A Data Mining Framework to Analyze Road Accident Data using Map Reduce CCMF and TCAMP Algorithms

S. Nagendra Babu [1], J. Jebamalar Tamilselvi [2]

[1] R &D Center, Bharathiar University, Coimbatore, India. E-mail: s.nagendrababu@gmail.com
[2] Jaya Engineering College, Thiruninravur, Chennai, India. E-mail: jjebmalar@gmail.com

***Abstract* - Accident prediction is an important safety issue to raise alarms before accidents happen. In this paper we formulate relevant questions to anticipate the occurrence of accidents and process the available information using Hadoop. We examine the execution time on Hadoop when compared with other methods, and propose 2 algorithms to use Congestion Control Machine Framework (CCMF) and Traffic Congestion Analyzer using Map Reduce (TCAMP) to effectively analyze the available data and assess road accident reasons and advise authorities to take appropriate actions to increase road safety. We apply information mining to examine recorded road attributes to reduce road accidents specifically in India, and formulate a set of standards that can be utilized by the National Highway Authority of India to improve safety.**

Keywords - Big data, data clustering, map reduce methods, Hadoop framework, road accident data.

I. INTRODUCTION

Lately the accumulation of data on movement volumes has turned into a noteworthy bit of crafted by road arranging programs as far as both cost and staff. Movement information is parceled into various administrations by distinguishing breakpoints for activity factors in the information. In two-administration movement models, basic inhabitance is utilized to isolate free stream and congested stream conditions. Perceptions with inhabitance esteems littler than basic inhabitance are thought to be in the free stream administration and perceptions with inhabitance esteems more prominent than basic inhabitance are thought to be in the congested stream administration. Distinguishing the basic inhabitance esteem from field perceptions isn't trifling. Information mining is the way toward finding intriguing learning, for example, designs, affiliations, changes, oddities and noteworthy structure from a lot of information put away in databases, information stockrooms, or other data archives.

When a dataset is considered it has to undergo different stages for getting structured data and relevant and complete dataset preprocessing techniques must be applied on dataset and then clustering is performed so only relevant information is formed as a group. Then Map Reduce methods are applied on the dataset so that accident prediction is accurate and effective. The block diagram of the proposed work is depicted below.

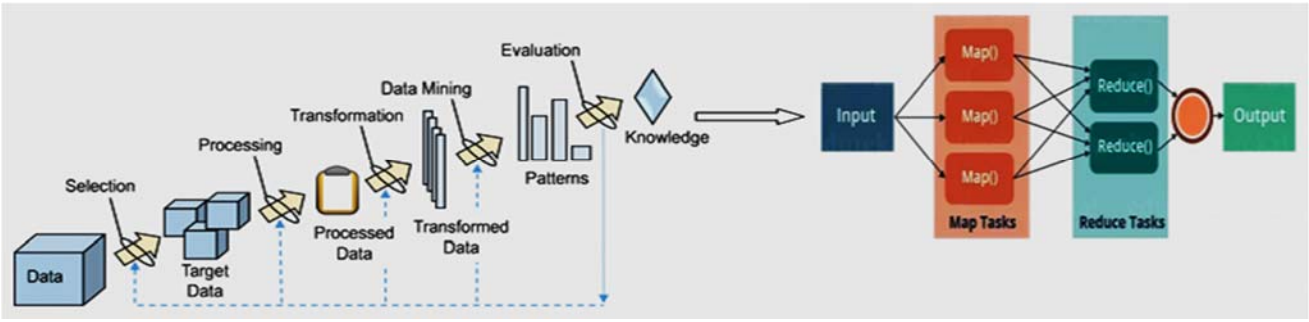

Figure 1. Block diagram of proposed method

A standout amongst the most valuable uses of the frameworks utilized in rush hour gridlock control is the enhanced capacity to control the road arrange movement. Following strategies are utilized to catch the position and additionally area of monstrous number of portable articles. With the assistance of that followed data, investigation and prediction of activity thickness in a given system is upgraded. This renders important data for controlling activity stream, forecast of clog and decreasing the quantity of Accidents in that system.

