# PREDICTION OF THE CAUSE OF ACCIDENT AND ACCIDENT PRONE LOCATION ON ROADS USING DATA MINING TECHNIQUES

Ms. Gagandeep Kaur
Department of Computer Engineering
Punjabi University
Patiala, India
gaganmalhotra1791@gmail.com

Er. Harpreet Kaur
Department of Computer Engineering
Punjabi University
Patiala, India
khasria.harpreet@gmail.com

***Abstract*-Road Accident is a specific instance of traumatic events that constitute major loss. Data mining tools and techniques are used to predict the likelihood of accident and accident prone locations. This paper sheds light on predicting the probability of accidents on roads with special emphasis on STATE HIGHWAYS (SHs) and ORDINARY DISTRICT ROADS (ODRs) by estimating the severity of accidents based on the type of accident, type of spot using the R tool. Pointing out the traffic collision data of roads the frequency of traffic collision of roads is analyzed using correlation analysis and exploratory visualization techniques. Finally, methodology has been proposed to analyze road traffic accidents. Using this methodology, improvement can be promised at the level and extent of road traffic safety management effectively and efficiently. The present study modeled accident and incident data gathered from the traffic data and data related to construction sectors.**

***Keywords-Data Mining using RStudio, Correlation Analysis, Road Accidents, Exploratory Visualization Techniques*.**

## I. INTRODUCTION

Road accidents severity is increasing at alarming rate. Regulating the traffic accidents on roads is a vital task. State Highways (SH's) and Ordinary District Road (ODR's) form the pecuniary cornerstone of the state. The importance of the study is to analyze the traffic accident data factors of State Highways (SH's) and Ordinary District Roads (ODR's). Perceiving the solemnness of the issue restorative measures are taken to curb the menace caused by road accident. The rationale behind investigation is to analyze the accidental data using data mining techniques on various categories of roads such as plan roads, link roads and central government funded roads (PMGSY Roads).Data mining is a approach for locating interesting trends as well as illustrative, comprehensible models from huge datasets [10].The knowledge driven version deals with harvesting and analyzation of data ,demand-driven collection of sources of information, security and privacy contemplation, modeling according to user interests [11]. Data dredging is applied on the traffic and construction sector data. Data mining is fitting a key technology for identifying doubtful activities [14].Higher severities in our nation is due to Road collisions [1].Statistical analysis suggests that higher causalities in the road collisions are due to various factors such as lack of proper analyzation of accidental factors, wrong pathway user behavior, imperfect road design. The specific objective of the research is to analyze and predict the black spots, accident prone zones and road conditions using exploratory visualization techniques and regression analysis. The steps of data mining which deals with analyzing the accidental data from different views and perspectives are depicted in Fig. 1.

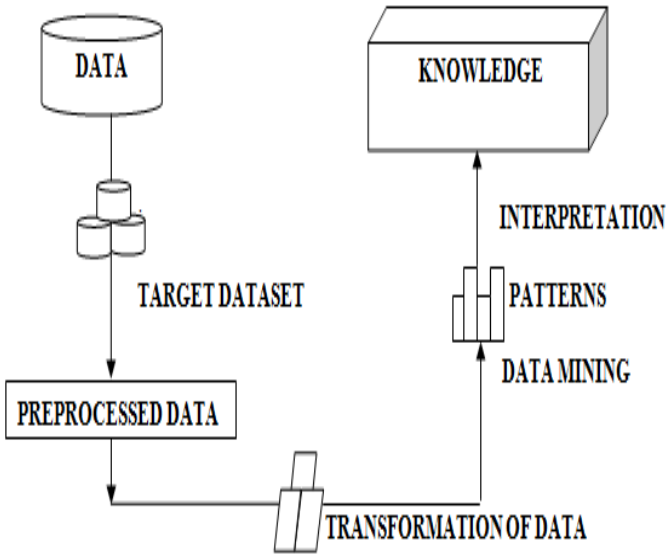


**Figure1. Representing the steps involved in Data Mining**

The paper has been organized into various sections. Section II describes the work related to the area of research, discussions related to prior work and contributions in the concerned field. Section III illustrates the methodology used in performing the analysis. Section IV is related to the simulation and hence predicts the results. Section V defines the proposed work. Then concluded with conclusion and references.

