A Review on Injury Severity in Traffic System using Various Data Mining Techniques

Dheeraj Khera
M.tech (CE) Research Scholar,
Punjabi University,
Patiala, India

Williamjeet Singh
Asst. Professor, Dept of
Computer Engineering, Punjabi
University, Patiala, India

ABSTRACT
Road Traffic Accidents (RTAs) are a major public health concern, resulting in an estimated 1.2 million deaths and 50 million injuries worldwide each year. In the developing world, RTAs are among the leading cause of death and injury. The objective of this study is to evaluate a set of variables that contribute to the degree of injury severity in traffic crashes. The issue of traffic safety has raised great concerns across the globe and it has become one of the key issues challenging the sustainable development of modern traffic and transportation. The study on road traffic accident causes can identify the key factors rapidly, efficiently and provide instructional methods to the traffic accidents prevention and road traffic accidents reduction, which could greatly reduce personal casualty and property loss caused by road traffic accidents. Using the method of traffic data analysis, can improve the road traffic safety management level effectively**.**

Keywords
Road Traffic Accidents, Data Mining, Influential Factors, Weka, Tanagra, R, Data Mining Techniques.

1. INTRODUCTION
In recently years, Because of too much travel speed of road traffic, the accidents have been increasing on yearly basis. So the traffic safety has raised great concerns across the globe. So, it has become one of the key for challenging the modern traffic and transportation so that traffic administrations can be more accurately informed and better policies can be introduced. The encyclopedia defines road traffic accident as "any vehicle accident occurring on a public highway (i.e. originating on, terminating on or involving a vehicle partially on the highway). These accidents therefore include collisions between vehicles and animals, vehicles and pedestrians or vehicles and fixed obstacles. Single vehicle accidents, in which one vehicle alone and no other road user involve, are included. A report by [17] estimated that the costs of fatalities and injuries due to Road Traffic Accidents (RTAs) have a tremendous impact on societal well-being and socio-economic development. RTAs are among the leading causes of death and injury worldwide, causing an estimated 1.2 million deaths and 50 million injuries each year. Moreover, by the year 2020 road accidents will be the third leading cause of death. This puts road safety well ahead of wars, HIV/AIDS, malaria and (other) acts of violence as world health problem. Among children aged 5-14years, and young people aged 15-29 years, road traffic injuries are the second-leading cause of death worldwide. In low-income countries, the majority of road deaths are among pedestrians, passengers, cyclists, users of motorized two wheelers and occupants of buses and minibuses. Globally, the risk of dying in a road crash is far higher for vulnerable road users like pedestrians, cyclists and motorcyclists than the car occupants. This paper structures in four sections; Section 1 covers the basic introduction whereas section 2 presents the Review of literature. Section 3 gives Methods and various data mining techniques are used. The conclusion is drawn in the section 4.

2. LITERATURE SURVEY
The costs of fatalities and injuries due to traffic accidents have a great impact on society. In recent years, researchers have paid a great attention at determining the factors that significantly affect driver injury severity in traffic accidents. The author in [1] identifies most important factors which affect injury severity by using classification & regression tree. The author in [4] presents a random forest & rough set theory to identify the factors significantly influencing single vehicle crash severity. The author in [5] presents a decision tree which predicts causes of accidents and accident prone locations. Papers [10] predict Traffic accident duration of incident and driver information system. The author in [13] used various data mining techniques and tells Random forest Outperforms than other classification algorithms. In paper [14], author talks about the Significance of data mining classification algorithms in predicting the factors which influence road traffic accident. The author in [15] used to explore the possible application of data mining technology for developing a classification model and the Results shows that developed model could classify accidents within a reasonable accuracy. It is important to analyze these datasets to extract useful knowledge. Data mining is an effective tool for analyzing data to extract useful knowledge. Table1 shows a sample of different data mining techniques and their influential factors used in traffic injury severity. The severity of injuries measured for crash records has both continuous and categorical characteristics. Hence many previous studies have used models with ordered structure to analyze risk factors and their effect on severity of injuries sustained in traffic crashes.

