classifiers outperformed other single classifiers for predicting injury risk.

TABLE XIV: ENSEMBLE PERFORMANCE AGAINST INDIVIDUAL MODELS

Model comparison in percentage of prediction Success				
	CART (test)	RF (test)	TN (test)	Ensemble
Over all	93.52%	90.75%	94.54%	95.47%
Fatal	66.67%	77.41%	64.29%	87.61%
Injury	73.40%	55.39%	77.94%	78.41%
Non Injury	100.00%	99.96%	100.00%	100.00%

#### 4.7 Models Comparison, Discussion and Lessons learned

As mentioned in previous sections, the experiment was done using Salford Predictive Miners suite (SPM). SPM mainly contains four methods. Three of these techniques namely CART, TreeNet and RandomForest, were employed in this specific empirical experimental research. The fourth method which is MARS is designed to handle binary target class and thus cannot be used in this specific experiment, where there are three target classes. With the intent of finding the best model, the predictive performance of these techniques are compared based on three important performance measures namely prediction successes, prediction error rate and ROC.

One of the major objectives of this empirical research was identifying human related determinant factors for accident severity. Accordingly the search and testing methods along with the top 10 important variables identified by the three modeling techniques are presented in table 15 below. It is learned that road user related factors are found to be more important in determining accident fatality or injury. On the other hand factors related with road and environment is found to be least important. This can be seen from the table as they are listed against “least important attributes” under each methods employed. Looking in to the determinant factors, as hypothesized, road users features like their category, occupation and age tell significant information about the possible result of a given accident collision. However this paper argues that the focus of attention in reducing accidents and risks should not be only drivers. As can be seen from the experiment for an accident to end in either fatal, injury or non injury scenario the most determinant factors are the nature of victims involved. On the other hand time and road and environment related factors are found to be irrelevant in determining the result risk of an accident. A good example of this could be the splitter variables and variable importance results of the above experiments. And it is learned that road user related factors need more investigation so as to guide proactive methods in reducing road accident and improving road safety in general.

TABLE XV: ENSEMBLE PERFORMANCE AGAINST INDIVIDUAL MODELS

	CART	TreeNET	Randomforest
<b>Evaluation</b>	Prediction Success, ROC, Error rate	Prediction Success, ROC, Error rate	Prediction Success, ROC, Error rate
<b>Search and testing Method</b>	Entropy with 80/20 percent for training and test sets respectively	cross entropy or likelihood with 80/20 percent for training and test sets respectively	out of bag data technique
<b>Top ten Evaluating attributes</b>	VICTIMOCCUP VICTIMHEALTHST\$ VICTIMAGES\$ VICTIMCATEGORY\$ ACCCOLLISIONTYPE\$ ACCIDENTCOUSE\$ ACCSUBCITY\$ VEHICLETYPE\$ HOURCATEGORY\$ ACCAREAS\$	VICTIMOCCUP\$ VICTIMCATEGORY\$ VICTIMHEALTHST\$ ACCSUBCITY\$ VEHICLETYPE\$ VICTIMAGES\$ ACCCOLLISIONTYPE\$ ACCIDENTCOUSE\$ VEHICLEPLATE\$ HOURCATEGORY\$	VICTIMOCCUP\$ VICTIMHEALTHST\$ VICTIMCATEGORY\$ VICTIMAGES\$ ACCCOLLISIONTYPE\$ VEHICLEPLATE\$ DRIVINGEXP\$ DRIVINGLICENS\$ ACCU DRIVEDULEVEL\$ ACCIDENTCOUSE\$
<b>Least important attributes</b>	PEDSTRIANMOVEM\$ VEHICLETECHSTATUS\$ ROADCONDITION\$ WEATHERCONDITIONS\$ ROADSURFACES\$	VEHICLETECHSTATUS\$ DRIVERSEX\$ WEATHERCONDITIONS\$ ROADSURFACES\$ ROADCONDITION\$	WEATHERCONDITIONS ROADORIENTATION ACCAREA VEHICLEMOVMENT ROADSURFACES\$

With regard to the performance of the models, all the three models perform less in case of fatal and injury classes while their performance is very good in determining non injury risk of an accident. In addition, they all exhibit better ROC scores for non injury class than the others. However TreeNet predictive modeling technique performs better by exhibiting lower error rate of 0.204 which is closer to 0, 94.54% over all prediction success and better ROC score which is closer to 1, than CART and RandomForest. The detail is presented in Table 16 below.