## II. RELATED WORK

The work presented here has focused on parametric analyzation of various factors contributing to collisions using various tools, techniques and methodologies. Table. 1 depicts the work related to the area of research which includes the method used, input attributes, simulation and the results of various research work.

Today, one of the main work of government is the traffic safety. Pointing out the vitalness of topic, identifying the reasons of road accidents has become the main focus to reduce the harm caused by traffic accidents. The data mining outcomes will help the various organizations such as transportation, to inquire the accidents data recorded by the police information system, discover hidden patterns and trends thus predict the future consequences. Efficient and effective decisions are taken to lessen accidents [8].

Categorizing the major aspect of traffic collisions and their severity assists the highway safety development in improving the road conditions according to the variation in vehicle occupancy need of specific sector of the population. Different algorithms of data harvesting such as K-Means clustering and Self Organizing Map clustering algorithms are applied on accidental data and the outcome of the study states the classification of major contributing parameters to traffic accidents [9].

According to MORTH-2013 India has the largest no. of accidents in the World. Severity of accidents has been increasing year by year. Hence safety of road is major concern. A study was taken which joins to various vital industries and mines. The interpretation states that main accident casualties are due to accused vehicle such as trucks. The major reasons of collisions are due to high vehicle occupancy of heavy automobiles, non restriction of speed, on street parking, edge drop, visibility restriction etc [12].

Data harvesting is a booming field that relates to huge volumes of data and discover interesting patterns and trends from the dataset. Various data mining techniques, tools and applications are analyzed in this paper. Crash analysis methodologies are based on the occurrence of accident scenarios and simulation of collision situation. Experts have attempted to build safer vehicles but the crashes are still unavoidable [7].

**Table 1. Depicting the work related to Area of Research**

| Year/Author | Method | Input | Simulator | Output |
|---|---|---|---|---|
| 1)2017 AshishDutt, Maizatul Akmar Ismail, Tutut Herawan | Educational Data Mining (EDM) applied to resolve education related problems through psycho- pedagogy, cognitive psychology, and recommender system methods and techniques. | - | - | Systematic and structured review on clustering algorithm has presented in this paper [2]. |
| 2)2016-Geltmar von Buxhoeveden, Uwe Becker | Simulation techniques, modeling for traffic safety are defined as the methodology. | Data from various sources were gathered to provide a comparison of train, bus, and car accidental data. | To construct a web based tool for interactive data exploration and analysis, R programming language is used. | The R language is used for the aggregation and analysis of public transport and individual transport data [3]. |
| 3)2015-Rupanjan Chakraborty, Sarabjt Bhattacharyya, Mrial Roy,Pinak Paul | Linear regression model, A log-linear model. | Data input in the form of factors affecting accidental data. | - | To predict the number of Accidents per stretch [5]. |
| 4)2015 Dheeraj Khera Willamjeet Singh | Performance is evaluated by using Naïve bayes, ID3& Random tree. | Data mining algorithm using various performance criteria. | Study is to scrutinize the performance of different taxonomy method using WEKA&Tanagra | Tanagra tool is the best and give the highest Accuracy measure [4]. |