Table 1. Summary of the Pertinent Literature

Author	Objective	Data Mining Techniques	Influential factors
Ali Tavakoli Kashani, Afshin Shariat, Andishe Ranjbari (2010)	To identify Most important factors which affect injury severity	Classification & Regression tree	Injury Severity, Gender, Age, Seat Belt, Cause Of crash, Collision type, Vehicle Type, Location type, Lighting conditions, Weather

			conditions, Road surface condition, Occurrence, Shoulder type, Shoulder Width
Chaozhong Wu, Ming Ma, Hu Lei, Xinping Yan (2009)	To identify the factors significantly influencing single vehicle crash severity.	Random Forest Rough set theory	Weather, Speed limit, Lightingconditions, collision factors, gender, Age, Experience, Safety belt, Vehicle type, Severity of svc
DipoT.Akomolafe, Akinbola Olutayo (2012)	To predict causes of accidents and accident prone locations	Decision tree: Id3, Functional tree	Vehicle Type, Time of the day, Season, Causes
Liping Guan, Weiming Liu, Xiangyuan Yin, Luping Zhang (2010)	To predict Traffic accident duration of incident and driver information system	Artificial neural Networks	No. of trucks involved, Rollover vehicle, Facility damage, Degree of traffic jam, No. of fatalities, No. of severe personal injuries, Road pollution involvement, Hazard material, Fire Involvement, Police involvement, Patrol vehicle involvement, Fire engine involvement
S.Krishnaveni, Dr. M.Hemalatha (2011)	To predict severity of injury using data mining techniques & compare algorithm performance.	Naïve Bayes AdaBoost M1 Meta Classifier Part J48, Random Forest	Casualty, Fatal accident, Slight accident, Killed causality, Serious injury, Slight injury, Road users, Vehicles involved
S.Shanthi, R.Geetha Ramani (2012)	Significance of data mining classification	Classfication techniques: C4.5,ID3, CS-CRT,CR-T,CS MC4,	Key Value, state, County, Month, date,
	algorithms in predicting the factors which influence road traffic accident.	Naive Bayes, Random forest	Time, Day, Harmful event, Manner of collision, Person type, Seating position, age, Gender, Injury Severity, Air bag, Protection System, Ejection, Ejection path, Year_ of_ death, Month_ of_ death, Alcohol Test, Drug test, Drug involvement, Accident Location, Related Factors
Tibebe Beshah tesema, Ajith Abraham, Crina Grosan (2013)	To Explore the possible application of data mining technology for developing a classification model	Classification & Regression tree	Accident_Id, Driver_Age, Driv_Exp, Vehic_Age, Vehic_Type, Road_Surf_Type, Road_Cond, Weather_Con, Light_Cond, Acci_Type, Acci_Cause, Injury_Severity

Different data mining techniques have been used to help traffic injury severity such as Decision tree, Naive bayes, artificial neural networks, Classification & Regression tree, J48, Part classifier, Random Forest. Those most frequently used focus on: Classification, Decision tree & artificial neural networks. Driver related factors that affect severe injury crashes have been recognized to have a great influence in the occurrence of crashes. There are various levels of injury severity like No injury means not bodily harm from the crash but only property damage, possible injury means no visible signs of injury but complaint of pain, fatal injury means an injury sustained in a crash that results in a death within 30 days of the crash. Each road traffic accidents records contain multiple data attributes. Each attribute value reflects a characteristic in a traffic accident. And the more data attribute data constitute the multiple dimensions of traffic accidents. In addition, data attributes are set from the basic accident information, personnel information, vehicle information, road information and environmental information.

Table 2. Summary of different data mining tools on different datasets

Author	Data Mining Tool	Work Done
Ali Tavakoli Kashani,Afshin Shariat,Andishe Ranjbari (2010)	Variable importance measure	Cause factors: seat belt, improper overtaking & speed.
Chaozhong Wu, Ming Ma, Hu Lei, Xinping Yan(2009)	Cross Validation Method	Cause factors: Lighting Conditions, vehicle type, driving experience, wearing belt or not. The efficiency of attribute reduction is not high.
DipoT.Akomolafe, Akinbola Olutayo (2012)	WEKA	Decision tree predict causes of accidents and accident prone locations accurately.
Liping Guan, Weiming Liu, Xiangyuan Yin, Luping Zhang(2010)	MATLAB	The incident duration is predicted in practice, as time goes & incident information gradually increases.
S.Krishnaveni,Dr. M.Hemalatha(2011)	WEKA	Random Forest Outperforms than classification algorithms.
S.Shanthi, R.Geetha Ramani (2012)	TANAGRA	Random tree classifier using Arc-x4 Meta Classifier outperforms & also improves accuracy.
Tibebe Beshah tesema, Ajith Abraham, Crina Grosan (2013)	WEKA	Results shows that developed model could classify accidents with in a reasonable accuracy.