TABLE XVI: MODELS COMPARISON (TEST SET)

Criteria per Target class		Predictive Methods/Techniques		
		CART	RandomForest	TreeNet
<b>Prediction Success (Test Set)</b>	Fatal	66.67%	77.41%	64.29%
	Injury	73.40%	55.39%	77.94%
	Non-Injury	100.00%	99.96%	100.00%
	Overall	93.52%	90.75%	94.54%
<b>ROC (Test set)</b>	Fatal	0.940	0.94260	0.95097
	Injury	0.9721	0.97671	0.97823
	Non-Injury	0.9962	0.98941	0.99395
<b>Error rate</b>	Overall	0.300	0.224	0.204

It is to be recalled from previous sections that another major purpose of this experiment is to get an overall understanding of accident data and getting sense of data quality issues. In connection with this, testing the data for identification of patterns without making significant preprocessing provides a good insight in to the nature of the data. This will guide the subsequent analyses and selection of better tools for this specific domain in a specific context. Accordingly, the role of various aspects of road accidents like vehicle status, time and environment, infrastructure which includes road and road signs, will still be explored to find empirical results that guides counter measures from the data point of view.

The subsequent experiments will result in more patterns. Making more preprocessing will provide better accuracy and explanation about the case at hand. This is especially important in increasing performance of a model like accuracy. This, along with the subsequent experiments, will be used in the design of education and enforcement measures in road safety domain.

Moreover though all the three techniques are found to be promising in identification of patterns in a road safety domain TreeNet is shown to be the best method to be used in the domain under study if the decision is to use a single method. However, ensemble result proves to be the best of all individual models.

To summarize, the following are issues that needs attention both at the organization and/or system level;

- Road accident data should be complete for the analysis to reflect important patterns and knowledge
- There should be a mechanism for data quality checks about each accident that requires architectural guideline.
- Once data collection is organized machine learning approaches to data analysis should be implemented
- Periodic analysis of the accident data is required to see changes through time and adjust the counter measures accordingly

These lessons will be reflected on the information architecture to be proposed as a guideline for accident data collection and analysis. Thus this research is trying to view accident data collection and analysis as a system that requires a special view towards understanding the whole and making sense out of it for improved decision makings in the effort of reducing the problem of road safety ultimately. That is why the issue of data quality and understanding gets more attention in addition to predictive modeling of some interesting patterns not addressed so far.

#### ***4.8 Trend Analysis and implications***

Another aspect of this experiment was to analysis the trend in the past few years. The best point of reference chosen was a study conducted by Tibebe, Abraham and Grosan [2] in 2005 and published on an International Journal of Simulation for its comprehensiveness and comparability from other studies focusing only on road related, driver or vehicle factors. Accordingly in a work by Tibebe, Abraham and Grosan [3], 16 attributes were used in predicting accident severity while the current study used 31 which are, by far more attributes. The best accuracy exhibited was 87% while in the current experiment overall accuracy of 90.57%, 93.52% and 94.54% were achieved in RandomForest, CART and TreeNet experiments respectively. It is also important to note that, the ensemble technique exhibited an overall accuracy of 95.47%.

Another worth mentioning issue was the evaluation method used. Whilst three evaluation techniques namely; Prediction success, ROC and classification error rates, are used in the current experiment, it was only accuracy used in the previous study. And important variables determining severity in the previous study *were accident type* and *accident cause*. On the other hand in the current study, with the inclusion of more attributes, road user related factors are found to be more important in

determining fatality and injury. Thus, it is apparent that periodical analysis of accident data will help to see the trends and reveals more pattern and knowledge about the domain.

## **5. CONCLUSION AND FUTURE WORK**

Through this paper, attempt has been made to explore CART, TreeNet , RandomForest and ensemble techniques for road accident data understanding and analysis. A review of literature enabled to create a good understanding of state of the art techniques and attempts in a road safety data quality and data analysis domain. The main goal was to empirically explore data quality issues, trend analysis and to identify the role of road user's factors, which is said to be the major factor, on the risk of injury for a road traffic accident.

Detection of accidents risks due to road users related factors could assist in designing appropriate counter measures in the effort of reducing the socio-economic impact of road accidents which ultimately improving road safety. Another advantage of this systemic view approach to road traffic accident data understanding and analysis through machine learning is that, hypothesis can be easily be formulated for future trends.

In addition to revealing patterns related with road users factors for accident severity, major contribution of this work includes comparison of predictive models for the domain, highlighting data quality issues, proposing ensemble technique to improve accuracy and trend analysis regarding factors for accident severity. With reference to the main objective, future work will focus on exploring and proposing possible solutions as a means of data quality problem mitigations. Moreover use and comparison of different soft computing techniques on the test bed will revile best approach and accuracy in understanding and predicting road safety patterns. In line with this, novel techniques and algorithms like non-negative matrix factorization and genetic algorithm will also be explored. We strongly believe that the result of theses successive experiments will be major ingredient of the information architecture to be proposed for accident data collection and analysis in developing countries in general and for Ethiopia in particular.