| | | | | |
|---|---|---|---|---|
| | | | tools. | |
| 5)2014<br>A.Priyanka<br>K.Sathiyakumari | Machine learning algorithms are used for developing a DSS to deal with traffic accident data analysis. | Databases of Road traffic accident provide the origin for road traffic accident investigation. | Weka version3.7.9 | SMO algorithm was accurate and provides 94% accuracy prediction [7]. |
| 6)2014<br>Manisha Birdi<br>Dr.R.C Gangwar<br>Prof GurpreetSingh | Various algorithms such as Self Organizing Map Clustering and K-means algorithms used to investigate the accidental data. | Traffic accident data of rural highways is gathered for a time duration of six months. | - | Generate human interpretable clusters after Pre-processing highway accidental data[9]. |
| 7)2013<br>A.N Dehruy<br>A.K Patnaik<br>A.K Das,U.Charttraj<br>P.Bhuyan,M.Panda | - | Data related to hourly, monthly and annual variation on selected stretch of NH. | - | Major cases are due to trucks. The major reason of accidents is due to on street parking, edge drop off [12]. |
| 8)2013<br>Vishrut Landge<br>A.K Sharma | Statistically methods and regression technique are used to identify the accident prone location. | Data in the form of factors responsible for accident. | - | Parameters like speed, percentage of heavy vehicle play major role in determining the rate of crashes [13]. |
| 9)2013-R.R Sorate<br>R.P Kulkarni, M.S. Patil,S.U. Bobade, A.M.Talathi, S.V.Apte, I.Y.Sayyad | Methodology adopted includes collecting secondary data from respective authorities. | Data in the form of y intersection, super elevation etc. | - | Locating the black spot suggest the remedial measure [6]. |
| 10)2009<br>K.Jayasudha<br>Dr.C.Chandrasekar | Accident analysis methodologies are based on scenarios of the occurrence of accidents and simulation of accident prone area. | Real life traffic accidental data is analyzed. | Orange<br>Tanagra<br>Rapid miner. | To overcome the death rate using various tools and techniques [15]. |
| 11) 2007<br>Jiawei Han<br>Hong Cheng<br>Dong Xin<br>Xifeng Yan | Pattern mining on structured and sequential data, correlation mining, clustering, frequent item set mining. | Frequent patterns item sets, substructures and subsequences. | - | Frequent item set harvesting has claimed tremendous progress [16]. |

## III. METHODOLOGY USED

Methodology selected comprises gathering the secondary data from the Executive Engineer Construction Division PWD B&R, conducting the physical survey and analyzing through data mining techniques.

The Methodology is applied using R's IDE Integrated Development Environment (Rstudio) which is graphical and statistical computing tool on road accidental dataset for analyzation. R is free software, GNU package used for analyzing the datasets to vaticinate the hidden patterns and trends. The flow chart that describes the whole data mining process is depicted in Fig. 2.

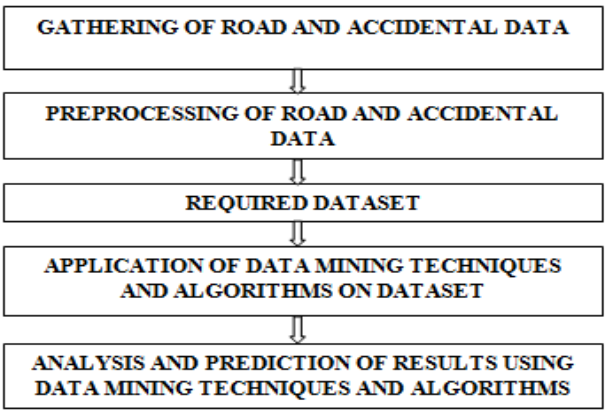


**Figure2. Presenting steps of Data Mining of Crash data**

## IV. Simulation And Results

The simulation is performed by using RStudio which is a Integrated Development Environment (IDE) for R tool. Correlation analysis and exploratory visualization techniques has been applied on various parameters of accidental data of roads to analyze and predict the useful results which help to minimize the frequency of accidents and determine the road conditions. Correlation is bivariate analysis that measures the strength of association between two variables. Pearson Correlation is the parametric measure of linear association. Correlation can be negative or positive. After analysis of correlation between two variables we conclude that positive correlation means if one variable increase or decrease the other tend to increase or decrease. Negative correlation is vice versa.