Table 2 illustrates a sample of different data mining tools used in the traffic injury severity over different datasets. Weka is a Waikato Environment for Knowledge Analysis. Weka is a collection of machine learning algorithms for data mining tasks and well suited for developing new machine learning schemes. Weka is java based software capable of working under various operating systems. These algorithms can either be applied directly to a dataset or can be called from your own java code. Weka is probably the most successful open source data mining software which has inspired by the development of other programs with more sophisticated graphical user interface and better visualization methods [2][8]. In Weka datasets should be formatted to the ARFF format. The Weka Explorer will use these automatically if it does not recognize a given file as an ARFF file the Preprocess panel has facilities for importing data from a database, a CSV file, etc., and for preprocessing this data using a filtering algorithm. These filters can be used to transform the data and make it possible to delete instances and attributes according to specific criteria. Tanagra is free data mining software for academic and research purposes. It offers several data mining methods like exploratory data analysis, statistical learning and machine learning. The first purpose of the Tanagra project is to give researchers and students easy to use data mining software. The second purpose of Tanagra is to propose to researchers an architecture allowing them to easily add their own data mining methods, to compare their performances. The third and last purpose is that novice developers should take advantage of the free access to source code, to look how this sort of software was built, the problems to avoid, the main steps of the project, and which tools and code libraries to use for. In this way, Tanagra can be considered as a pedagogical tool for learning programming techniques [16]. Revolution is a free software programming language and software environment for statistical computing and graphics. R provides a wide variety of graphical and statistical techniques such as linear and non-linear modeling, classical statistical tests, time- series analysis, classification clustering and is highly extensible. Extensibility and superb data visualization are the two main reasons for the success of R [12][9].

Table 3. Summary of injury Severity accuracy using crash involvements

Author	Sample Size	Accuracy
Ali Tavakoli Kashani,Afshin Shariat,Andishe Ranjbari (2010)	169648	72.49%
Chaozhong Wu, Ming Ma, Hu Lei, Xinping Yan(2009)	59	0.73% 0.54%
DipoT.Akomolafe, Akinbola Olutayo (2012)	148	77.70% 70.27%
Liping Guan, Weiming Liu, Xiangyuan Yin, Luping Zhang(2010)	170	85.35%
S.Krishnaveni,Dr. M.Hemalatha(2011)	34575	84.66% 84.64% 84.64% 85.18% 88.25%
S.Shanthi, R.Geetha Ramani (2012)	457549	99.73%

Tibebe Beshah tesema, Ajith Abraham, Crina Grosan (2013))	5207	87.47%

A summary of past studies where crash involvements were used in the injury severity analysis is presented in Table 3 above. The amount of past research demonstrates the capabilities of different methods to model crash injury severity versus a set of continuous and discrete independent variables. Interactions were found significant in several studies such as: light-weather, alcohol-seat belt, among others. Although there is a substantial amount of literature demonstrating different uses of severity analysis, only a few studies dealt with a sample as large as the one undertaken in this research. The crash data [1] from the records of the Information and Technology Department of the Iranian Traffic Police from 2006 to 2008 was used to study hundreds of drivers who were involved in traffic crashes on the main two-lane two-way rural roads of Iran. The results indicated that seat belt is the most important factor associated with injury severity of traffic crashes and not using it significantly increases the probability of being injured or killed. The crash data [13] from the records of the Transport Department of Government of Hon Kong from 2008 was used. The total record set used was 14,576 investigates the performance of Naive Bayes, J48, AdaBoostM1, PART and Random Forest classifiers for predicting classification accuracy The attributes involved in this case are Severity, District Council, Hit and Run, Weather, Rain, Natural Light, Junction Control, Road Classification, Vehicle Movements, Types of Collision, Number of Vehicles Involved and Number of Causalities Injured. The results indicated that Random Forest outperforms than other classification algorithms instead of selecting all the attributes for classification.

3. METHODS

3.1 Data Mining

This fast growth and tremendous amount of data, collected and stored in large and numerous databases need a powerful tool to elicit useful information. The tool helps to get benefit from the collected data by identifying relevant and useful information. Data mining is one of the solutions to analyze huge amount of data and turn such data into useful information and knowledge. Data mining [6] refers to extracting or "mining" knowledge from large amounts of data. There are some other terms which carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data or pattern analysis, and data archaeology. Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, classified data mining tasks into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database where as predictive mining tasks perform inference on the current data in order to make predictions. Data mining, also popularly known as Knowledge Discovery in Database refers to extracting or "mining" knowledge from large amounts of data. Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. While data mining and knowledge discovery in database are frequently treated as synonyms, data mining is actually part of the knowledge discovery process [3]. The sequences of steps identified in extracting knowledge from data are shown in Figure 1.