The result of this research will help road safety organizations to revisit their focus of attention in crafting and implementing measures to reduce road safety danger. More specifically the research indicated that in addition to drivers, education and enforcement measures should address well other road users like pedestrians. It is also worth mentioning that systematic data collection and quality check along with periodic analysis should get due attention, so that other measures will be knowledge driven. The research result can also be used as a hypothesis and/or replicated to other developing countries with similar context in the area of road accident data collection and analysis.

Finally the result of this study can also be used to support future research related to machine learning approach, especially in the context of road safety.

## **ACKNOWLEDGMENT**

The authors are grateful to the national and regional road traffic and road safety departments of Ethiopia for letting the researchers get access to the road traffic data, VSB-Technical University of Ostrava, Faculty of Electrical Engineering and Computer Science, Czech Republic for providing computational resource and work environment throughout the research, Addis Ababa University, IT Doctoral Program, for providing fund for research visit.

## REFERENCES

- [1] N. Saunier, N. Mourji, and B. Agard, "Investigating collision factors by mining microscopic data of vehicle conflicts and collisions," 2010.
- [2] P. T. Savolainen, F. L. Mannering, D. Lord, and M. A. Quddus, "The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives," *Accident; Analysis and Prevention*, vol. 43, p. 1666–1676, 2011.
- [3] T. B. Tesema, A. Abraham, and G. Crina, "Rule mining and classification of road traffic accidents using adaptive regression trees," *International Journal Of Simulation*, vol. 6, no. 10 and 11, ISSN- 80 1473-8031 print. 2005.
- [4] Z. Regassa, "Determining the degree of driver's responsibility for car accident: the case of addis ababa traffic office," Master's thesis, Addis Ababa University, 2009.
- [5] H. Kifle, "Application of data mining technology to support the prioritization of dangerous crash location: the case of addis ababa traffic office," Master's thesis, Addis Ababa University, 2009.
- [6] G. Mossie, "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," Master's thesis, Addis Ababa University, 2009.
- [7] A. Tabor, "Bayesian approach for analysis of road traffic accidents: The case of addis ababa." Master's thesis, Addis Ababa University, 2009.
- [8] T. Beshah and S. Hill, "Mining road accidents data to improve safety: the role of road related factors on accident severity." in *Proceeding of AAAI Symposium on Artificial Intelligence for Development*. Stanford University, 2010.
- [9] K. Janecka and H. Hulova, "Using spatial data mining to discover the hidden rules in the crime data," in *GIS Ostrava*, Jan. 2011.
- [10] P. Daigavane and P. Bajaj, "Analysis of selective parameters contributing to road accidents on highways for establishing suggestive precautionary strategies," in *Second International Conference on Emerging Trends in Engineering and Technology, ICETET-09*, I. C. Society, Ed., 2009.
- [11] X. Hongguo, Z. Huiyong, and Z. Fang, "Bayesian network-based road traffic accident causality analysis," in *2010 WASE International Conference on Information Engineering*, vol. 3, Aug. 2010, pp. 413 – 417.
- [12] S. Krishnaveni and M. Hemalatha, "A perspective analysis of traffic accident using data mining techniques," *International Journal of Computer Applications*, vol. 23, no. 7, Jun. 2011.
- [13] Y. Haixia and N. Zhihong, "An application of factor analysis on road traffic accident," in *The 5th International Conference on Computer Science & Education Hefei, China.*, Aug. 2010.
- [14] T. Li, Y. Chen, S. Qin, and N. Li, "Highway road accident analysis based on clustering ensemble," *CSEEE 2011, Part II, CCIS 159*, p. 212–217, 2011.
- [15] W. Jinlin, C. Xi, Z. Kefa, W. Wei, and Z. Dan, "Application of spatial data mining in accident analysis system," in *2008 International Workshop on Education Technology and Training & 2008 International Workshop on Geoscience and Remote Sensing*, I. C. Society, Ed., 2008.