Map Reduce is a programming, show. It comprises of the mapper and reducer period of creating and handling the extensive arrangement of information. The capacity of the mapper is to take the contribution to a couple of key and

esteem and the capacity of reducer is to deal with that middle of the road key and esteem. The key esteem is only the information which is identified with that specific assignment, i.e. the esteem and the gathering of the no. of esteem is a key. At that point it is sent to the reducer. The consolidation procedure is done by reducer which combines these arrangement of significant worth with a specific end goal to get the little arrangement of qualities into a similar hub. The distinctive esteems that the reducers stage get having a similar key into hubs.

Map Reduce is a system for playing out the parallel preparing of an expansive arrangement of information i.e unstructured and organized. It comprise of two stages, which are mapper and reducer stage, it takes the key/esteem match as an information, and play out some task on this info and deliver the applicable outcome as key/esteem. To process these outcomes lessen stage is required as determined by the diminish work. The information from the mapper stage is rearranged, which implies the information is traded and consolidate arranged, to the machine with a specific end goal to play out the lessen stage.

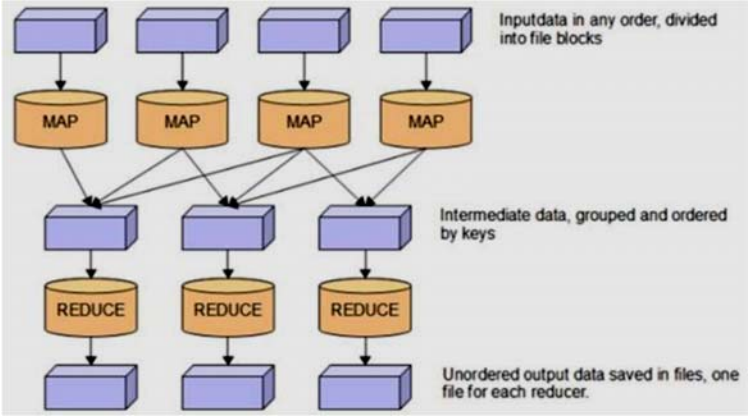

Figure 2. Map-Reduce framework

II. LITERATURE REVIEW AND LIMITATIONS OF CURRENT METHODS

At exhibit, activity data is ordinarily shared between various movement offices by methods for voice or information interchanges. For such information correspondence between administration focuses, a typical dialect and a casing of reference is required. There are numerous conceivable approaches to examine activity information like physically checking every vehicle or through picture preparing, video examination et cetera. Manual checking includes a group to tally every vehicle going out and about and refresh the insights and a different group is required to break down these aftereffects of measurements.

Sun and Zhou (2005) apply gathering techniques in the appearing of multi-administration speed-thickness relations. Grouping procedures are utilized to perceive the disappointment focuses in a speed-thickness graph, speed-thickness information is then partitioned in light of the perceived partitions, and wrinkled return strategies is utilized to make multi administration speed-thickness relations. K-implies gathering technique is connected to three informational collections from three conduit segments in San Antonio, Texas. Speed-thickness information is then gathered to two and three gatherings.

Xia and Chen (2007) arrange throughway working settings utilizing an agglomerative social event calculation. Stream, speed, and occupancy information from a freeway sensor in California are utilized as a contextual analysis. Bayesian Information Criterion (BIC) and dispersal estimation strategies are utilized to perceive the quantity of bunches. The investigation recommend that each group could speak to a freeway stream stage.

A. Existing Methods

Guide Function: The Map work takes the contribution from the info per user as key/esteem, play out the activity of guide work on it, and give the yield as another key/esteem combine.

Decrease Function: The combine with a similar key will get prepared in a gathering, the diminish work is called once for each unique key. It is ensured that the contribution to each diminish errand is handled to expand the key request. Amid the arranging procedure, the client characterizes its correlation capacity to be utilized.

There are numerous conceivable approaches to break down movement information like physically tallying every vehicle or through picture handling, video investigation et cetera. Manual checking includes a group to tally every vehicle going out and about and refresh the insights and a different group is required to break down these aftereffects of measurements.