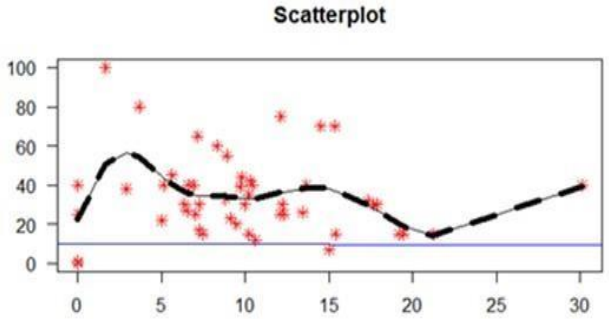


**Figure3. Showing correlation analysis where X axis depicts Length in kms and Y axis depicts Progress in percentage**

Fig. 3 shows the relation between length and progress of road through Scatter plot of PLAN ROAD (PMGSY). It has been observed from the plot that there is negative relationship between these two parameters. As the length of road increased the progress is not uptomark or less progress on the road.

**Table 2. Representing the accidental data related to Road and the Type of Spot**

| Road Name / Type Of Spot | SH-Samana Patiala Road | SH-Patiala Pehowa road | ODR-Patran Moonak road | ODR-Khanouri Arno road |
|---|---|---|---|---|
| Straight Road | 20 | 20 | 2 | 0 |
| Cross Intersection | 0 | 0 | 7 | 20 |
| R-Intersection | 0 | 0 | 7 | 0 |
| Y-Intersection | 0 | 0 | 0 | 0 |
| Toll Plaza | 0 | 0 | 0 | 0 |
| Narrow Bridge | 0 | 0 | 0 | 0 |
| Near the Curve | 0 | 0 | 4 | 0 |

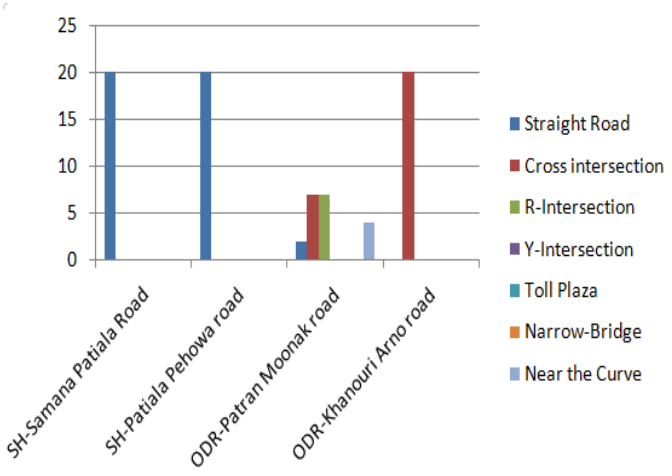


**Figure4. Showing accidental data of State Highways and Ordinary District Roads on the basis of Type of Spot through exploratory visualization technique where X axis depicts Road name and Y axis depicts frequency of Type of Spot**

In Fig. 4 the accidental data of four roads are analyzed out of which two are State Highways and other two are Ordinary District Roads. The numerical data is given in Table 2. From the figure it is concluded that frequency of accidents on the State Highways (SHs) are maximum on the straight roads and minimum on other type of spots. Frequency of accidents on Ordinary District Roads can be on straight roads, cross-intersection, r-intersection but mainly on cross-intersection.

**Table 3. Depicting accidental data of Straight Highway Road I SH-Samana Patiala**

| Accident Type → / Accident Spot ↓ | Head on | Head tail | Sideswipe | Overturn | Hit and Run |
|---|---|---|---|---|---|
| Straight Road | 6 | 3 | 3 | 5 | 3 |

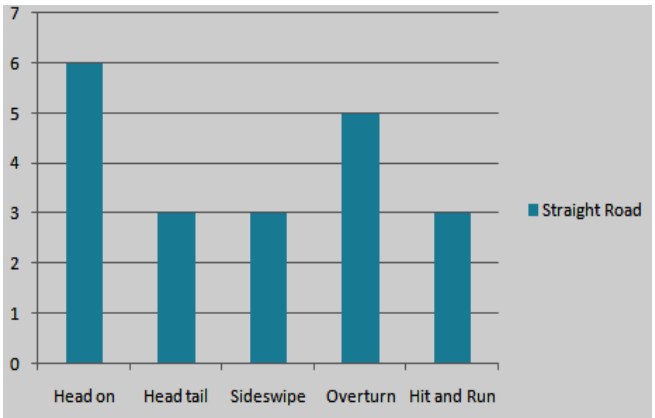


**Figure5. Depicting Road I SH-Samana Patiala Road accidental data where X axis depicts Accident Type and Y axis depicts the frequency of Type of Spot**

