A REVIEW ON ROAD ACCIDENT DETECTION USING DATA MINING TECHNIQUES

Arun Prasath N
Ph.D Research Scholar
Department of Computer Applications
Bharathiar University
Coimbatore, India

Dr. M. Punithavalli
Associate Professor
Department of Computer Applications
Bharathiar University
Coimbatore, India

Abstract: Transportation has evolved greatly over time. With modern technology, the automobile industry has obtained new heights with respect to comfort, speed, efficiency and security. Despite improvement in technology, there has been increase in the rate of accidents. A large number of precious lives are lost because of road traffic accidents every day. The common reason behind road accident is driver’s mistake. It is essential to have effective road accident detection mechanism to save life. Data mining techniques are widely used for road accident detection. The main focus of this survey is to provide an overview of the literature in road accident detection with various techniques and approaches implemented in them, their merits and demerits etc. Comparison based on parameters is also done to prove the efficiency of the various road detection techniques and approaches. The comparison result shows the best road accident detection method.

Keywords: Road accident detection; data mining; accident prediction; road safety; transportation.

I. INTRODUCTION

Nowadays due to road accidents [1] a large number of lives are lost. From an analysis it has been estimated that for every year over 3,000,000 persons die and 10 to 15 million people are injured due to road accidents in the entire world. The accidents are classified into the following types are [2]:

- **Fatality**: an accident or incident resulting in a fatality either immediately.
- **Major injury Accidents**: are accidents which results in a significant injury, damage or loss.
- **Minor accidents**: are accidents which result in an injury, damage or loss but do not cause significant harm to a person.
- **Lost time accidents**: are accidents which result in an employee being absent from work for more than half day.
- **Never miss incidents**: result in no apparent damage or injury.
- **Dangerous Occurrences**: are specific incidents as defined by Reporting of Injuries, Diseases and Dangerous Occurrences Regulations.

Data Mining [3] has been proven as a consistent technique to analysis road accidents and it also provide productive results. Most of the road accident data analysis are done by the data mining processes such as feature selection, clustering and classification to identify factors that affect the severity of an accident. In this paper, various data mining techniques for detection of road accidents are analyzed based on their merits and demerits and compared in terms of accuracy, precision, recall and F-measure.

II. RESEARCH METHODOLOGY

The road accidents were detected by using deep learning algorithm [4]. The paired tokens in the collected three million tweets captured the association rules which improve the traffic accident detection accuracy. Then, Deep Belief Network (DBN) and Long Short-Term Memory (LSTM) were applied on the extracted tokens. The results of these learning algorithms are given as input to the classification algorithms supervises Latent Dirichlet Allocation (sLDA) and Support Vector Machine (SVM) which detect the traffic accidents.

A framework [5] was proposed to analysis the road accidents using the rule mining approach. A raw data was collected from Emergency Management research Institute (EMRI). It serves and keeps track of every accident record on every type of road. The collected data are converted into structured format by applying filtering techniques. Then the accident data were clustered by applying the hybrid clustering based on enhanced K-means clustering. It clusters the accident data by splitting the input array into sub arrays based on the distance between the elements in clusters. Finally, association rule mining was applied to identify the circumstance in which an accident may occur for each cluster. The outcome of this technique was utilized to take some accident prevention efforts in the areas identified for different categories of accidents in a way to reduce the number of accidents.

Some classification techniques [6] were used to predict the severity of injury occurred during traffic accidents. The classification algorithms such as Random Forest, AdaBoostM1, Naive Bayes, J48 and PART were investigated and compared these algorithms performance based on injury severity. It includes labels of severity, road classification, district council district, rain, hit and run, type of collision, natural light, number of vehicles involved degree of injury, number of causalities injured, pedestrian action, casualty sex, casualty age, vehicle class of driver or passenger casualty, year manufacture, role of casualty, driver age, driver sex, vehicle class and severity of accident. The three classes of severity of injury are based on casualty, based on accident and based on vehicle.

For accident detection, a preliminary real time autonomous accident detection system [7] was proposed. For the accident detection, data were collected from the sensors and it was integrated with the event log to extract the most discriminative features. It extracted features such as average velocity difference between reading at time T and T+1, weekday or weekend, average capacity usage difference.
between reading at time T and T+1, Average occupancy difference between reading at time T and T+1, occurrence of accident or event at rush hours. These features were fed into a regression tree, neighbor model and feed forward neural network model. It predicts the possibility of occurrence of an accident.

Data mining algorithms [8] were introduced for classification of vehicle collision patterns in road accidents. It derived the classification rules which can be utilized for prediction of vehicle collision patterns. Initially the training set was taken and then noisy, inconsistent and incomplete data were removed by applying data cleaning process. The preprocessed data is converted into an appropriate form for mining. Then the attribute space of a feature set was reduced for classification of vehicle collision. It can be achieved by applying the feature selection algorithms such as Multi valued Oblivious Decision Tree (MODTTree) filtering, feature ranking, Correlation based Feature Selection (CFS), Mutual Information Feature Selector (MIFS) and Fast Correlation Based Filter (FCBF) algorithm. The selected features are used in different classification algorithms namely Naïve Bayes, C4.5, Classification and Regression Trees (C&RT), RndTree, Decision List, rule induction and random tree.

A Multi-class Support Vector Machine [9] was introduced to predict causes of traffic road accidents. A real time data is collected from police department in Dubai. Then, a typical data mining framework was applied on the collected data. The framework consists of three steps namely pre-processing, mining patterns and post processing. In the pre-processing step the data gets into the processes of data cleaning, deals with unknown and missing data, feature selection and also it take stock of unbalanced data. The format of the data was converted into such a form which can be accepted by SVM. Finally, in the post processing step Multi-SVM was applied which predict the causes of traffic road accidents.