III. PROPOSED METHOD

A. Data Preprocessing

The reason for information preprocessing is to separate helpful information from rough data set and after that change these information in to dataset, the first record can't be straightforwardly utilized in the dataset mining technique, thus in information preprocessing stage, rough dataset should be cleaned, dissected and changed over for additionally step. The information recorded in server records, for example, the road type, accident cause, time, and so forth, are accessible to distinguish users. In any case, since some parameters might be stored by the user program or by an intermediary server, we ought to realize that the information gathered by server are not by any stretch of the imagination. This issue can be mostly illuminated by utilizing a few different sorts of use data.

B. Data Clustering

In various leveled clustering, the number of things are indistinguishable to the measure of clusters(say n). The sets which are nearest to each other are joined into single cluster. After this estimation of the packet between new pack and every single one of old clusters. Repeating of the techniques is done until the point that everything is gathered into m no. of gatherings.

Distributing Clustering Methods like k-mean, Bisecting K Means Method, Medoids Method, PAM (Partitioning Around Medoids), CLARA (Clustering LARGE Applications) and the Probabilistic Clustering are goes under allocating name itself suggests that the data is isolated into number of subsets. Since it isn't computationally possible to check the all possible subset of the structures available that is the reason this procedures can be used to grouped sweeping information.

C. Association Rules

In the proposed work Improved (Association Rule Mining) ARM algorithm is characterized for association rule mining for road accident information expectation. The proposed algorithm is quite enhanced in association rule mining when contrasted with FP-Growth algorithm. This algorithm characterizes how standards can be set to characterize new techniques for mining regular designs.

D. Map Reduce Methodology

MapReduce is a product structure which was acquainted by Google with complete parallel preparing on vast informational indexes. This substantial informational index is dispersed over an extensive number of machines introduce in a bunch. For speedy access, each machine figures and stores the information locally, this thus adds to appropriated parallel handling. Such a calculation includes two sections – Map and Reduce. In the Map stage, information hubs take crude info information and create middle of the road information in light of the sort of calculation and afterward that information is put away locally. In Reduce stage, middle of the road yields from delineate is brought by the hubs and afterward it is consolidated to determine last yield that is put away in HDFS. Name hub with its earlier information on the information appropriation, tries to appoint the undertaking to a specific hub in light of the area of information. Designers can compose custom guide and lessen capacities appropriate to the application and the MapReduce work at that point deals with disseminating and parallelizing undertakings over a rack on item equipment in the group underneath.

A Map Reduce fill in generally speaking parts the data into various pieces and each of these are dealt with by the guide errands in parallel way. The Mapper maps the little endeavors by impacting usage of the key and motivator to consolidate thought and the yields are orchestrated. By then the Reducer diminishes the procured yields from the maps to get the last yield. The MapReduce structure contains a solitary Job Tracker as the expert and a singular Task Tracker as the slave for each pack center point. All data and yield in MapReduce are <key, value> sets. The Hadoop is a Java based appropriated programming condition bolstered by Apache that can be used to process and handle a considerable measure of data. Hadoop has been made using the possibility of MapReduce for broad getting ready by using a broad number of center points and clusters.

In the proposed technique at first we take the informational index and we give that informational collection for pre-handling, after pre-preparing we will get a perfect informational index which contains heterogeneous information writes. In the original copy two calculations are proposed Congestion control utilizing Machine Framework (CCMF) which is utilized for considering activity information by the machine system. Another calculation Traffic clog Analyzer utilizing Map Reduce (TCAMR) which understudy investigate the activity information for prediction of road Accidents. Take every one of those information and go for delineate based unsupervised grouping, from that point we will get a handled clusters, and at that point apply affiliation leads then we are now planned a machine with all the conceivable sources of info and furthermore tried, after pre-preparing we will ready to create the forecast. On the off chance that on the off chance that we may missed any sort of unique info or circumstance we will again create the new preparing and testing instruments to the machine.