**Table 4. Representing accidental data of Straight Highway Road II SH-Patiala Pehowa**

| Accident Type → / Accident Spot ↓ | Head on | Head tail | Sideswipe | Overturn | Hit and Run |
|---|---|---|---|---|---|
| Straight Road | 6 | 3 | 3 | 5 | 3 |

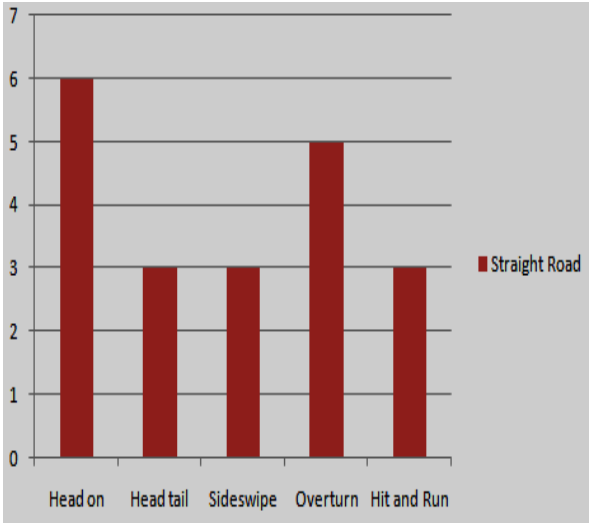


**Figure6. Depicting Road II SH-Patiala Pehowa Road accidental data where X axis depicts Accident Type and Y axis depicts the frequency of Type of Spot**

**Table5. Showing Accidental data of Ordinary District Road III ODR-Patran Moonak Road**

| Accident Type → / Accident Spot ↓ | Head on | Head tail | Hit pedestrian | Right angle |
|---|---|---|---|---|
| Cross Intersection | 5 | 0 | 2 | 0 |
| R-Intersection | 2 | 1 | 0 | 4 |
| Straight Road | 0 | 1 | 1 | 0 |
| Near the Curve | 1 | 1 | 2 | 0 |

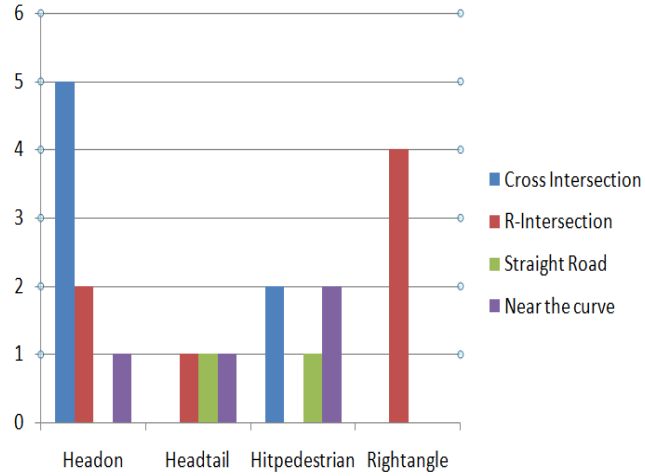


**Figure7. Depicting Road III ODR-Patran Moonak accidental data where X axis depicts Accident Type and Y axis depicts the frequency of Type of Spot**

**Table 6. Displaying Accidental data of Ordinary District Road IV ODR-Khanouri Arno Road**

| Accident Type → / Accident Spot ↓ | Head on | Head tail | Overturn |
|---|---|---|---|
| Cross Intersection | 10 | 6 | 4 |

