**Chapter 2**

**Literature Review**

**2.1. Reducing Traffic Accidents by using Analysis:**

Statistical analysis of the traffic incident is highly expected to evaluate the reduction effect of the traffic accident. This study introduced Internet survey by delivering questionnaires to respondents. From the analysis of the collected answer, major psychological state of driver is hasty and distraction. As a first step this study focuses on driver's distraction, which causes severe traffic accidents. By using pattern recognition, the detection accuracy of driver's distraction is acquired. The reduction effect of the driver's distraction in the traffic accident is estimated by referring the reduction rate of both ASV (Advanced Safety Vehicle) and Intelligent Transportation Systems.[5]

**2.2. Using Partial Vehicle Trajectories approach:**

This approach describes traffic accident events at intersections in a human-understandable way using automated video processing techniques. This approach relies on learning normal traffic flow using trajectory clustering techniques, then analyzing accident events by observing partial vehicle trajectories and motion characteristics. In first phase, the model implements video preprocessing, vehicle detection and tracking in order to extract vehicle trajectories at road intersections. Second phase is to determine motion patterns by implementing trajectory analysis and then differentiating normal and abnormal events by defining descriptors, and last phase executes semantic decisions about traffic events and accident characteristics.[6]

**2.2. Naïve Bayes Implementation:**

Traffic accident casualty analysis is an important aspect in the traffic safety research field. Based on data survey and statistical analysis, a Bayesian network for traffic accident causality analysis is developed**.** The results show that the Bayesian network can express the complicated relationship between the traffic accident and the causes, as well the correlations among the factors of causes. The results of analysis provide the valuable information on how to reveal the traffic accident causality mechanisms. Naïve Bayes Classifier uses Bayesian theorem with strong independent assumptions. In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. [1][2].

Abstractly, the probability model for a classifier is a conditional model [[7]](http://en.wikipedia.org/wiki/Naive_Bayes_classifier#cite_note-7)

p(C \vert F_1,\dots,F_n)\,

over a dependent class variable C with a small number of outcomes or *classes*, conditional on several feature variables F1through Fn. The problem is that if the number of features n is large or if a feature can take on a large number of values, then basing such a model on probability tables is infeasible. We therefore reformulate the model to make it more tractable. [4].

Using Bayes theorem, this can be written

p(C \vert F_1,\dots,F_n) = \frac{p(C) \ p(F_1,\dots,F_n\vert C)}{p(F_1,\dots,F_n)}. \,

………………Equation (I)

**2.3. Random Forest Algorithm Implementation:**

In order to determine the major factors contributing to traffic accidents and their severity, the data mining model uses Random Forest tree algorithms to analyze the traffic collision data. The experiment results from this study shows that the developed data mining model using decision trees can effectively classify the probability of an accident occurring as high, low or medium.[3]

Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest). Each tree is grown as follows:

* If the number of cases in the training set is N, sample N cases at random - but with replacement, from the original data. This sample will be the training set for growing the tree.
* If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
* Each tree is grown to the largest extent possible. There is no pruning.[1]

**2.4. Advantages of Naïve Bayes:**

It is fast and highly scalable model building and scoring. An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the necessary for classification. Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood.[1]. In spite of their naive design and apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. In 2004, analysis of the Bayesian classification problem has shown that there are some theoretical reasons for the apparently unreasonable efficacy of naive Bayes classifiers.[4]

**2.5. Disadvantages of Naïve Bayes:**

Even though Naïve Bayes has good efficiency,a comprehensive comparison with other classification methods in 2006 showed that Bayes classification is outperformed by more current approaches, such as boosted trees or random forests.[4]. It does not estimate any missing data. It is not suitable to work with Naïve Bayes algorithm on large datasets. [3].

**2.6. Advantages of Random Forest:**

Random Forest is unexcelled in accuracy among current algorithms. It runs efficiently on large data bases and it can handle thousands of input variables without variable deletion. It gives estimates of what variables are important in the classification. It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing. Generated forests can be saved for future use on other data. Random Forest can be used to perform unsupervised clustering, data views and outlier detection.

**2.7. Disadvantages of Random Forest:**

The major disadvantage of Random Forest is its complexity. The learners can create over-complex trees that do not generalize well from the training data. This is known as over-fitting. Mechanisms such as pruning are necessary to avoid this problem of over-fitting.[1]. There are concepts that are hard to learn because decision trees do not express them easily. In such cases, the decision tree becomes prohibitively large. Approaches to solve the problem involve either changing the representation of the problem domain or using learning algorithms based on more expressive representations such as statistical relational learning or inductive logic programming. [3]

**References**

1. “A decision tree approach for traffic accident analysis of saskatchewan highways” By Xue-Fei Zhang ; Fan, L. published in Electrical and Computer Engineering (CCECE), 2013 26th Annual IEEE Canadian Conference.
2. “Accident cause analysis method based on traffic accident information system “ By Xi Jianfeng Chen Xiaodong ; Wang Shuangwei ; Zhurong Tao published in Computer Application and System Modeling (ICCASM), 2010 International Conference.
3. “Bayesian Network-Based Road Traffic Accident Causality Analysis” By Xu Hongguo ; Zhang Huiyong ; Zong Fang published in Information Engineering (ICIE), 2010 WASE International.
4. “A Perspective Analysis of Traffic Accident using Data Mining Techniques” By S.Krishnaveni;Ph.D (CS) Research Scholar published in International Journal of Computer Applications (0975 – 8887) Volume 23– No.7.
5. **Chapter 3**

**System Analysis**

**3.1 Functional Requirements**

• Acquiring Dataset: The dataset containing accident records is obtained from the internet in the form of statistics. This statistical data is converted into synthetic data. The Dataset contains attributes such as weather, type of vehicle, casualty age, casualty sex, year of manufacture, severity of accident and location of the accident.

• Performing Data Mining: From the data available in the database, information is extracted by the system. This is executed by performing Data Mining using Naïve Bayes and Random Forest Algorithm. An analysis of both algorithms is done based on specific parameters after performing Data Mining.

• Displaying output: The output will contain the prediction of the chances of the accident happening as high, low or moderate. The output will also contain a comparison of Naïve Bayes and Random Forest Algorithm based on the time to develop the classifier, the time to retrieve the records and the accuracy of each algorithm.

**3.2 Non-Functional Requirements**

**3.2.1 Performance:**

There should be no over-fitting of data. The product should perform the comparison of the Naïve Bayes and Random Forest Algorithms..

**3.2.2 Reliability:**

The product should not crash under any circumstances such as user entering invalid values, user trying to load unsupported files etc. It should show appropriate message for every user generated message.

**3.2.3 Portability:**

The product should be portable and should run on most personal computers or machines which support windows operating systems.

**3.3 Specific Requirement:**

• Microsoft Windows (XP, Windows 7 , Windows 8).

• Core i3 processor.

• 2 GB RAM.

• Java Language for coding.

• SQL for database.

• All platforms of Microsoft Windows.

• Hard Disk with at least 250GB.

• Windows Compatible Mouse and Keyboard.