Algorithm CCMF ()

{

1. Initially take the training data $t_1, t_2 \ldots . t_n$
2. Give each and every aspect of training data to the machine
3. After the completion of training go for testing
4. Give test data to the machine
    a. If any failure
    b. Go back to step 1 and include the failure case to train data
    c. Or test all the test case
5. And derive the prediction's

}

Algorithm TCAMP ()

{

Input Traffic data set

Output predicting the accidents

Step-1: take the traffic data set

Step-2: apply pre-processing

Step-3: apply Map function with an clustering algorithm

1. Partitioning the traffic data
2. Send the partitions onto different machines
3. Map the each partitions value into a key value pair

Styep-4 apply Reduce function

1. Shuffling
2. Reduce into unique key value pairs
3. Get the clusters

After

Step-5: take the clusters and apply the regressions

}

E. Accident Prediction Using Map/Reduce

We analyze our strategy in the light of the MapReduce for comprehending client asked questions from movement prediction information. The framework was based on Apache Hadoop for expanding preparing execution utilizing multi-hub bunch. In our investigation numerous questions are prepared which are outlined in table I.

The handling capacities of big data can precisely test road accident cases, its prognostic capacity can viably foresee the event of accident incident, utilizing microwave identification frameworks, video observation frameworks, versatile location framework, we can fabricate a successful security model to enhance the safety measures. At the point when security incident honed, and emergency is required, Because of its comprehensive handling and basic decision taking capacity, speedy answer ability, big data can significantly recoup the capacity of emergency, and decrease accidents.

Smart transportation framework on big data stage is a blend of various frameworks, models, division, innovations. One might say, It is an exhaustive arrangement of framework knowledge, administration knowledge, arithmetic, financial aspects, conduct knowledge, and data innovations.

TABLE I. ISSUES INCLUDED IN THE INVESTIGATION

Queries	Data set
Q1: Area where maximum accident occurs	Grid Reference Easting and Grid Reference Northing and Road surface
Q2: On which Highway Accidental Timing.	Accident date and Time and Road class
Q3: On which road Maximum Accident Occur due to Road Surface.	Road class and Road surface and Number of causality
Q4: Due to Lighting Problem on which road Maximum Accident occurs	Road class and Lighting condition and Number of causality
Q5: Probability of Accident at any location when drive is male or female.	Sex of causality and Age of Causality and Causality class