- [16] R. Nayak, D. Emerson, J. Weligamage, and N. Piyatrapoomi, "Road crash proneness prediction using data mining," *EDBT 2011*, Mar. 2011.
- [17] A. Pakgohar, R. S. Tabrizi, M. Khalili, and A. Esmaeili, "The role of human factor in incidence and severity of road crashes based on the cart and lr regression: a data mining approach," in *Procedia Computer Science* 3, 2011, p. 764–769.
- [18] L.-Y. Chang and H.-W. Wang, "Analysis of traffic injury severity: An application of non-parametric classification tree techniques," *Accident; Analysis and Prevention*, p. 1019–1027, 2006.
- [19] W. Duch, N. Jankowski, and K. Grabczewski, "Computational intelligence methods for information understanding and information management." Available online; <http://www.fizyka.umk.pl/publications/kmk/05-CI-info.pdf>. 2006.
- [20] J. de Ona, R. O. Mujalli, and F. J. Calvo, "Analysis of traffic accident injury severity on spanish rural highways using bayesian networks," *Accident; Analysis and Prevention*, p. 402–411, 2011.
- [21] P. C. Anastasopoulos and F. L. Mannering, "An empirical assessment of fixed and random parameter logit models using crash- and non-crash-specific injury data," *Accident; Analysis and Prevention*, vol. 43, p. 1140–1147, 2011.
- [22] P. Liu, "A self-organizing feature maps and data mining based decision support system for liability authentications of traffic crashes," *Neurocomputing*, vol. 72, 2009.
- [23] D. Delen, R. Sharda, and M. Bessonov, "Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks," *Accident; Analysis and Prevention*, p. 434–444, 2006.
- [24] A. Morgan and F. L. Mannering, "The effects of road-surface conditions, age, and gender on driver-injury severities," *Accident; Analysis and Prevention*, vol. 43, p. 1852–1863, 2011.
- [25] T. Beshah, A. Abraham, D. Ejigu, V. Snasel, and P. Kromer, "Pattern recognition and knowledge discovery from road traffic accident data in ethiopia: Implications for improving road safety," in *WICT 2011 Conference Proceedings*. Mumbai, India. ISBN: 978-1-4673-0125-1. IEEE, Dec. 2011,
- [26] E. Januzaj and V. Januzaj, "An application of data mining to identify data quality problems," in *Third International Conference on Advanced Engineering Computing and Applications in Sciences*. IEEE computer Society, 2009.
- [27] B. Chen, B. Wang, X. Weng, and X. Hu, "Analysis and solution of data quality in data warehouse of chinese materia medica," in *Proceedings of 2009 4th International Conference on Computer Science & Education*. IEEE, 2009.
- [28] S. Farzi and A. B. Dastjerdi, "Data quality measurement using data mining," *International Journal of Computer Theory and Engineering*, vol. 2, no. 1, pp. 1793–8201, February 2010.
- [29] H. Xiong, G. Pandey, M. Steinbach, and V. Kumar, "Enhancing data analysis with noise removal," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 3, 2006.
- [30] G. S. and N. E., "Model selection for cart regression trees," *IEEE Trans. Inf. Theory*, vol. 51, p. 658–670, 2005.
- [31] S. Systems, *Salford Systems Data Mining Solutions*, Salford Systems, 2011.
- [32] S. Sulaiman, S. M. Shamsuddin, A. Abraham, and S. Sulaiman, "Intelligent web caching using machine learning methods," *Neural Network World*, vol. 21, no. 5, pp. 429–452, 2011.
- [33] J. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. vol. 38, pp. 367–378, 2002.

- [34] M. Elish and K. Elish, "Application of treenet in predicting object-oriented software maintainability: A comparative study," in *Software Maintenance and Reengineering CSMR '09*. CSMR '09. 13th European Conference, 2009, pp. pp. 69–78.
- [35] J. Friedman and J. Meulman, "Multiple additive regression trees with application in epidemiology," *Statistics in Medicine*, vol. 22, pp. 1365–1381, 2003.
- [36] R. Schapire, "A brief introduction to boosting," in *Proc. 16th International Joint Conference on Artificial Intelligence*, 1999, pp. 1401–1405.
- [37] S. Miaou and L. Harry, "Modeling vehicle accidents and highway geometric design relationships," *Accident Analysis & Prevention* Volume 25, Issue 6, December 1993, Pages 689–709
- [38] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," in *IEEE Transactions on Pattern Analysis and machine Intelligence*, vol. 20(3), 1998, pp. 226–239.
- [39] R. Ranawana and V. Palade. "Multi-classifier systems - review and a roadmap for developers". *International Journal of Hybrid Intelligent Systems* Vol. 3 Issue 1, January 2006.
- [40] N. M. Wanas, "Multiple classifiers systems," U. W. Multiple Classifiers Focus Group, Ed., Dec. 2003.
- [41] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Addison-Wesley, 2006.
- [42] L. Hall, K. Bowyer, W. Kegelmeyer, T. Moore, and C. Chao, "Distributed learning on very large data sets," in *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, SIGKDD, Ed., 2000.
- [43] I. H. M. Paris, L. S. Affendey, and N. Mustapha, "Improving academic performance prediction using voting technique in data mining," *World Academy of Science, Engineering and Technology*, no. 62, 2010.