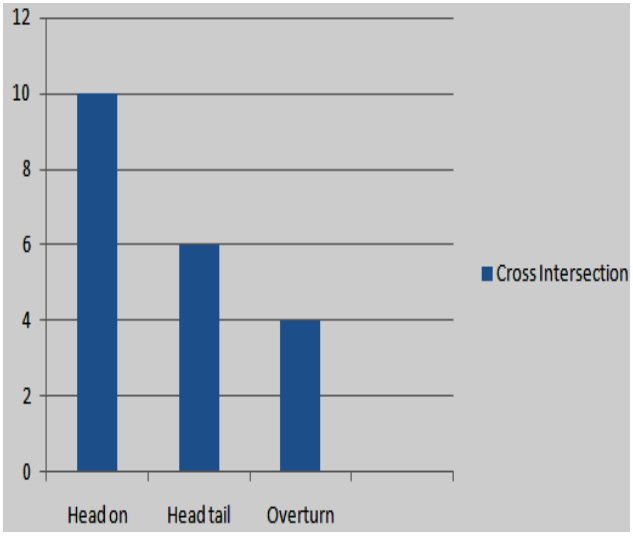


**Figure8. Depicting Road IV ODR-Khanouri Arno Road accidental data where X axis depicts Accident Type and Y axis depicts the frequency of Type of Spot**

Table 3 and 4 and its corresponding graphs i.e. Fig. 5 and 6 depicts that mainly accidents on State highways occur on Straight roads are of Head on collision type. Table 5 and 6 and its corresponding graphs i.e. Fig. 7 and 8 shows that majorly the accidents on Ordinary District Roads occur on Cross intersection are of Head on collision type.

## V. Proposed Work

The problem is formulated in order to do parametric investigation on attributes by considering traffic data such as the type of accident, the type of spot, type of intersection etc. The traffic accidental data over the last 3 years has been analyzed from time period (2012-2015) of State Highways and Ordinary District Roads and the Division name is Executive Engineer Construction Division PWD B&R using various data harvesting approaches such as Exploratory visualization techniques. The analysis of road conditions of PLAN ROAD (PMGSY) helps to derive the relation between the two numerical variable of road using Regression analysis by using a specialized tool called R tool. RStudio is an IDE for using R. The limitation of this approach is that though Exploratory visualization techniques depicts many crucial aspects such as frequency distribution of enormous data categories and summarize dataset in pictorial form, still additional explanation is required that expose the key trends, hidden patterns, classify new instance, defines similar and homogeneous , interesting relationship areas which can achieved through various algorithms such as K nearest neighbor(KNN) , K-means in the proposed work. The analysis of K-means and KNN will be performed in the future. Fig. 9 presents the proposed work steps.