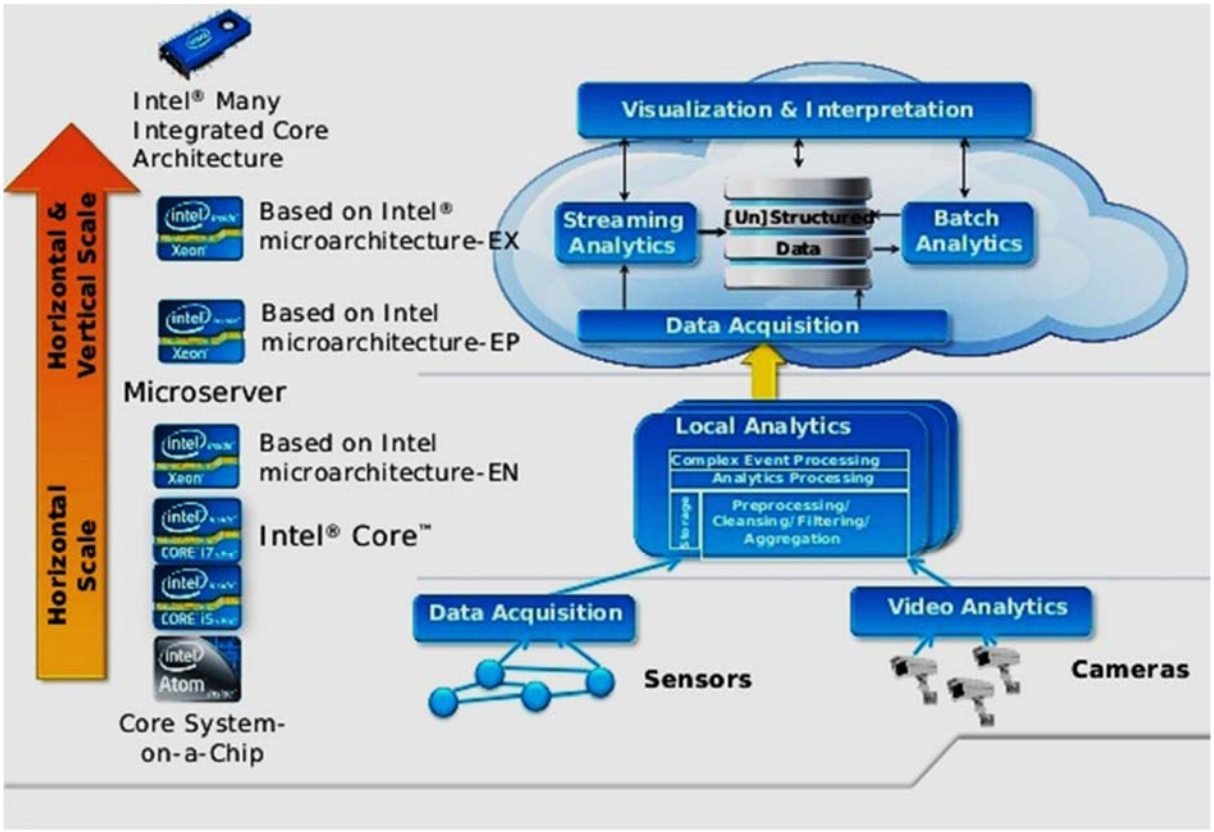

Figure 3. Proposed Model Framework.

The essential visualization layer is the establishment of pre-processing layer and data distributing layer, its fundamental capacity is to finish the work of different specialty units, and to create essential data acquisition models. It incorporates activity data gathering framework, flag control frameworks, video observation frameworks, , GPS vehicle area following framework, movement direction framework, vehicle data administration framework, driver data administration framework with maintaining accurate information.

IV. RESULTS AND DISCUSION

Examination like sort of vehicles (bike, auto, transport, lorry, jeep, truck, and so forth.) is done present rate dissemination of accidents on different criteria, speed point of confinement, and damage seriousness. Comparable investigation is done on other criteria, for example, conveyance of accidents by time of accidents and expired age, dispersion of accidents by month and climate amid the mishap, dissemination of accidents by softness and speed confine, circulation of accidents coincidentally type (human elements), appropriation of accidents by day of mischance and perished age, circulation of accidents by expired feelings.

TABLE II. TOP FACTORS FOR ROAD ACCIDENTS

Contributing factor	Percentage of accidents (%)
Rash driving	62.57
Object hit	26.67
Lane change	8.1

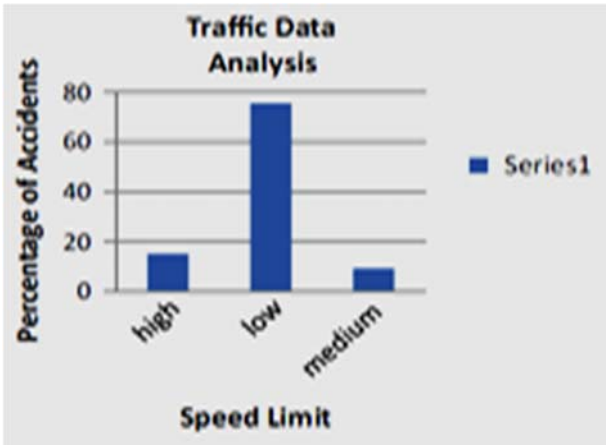

Figure 4. Accidents by speed limit.