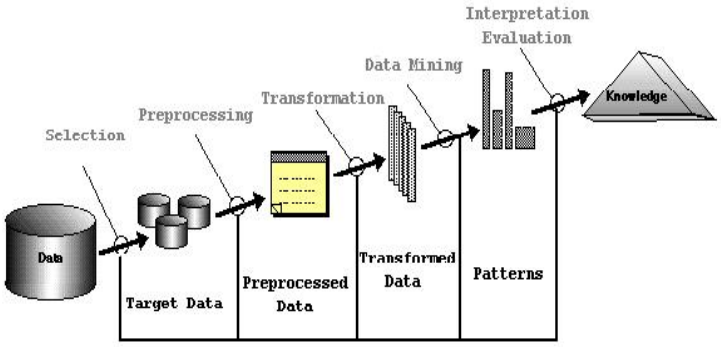

Fig 1: An overview of the steps that compose the KDD process

3.2 Data Mining Tasks

The cycle of data and knowledge mining comprises various analysis steps, each step focusing on a different aspect or task. It proposes the following categorization of data mining tasks [7].

3.2.1 Description and Summarization

At the beginning of each data analysis are the wish and the need to get an overview on the data to see general trends as well as extreme values rather quickly. It is important to familiarize with the data to get an idea what the data might be able to tell you where limitations will be and which further analyses steps might be suitable. Typically, getting the overview will at the same time point the analyst towards particular features, data quality problems and additional required background information. Summary tables, simple univariate descriptive statistics and simple graphics are extremely valuable tools to achieve this task.

3.2.2 Descriptive Modeling

Descriptive modeling tries to find models for the data. The aim of this model is to describe not to predict models. Even these models are used in the setting of unsupervised learning. Various methods of descriptive modeling are density estimation, smoothing, data segmentation, and clustering. The most widely used method of clustering is k-means clustering. The reasoning behind cluster analysis is the assumption that the data set contains natural clusters which, when discovered, can be characterized and labeled. While for some cases it might be difficult to decide to which group they belong, we assume that the resulting groups are clear-cut and carry an intrinsic meaning. In segmentation analysis, the user typically sets the number of groups in advance and tries to partition all cases in homogeneous subgroups.

3.2.3 Predictive Modeling

The aim of this task is to build a model that will permit the value of one variable to be predicted from the known values of other variables. Predictive modeling falls into the category of supervised learning; hence, one variable is clearly labeled as target variable and will be explained as a function of the other variables. The nature of the target variable determines the type of model: classification model, if it is a discrete variable or regression model, if it is a continuous one. Many models are typically built to predict the behavior of new cases and to extend the knowledge to objects that are new or not yet as widely understood.

3.2.4 Discovering Patterns and Rules

The area of the previous tasks has been much within the statistical tradition in describing functional relationships between explanatory variables and target variables. There are situations where these functional relationships are too hard to achieve in a meaningful way. So Association Rules are a method originating from market basket analysis to elicit patterns of common behavior.

3.2.5 Retrieving Similar Objects

The World Wide Web contains an enormous amount of information in electronic journal articles, electronic catalogs and private and commercial homepages. Having found an interesting article or picture, it is a common desire to find similar objects quickly. Based on key words and indexed meta-information search engines are providing us with this desired information. They can not only work on text documents, but to a certain extent also on images.

3.3 Data Mining Techniques

3.3.1 Association

Association is the discovery of togetherness or connection of objects. Such kind of togetherness or connection is termed as Association rule. An association rule reveals the associative relationship between objects, i.e. the appearance of set of objects. The association rule can be useful for marketing, commodity management, advertising, etc. For example, a retail store may discover that people tend to buy soft drinks together with potato chips and then put the potato chips on sale to promote the sale of soft drinks.