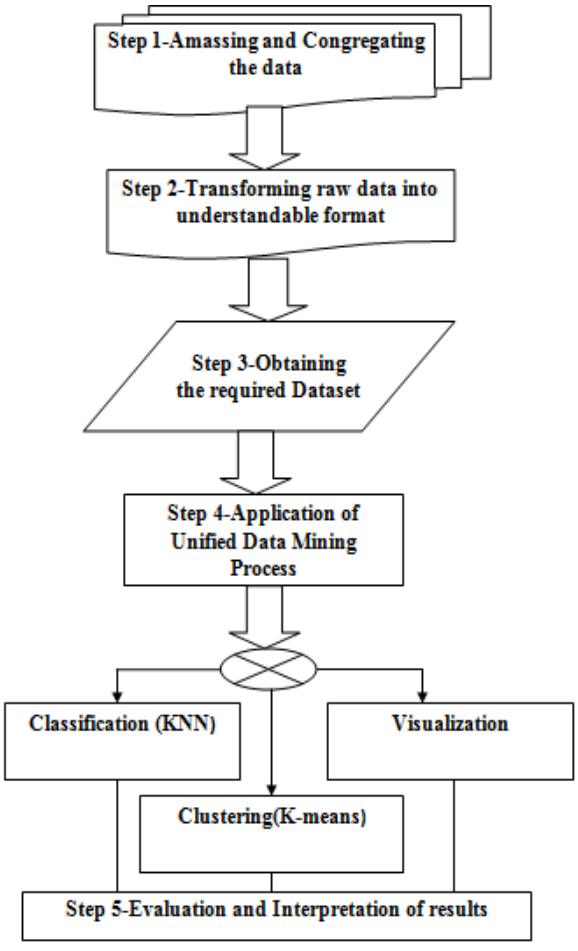


**Figure 9. Steps representing the Proposed Work**

## VI. Conclusion

The study helps us to derive the statistical model by using various techniques. Correlation analysis examines the road conditions that help to derive the relation between two numerical variables i.e. length and progress of road which is negative that shows the inverse relation. From the study of Exploratory visualization techniques we concluded that accident on State Highways occur on Straight roads and on Ordinary District Roads the accidents can occur on other type of spots such as Cross-intersection, R-intersection and Straight road but majorly on Cross- intersection. Mainly the type of accident occurs is Head on collision type on both Roads. Future work will be to analyze the cause of severity of accidents by considering other parameters such as non restriction of speed, old girth trees on shoulder, shoulder drop-off etc.

## References

[1] https://en.wikipedia.org/wiki/Pradhan_Mantri_Gram_Sadak_Yojana.

[2] Maizatul Akmar Ismail, Tutut Herawan, Ashish Dutt, "A Systematic Review on Educational Data Mining," *IEEE*, 2017.

[3] Uwe Becker, Geltmar von Buxhoeveden, "Comparison of Traffic Incident Data in Individual and Public Transport," in *International Conference on Systems and Informatics* , 2016, pp. 1067-1071.

[4] Williamjeet Singh, Dheeraj Khera, "Prediction and Analysis of injury Severity in traffic systemusing data mining technique," *National Journal of Computer Applications*, 2015.

[5] Sarbajit Bhattacharyya, Mrinal Roy, Pinak Paul , Rupanjan Chakraborty, "Accident Analysis and the Suggestion of an Accident Prediction Model for Guwahati city," *Internatinal Journal of Innovative research in Science,Engineering and Technology*, 2015.

[6] R.P. Kulkarni, S.U. Bobade, M.S. Patil, A.M. Talathi, I.Y.Sayyad, S.V.Apte, R.R.Sorate, "Identification of Accident Black Spots on National Highway 4," , 2015.

[7] K. Sathiyakumari, A. Priyanka, "A Comparative Study Of classification algorithm using accident data," *International Journal of Computer Science & Engineering Technology (IJCSET)*, 2014.

[8] Somayya Ebrahimkhan, Bahram Sadeghi Begham, Farzaneh Moradkhani, "Road Accident Data Analysis-A Data mining Approach," 2014.

[9] Prof. Dr. R C Gangwar, Prof. Gurpreet Singh, Manisha Birdi, "A Data Mining Clustering Approach for Traffic Accident Analysis of National Highway-1," *International Journal of Advanced Research in Computer Science and Software Engineering*, 2014.

[10] Manish Mann, Bharti Thakur, "Data Mining for Big Data: A Review," *International Journal of Advanced Research in Computer*

*Science and Software Engineering* , pp. 469-473, 2014.

[11] Xingquan Zhu, Gong-Qing Wu, Wei Ding , Xindong Wu, "Data Mining with Big Data," *IEEE*, vol. 26, no. 1, pp. 97-107, 2014.

[12] A.K Patnaik, A.K Das, U.Charttraj, P Bhuyan, M.Panda, A.N Dehruy ,"Accident Analysis and Modeling on NH-55(India)," *International Journal of Engineering Inventions*, pp. 80-85, 2013.

[13] A.K.Sharma , Vishrut Landge, "Identifying Accident Prone Spot on Rural Highways - A Case Study of National Highway No 58," , 2013.

[14] Dr. Arun Sharma, Abu Sarwar Zamani, Ali Akhtar, Shakir Khan, "Data Mining For Security Purpose & Its Solitude Suggestions," *International Journal of Scientific & Technology Research*, vol. 1, no. 7, 2012.

[15] Dr.C.Chandrasekar, K.Jayasudha, "An Overview of Data Mining in Road Traffic and Accident Analysis," *Journal of Computer Applications*, 2009.

[16] Hong Cheng , Dong Xin , Jiawei Han, Xifeng Yan,"Frequent pattern mining: current status and future," , 2007.