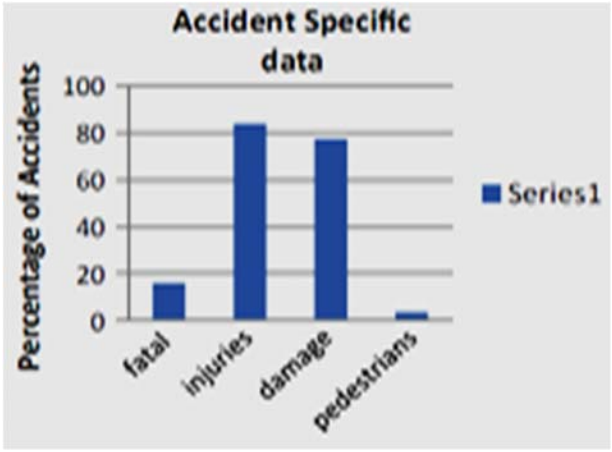

Figure 5. Accidents by injury severity.

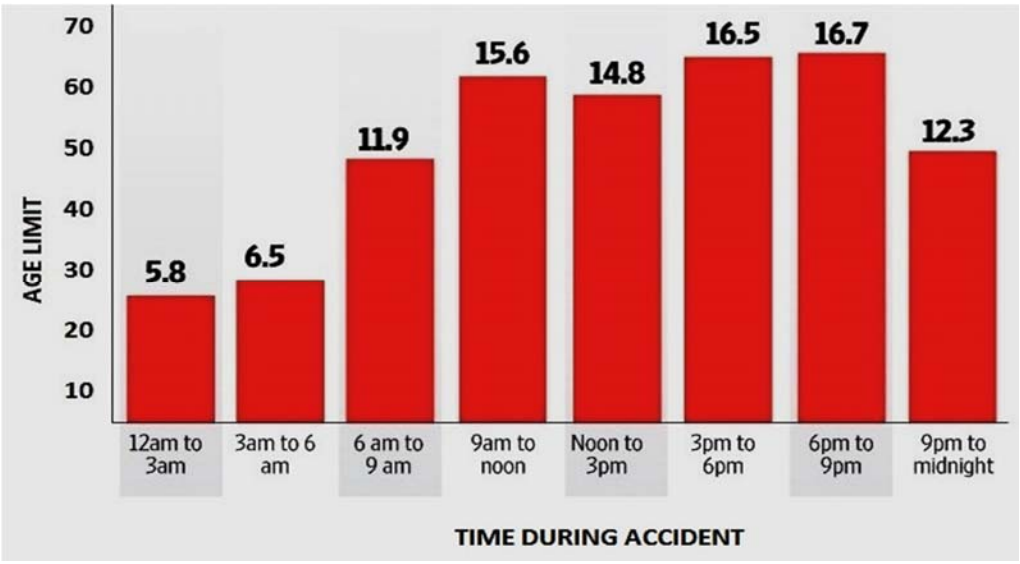

Figure 6. Final Road accident prediction based on Time and Age.

From figure 6, we can propose a solution for road accident prediction based on time on the day and the age of the driver who is driving a vehicle which meets with accident. Based on the above graph users are guided about the time in which majority accidents takes place and also warn about age limit for driving.

V. CONCLUSION

We proposed algorithms for road accident prediction and applied them to Indian road accidents data. The results can be used to recommend methods to national highway authorities to reduce road accidents. Our system utilizes big data analytics methods for accurate prediction. Initially pre-processing and clustering techniques are applied to the multi-dimensional data and then association rules are developed separately for vehicle types. We used the CCMF and TCAMP methods for automatic prediction by applying association rules to every parameter. We played several scenarios to conduct several tests on road accident data to estimate reasons for accidents and analyzed the data using the proposed methods. The results show that our proposed approaches are far better in predicting road accidents than other existing methods.

REFERENCES

[1] Seoung Hun Park, Young Guk Ha, "Large Imbalance Data Classification Based on MapReduce for Traffic Accident Predication" , IEEE International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, pp 45-49,2014.

[2] J. Conejero, P. Burnap, O. Rana, and J. Morgan, "Scaling Archived Social Media Data Analysis using a Hadoop Cloud", IEEE, 2013.

[3] S. G. Manikandan and S. Ravi, "Big Data Analysis Using Apache Hadoop," 2014 International Conference IT Converge. Security, , pp. 1–4, Oct 2014.

[4] J. Nandimath, "Big Data Analysis Using Apache Hadoop", pp. 700–703, 2013.