3.3.2 Clustering

Clustering is the identification of classes, also called clusters or groups, for a set of objects whose classes are unknown. The objected are so clustered that the intraclass similarities are maximized and the interclass similarities are minimized based on some criteria defined on the attributes of the objects. Once the clusters are decided, the objects are labeled with their corresponding clusters and common features of the objects in a cluster are summarized to form the class description. For example, a bank may cluster its customers into several groups based on the similarities of their age, income, residence, etc. and the common characteristics of the customers in a group can be used to describe that group of customers. The clusters will help the bank to understand its customers better and thus provide more suitable products and customized services.

3.3.3 J48

J48 is a version of an earlier algorithm developed by J. Ross Quinlan, C4.5. Decision trees are a classic way to represent information from a machine learning algorithm and this offer a fast and powerful way to express structures in data. It is important to understand the variety of options available when using this algorithm, as they can make a significant difference in the quality of results. J48 in WEKA3.6.0 employs two pruning methods. The first is known as sub tree replacement. This means that nodes in a decision tree may be replaced with a leaf basically reducing the number of tests along a certain path. This process starts from the leaves of the fully formed tree, and works backwards toward the root. The Second type of pruning used in J48 is termed sub tree raising. In this case, a node may be moved upwards towards the root of the tree, replacing other nodes along the way. Sub tree raising often has a negligible effect on decision tree models. There is often no clear way to predict the utility of the option, though it may be advisable to try turning it off if the induction process is taking a long time. This is due to the fact that sub tree raising can be somewhat computationally complex [2].

3.3.4 Classification

Classification trees are used to predict the classes of a categorical dependent variable from their measurements on one or more predictor or independent variables. Decision Trees have emerged as a powerful technique for modeling general input / output relationships. They are tree – shaped structures that represents a series of roles that lead to sets of decisions. They generate rules for the classification of a dataset and a logical model represented as a binary (two – way split) tree that shows how the value of a target variable can be predicted by using the values of a set predictor variables. Decision trees, which are considered in a regression analysis problem, are called regression trees. Thus, the decision tree represents a logic model of regularities of the researched phenomenon.

4. ANALYSIS

To predict Accident Severity, various classification models were built using Decision tree, Random Forest, ID3, Functional Tree, J48 and Naive Bayes. Decision trees are easy to build and understand can manage both continuous and categorical variables and can perform classification as well as regression.

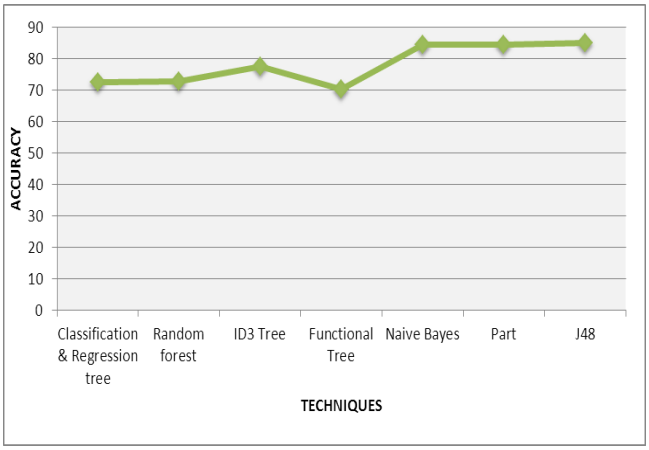

Fig 2: Graph of Accuracy against various Techniques

Figure 2 shows the graph of Accuracy against various techniques used. The statistics shows that having a means of predicting likely accuracy of different techniques base on some input values. It is evident from the line graph that value of Classification & Regression tree is slightly less than the random forest and ID3 tree. Classification models are generated on the basis of the training data whose independent variables and target variables are known, to be applied for the new dataset whose objective is the prediction of the target variable. The principle of CART method in developing the classification tree is described in the following: at first all data are concentrated at the root node, situated at the top of the tree. Further, it will be divided into 2 child nodes, on the basis of an independent variable (splitter), which create the best purity. A decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. It further reveals that value of Naive Bayes and PART is slightly greater than the J48 technique.

Table 4. Detailed accuracy by different techniques

S.no	Techniques	Accuracy
1	Classification & Regression tree	72.49
2	Random forest	73
3	ID3 Tree	77.7
4	Functional Tree	70.27
5	Naive Bayes	84.66
6	Part	84.64
7	J48	85.18

Table 4 illustrates a detailed accuracy of different data mining techniques used in the traffic injury severity over different datasets. It further reveals that value of Naïve Bayes, Part and J48 techniques are approximately same accuracy. The values are nearby 84.66%, 84.64%, 85.18%. In the end, I have to conclude that J48 technique has higher accuracy than other techniques.

5. CONCLUSION

Data mining in recent years with the database and artificial intelligence developed a new technology, its aim the large amount of data from the excavated useful knowledge, to achieve the effective consumption of data resources. A thorough literature review revealed a gap in published studies on the relationship between road characteristics and RTA severity. The study on road traffic accident causes can identify the key factors rapidly and efficiently and provide instructional methods to the traffic accidents prevention and road traffic accidents reduction, which could greatly reduce personal casualty and property loss caused by road traffic accidents. Meanwhile, it would be helpful for improving the efficiency and security service level of the road transportation system.

6. REFERENCES

[1] Ali Tavakoli Kashani, Afshin Shariat, Andishe Ranjbari ," A Data Mining Approach to identify key factors of traffic injury severity" Traffic & Transportation, Vol. 23, 2011, No. 1, 11-17.

[2] Bouckaert Remco, Eibe Frank, Mark Hall, Richard Kirkby, Peter Reutemann, and Alex Seewald, 2008.WEKA Manual for Version 3-6-0. University of Waikato, New Zealand.

[3] Brijesh Kumar Baradwaj, Saurabh Pal," Mining Educational Data to Analyze Students Performance" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011

[4] Chaozhong Wu, Ming Ma, Hu Lei, Xinping Yan,"Severity Analyses of Single-Vehicle Crashes Based on Rough Set Theory" 2009 International Conference on Computational Intelligence and Natural Computing.

[5] DipoT.Akomolafe, Akinbola Olutayo," Using Data Mining Technique to Predict Cause of Accident and Accident Prone Locations on Highways" American Journal of Database Theory and Application 2012, 1(3): 26-38.

[6] Han, Jiawei and Kamber, Micheline. (2006). Data Mining: concepts and Techniques. San Fransisco; Morgan kufman Publishers

[7] Hand, D.J., Mannila, H., and Smyth, P. (2001). Principles of Data Mining, MIT Press

[8] http://en.wikipedia.org/wiki/weka (machine learning)/ accessed on May 2014.

[9] http://rapid-i.com/content/view/181/190/ accessed on May 2014.

[10] Liping Guan, Weiming Liu, Xiangyuan Yin, Luping Zhang," Traffic Incident Duration Prediction Based on Artificial Neural Network" 2010 International Conference on Intelligent Computation Technology and Automation

[11] Mehmed Kantardzic (2003).Data mining: Concepts, Models, Methods, and Algorithms, ISBN13: 9780471228523, John Wiley & Sons Publisher

[12] Pasko Konjevoda and Nikola Stambuk, "Open-Source Tools for Data Mining in Social Science," Theoretical and Methodological Approaches to Social Sciences and Knowledge Management, pp.163-176

[13] S.Krishnaveni, Dr. M.Hemalatha," A Perspective Analysis of Traffic Accident using Data Mining Techniques", International Journal of Computer Applications (0975 – 8887) Volume 23– No.7, June 2011.

[14] S.Shanthi, R.Geetha Ramani " Feature Relevance Analysis and Classification of Road Traffic Accident Data through Data Mining Techniques" Proceedings of the World Congress on Engineering and Computer Science 2012 Vol. I WCECS 2012, October 24-26, 2012, San Francisco, USA

[15] Tibebe Shah, Shawndra Hill (2013)," Mining Road Traffic Accident Data to Improve Safety: Role of Road- related Factors on Accident Severity in Ethiopia".

[16] Tanagra – a Free Data Mining Software for Teaching and Research, Available at: http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html.

[17] WHO (2004),"World report on road traffic injury prevention", Switzerland, Geneva

7. AUTHOR'S PROFILE

Dheeraj Khera completed the B-Tech. degree in Computer Science Engineering in 2011 Punjab Technical University, Jalandhar (Punjab). Presently he is working as Lecturer in Bhai Gurdas Global Polytechnic College, Rakhra and pursuing his M-Tech. in Computer Engineering Department, Punjabi University, Patiala (Punjab).

Williamjeet Singh received his B.Tech, M.Tech degree in Computer Science and Engineering in 2005 and 2007 respectively from Punjab Technical University, Jalandhar (Punjab) and Punjabi University, Patiala (Punjab). Presently he is working as Assistant Professor and pursuing his PhD from Computer Engineering Department, Punjabi University, Patiala. His current research interests are primarily in the area of Wireless networks, Algorithms and Data mining**.**