[5] S. Maitrey and C. K. Jha, "Handling Big Data Efficiently by Using Map Reduce Technique", IEEE Int. Conf. Comput. Intell. Commun. Technology, pp. 703–708, Feb. 2015.

[6] S. Humbetov, "Data-Intensive Computing with," 2012.

[7] L. P. Thompson and D. P. Miranker, "Fast Scalable Selection Algorithms for Large Scale Data," pp. 412–420, 2013.

[8] D. Chung, X. Rui, D. Min, and H. Yeo, "Road traffic big data accident analysis processing framework," 2013 7th Int. Conf. Appl. Inf. Commun. Technol., pp. 1–4, Oct. 2013.

[9] J. Shafer, S. Rixner, and A. L. Cox, "The Hadoop Distributed File system : Balancing Portability and Performance."

[10] M. Wang, S. B. Handurukande, and M. Nassar, "RPig : A Scalable Framework for Machine Learning and Advanced Statistical Functionalities", IEEE 4th International Conference on Cloud Computing Technology and Science ,2012, pp. 3–10, 2012.

[11] C. Chen, J. Hu, Q.Meng, and Y. Zhang, "SHORT-TIME TRAFFIC FLOW PREDICTION WITH ARIMA-GARCH MODEL," in Proc. IEEE Intell. Veh. Symp., 2011, pp. 607–612.

[12] M. Lippi, M. Bertini, and P. Frasconi, "SHORT-TERM TRAFFIC FLOW FORECASTING: AN EXPERIMENTAL COMPARISON OF TIME-SERIES ANALYSIS AND SUPERVISED LEARNING" IEEE Trans. Intell. Transp. Syst., vol. 14, no. 2, pp. 871–882, Jun. 2013.

[13] E. I. Vlahogianni, M. G. Karlaftis, J. C. Golias, "SHORT-TERM TRAFFIC FORECASTING: WHERE WE ARE AND WHERE WE'RE GOING" Transportation Research Part C: Emerging Technologies, vol. 43, pp. 3-9, 2014

[14] J. Rice and E. V. Zwet, "A simple and effective method for predicting travel times on freeways," IEEE Trans. Intell. Transp. Syst., vol. 5, no. 3, pp. 200–207, Sep. 2004.

[15] M. S. Bascil and F. Temurtas, "A study on hepatitis disease diagnosis using multilayer neural network with Levenberg Marquardt Training Algorithm," J. Med. Syst., vol. 35, no. 3, pp. 433–436, Oct. 2011.

[16] Gagandeep Kaur, Er. Harpreet Kaur: "Prediction of the cause of accident and accident prone location on roads using data mining techniques", 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE Conferences, 2017.

[17] Liling Li, Sharad Shrestha, Gongzhu Hu: "Analysis of road traffic fatal accidents using data mining techniques", 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA). IEEE Conferences, 2017.

[18] Rishi Sai Reddy Sudireddy; Uttam Mande: "Prediction of Road Accidents Using Correlation Based on Map Reducing", 2016 8th International Conference on Computational Intelligence and Communication Networks (CICN), IEEE Conferences, 2016.

[19] Nikhat Ikram, Shilpa Mahajan: "Road accidents: Overview of its causes, avoidance scheme and a new proposed technique for avoidance", 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), IEEE Conferences, 2016.

[20] Suwarna Gothane; M. V. Sarode: "Analyzing Factors, Construction of Dataset, Estimating Importance of Factor, and Generation of Association Rules for Indian Road Accident", 2016 IEEE 6th International Conference on Advanced Computing (IACC), IEEE Conferences, 2016.

[21] Sheeba Razzaq, Faisal Riaz; Tahir Mehmood; Naeem Iqbal Ratyal, "Multi-Factors Based Road Accident Prevention System", 2016 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube), IEEE Conferences, 2016.

[22] Sachin Kumar, Durga Toshniwal, "Analysing road accident data using association rule mining", 2015 International Conference on Computing, Communication and Security (ICCCS), IEEE Conferences, 2015.