A new method [10] was proposed to detect the road accidents based on temporal data mining. This method employed ternary numbers time series model was constructed that reflected the state of the traffic flow based on cell transmission model. The computational cost and the linear drift between time series were handled by Discrete Fourier transform. It transformed the time domain data into frequency domain data. Then Euclidean distance was calculated for transformed time series data and based on this measure accident was detected.

In order to analysis and predict the nature of road accident a method [11] was proposed based on data mining techniques. Here, Random Forest, Naïve Bayes and J48 algorithm were chosen to analysis road accident data in the state of Maharashtra. Finally, the Apriori association rule mining algorithm was applied to determine the relationship between independent variables with respect to the nature of accidents.

For analysis of traffic accident, Artificial Neural Network and Decision trees techniques [12] were employed. For the analysis of traffic accident, the data were collected from one of the busiest roads of Nigeria. The collected data was arranged into categorical and continuous data. The categorical data of road accident were analyzed by using Decision tree technique. Artificial Neural Network was applied on the continuous data of accidents.

The road accident data were analyzed by using proposed data mining framework [13]. In this framework, K-modes clustering was used as a preliminary task to segment the road accident data. By applying the association rule mining technique, the various circumstances which were associated with the occurrence of the accident were identified. It was identified for both the dataset and the clusters were identified by introducing K-modes clustering algorithm. Then the results of cluster based analysis and dataset analysis were compared and it was captured from the analysis that was the combination of association rule mining and k modes clustering was producing crucial information effectively.

A data mining approach [14] was proposed to analyse road accidents in India. The intention of this approach was to create a model which sort out the heterogeneity of the data by grouping the similar objects together to find the accident prone areas in the country with respect to different accident factors. This was also used to determine the association between these factors and casualties. To group the similar objects of the heterogeneous data, K means clustering was employed. In K-means clustering, K was chosen randomly which is considered as initial centroids. Then, Euclidean distance between each data point and the centroids is calculated. The changes in the centroids are based on the Euclidean distance. This was continued until there is no change in the centroids. Finally, the decision tree classification was applied to analysis the road accidents.

For automatic road detection, a novel approach [15] was proposed. The novel approach was based on detection of damage vehicles from the collected footage from surveillance cameras. It observed the occurrence of road accident. A new supervised learning method with three stages was proposed for road accident detection. These three stages were comprised into a single framework in a serial manner. These were used five Support Vector Machine trained with Histogram Of Gradient (HOG) and Gray Level Co-occurrence features. The supervised learning was worked as a binary classifier which distinguished the data containing a damaged car as class 1 and data not containing damaged car as class 2.

### A. Comparison of Research Methodologies

The road accident detection methods described in the above section is analyzed and compared based on methods used, their merits, demerits and the parameters used in experimental results. The comparison is given in Table I.

In Table I, the different methods for road accident detection are analyzed based on accuracy, precision, recall and F-measure. The Preliminary real time autonomous accident detection system [7] has better accuracy of 99.79% than other methods, Naïve Bayes, J48, Random Forest algorithm, Apriori association rule mining [11] method has better precision of 0.983 than other methods, Naïve Bayes, J48, Random Forest algorithm, Apriori association rule mining [11] method has better recall of 98.3 than other methods and Naïve Bayes, J48, Random Forest algorithm, Apriori association rule mining [11] method has better F-measure of 98.3 than other methods.
| Ref No. | Methods                                                                 | Merits                        | Demerits                                          | Performance Metrics                        |
|--------|-------------------------------------------------------------------------|-------------------------------|--------------------------------------------------|--------------------------------------------|
| [4]    | Long Short-Term Memory, Deep Belief Network, supervised Latent Dirichlet Allocation, Support Vector Machine | Better accuracy              | Noisy and unreliable social data affect the performance | Accuracy  
1. DBN = 0.856  
2. ANN = 0.842  
3. LSTM = 0.835  
4. SVM = 0.791  
5. sLDA = 0.758  
**Precision**  
1. DSM = 0.93  
2. ANN = 0.824  
3. LSTM = 0.862  
4. SVM = 0.834  
5. sLDA = 0.949  
**Execution Time**  
1. K-modes clustering (k=2) = 81.37 ms  
2. Hybrid Clustering (k=2) = 87.2961 ms |
| [5]    | Hybrid Clustering, association rule mining                               | Better Prediction             | Less efficiency                                   | Accuracy  
1. K-modes clustering (k=2) = 81.37  
2. Hybrid Clustering (k=2) = 87.2961 |
| [6]    | Random Forest, AdaBoostM1, Naïve Bayes, J48, PART                       | Injury severity is predicted based on three different classes | Low accuracy                                    | Accuracy  
1. Naïve Bayes = 68.44%  
2. J48 = 70.86%  
3. AdaBoostM1 = 63.03%  
4. PART = 71.28%  
5. Random Forest = 74.34%  
**Execution Time**  
1. K-modes clustering (no. of clusters =6) = 160 ms  
2. Hybrid Clustering (no. of clusters =6) = 110 ms |
| [7]    | Preliminary real time autonomous accident detection system               | Provide useful information    | False alarms is considerably high               | Accuracy = 99.79%                          |
| [8]    | Naïve Bayes, C4.5, C&RT, RndTree, Decision List, rule induction, random tree | High accurate results        | More features leads to more complexity in classifiers | Accuracy  
1. C4.5 = 80.59%  
2. C&RT = 76.24%  
3. CS-MS4 = 71.09%  
4. Decision List = 67.92%  
5. ID3 = 75.54%  
6. Naïve Bayes = 72.28%  
7. RndTree = 94.38%  
8. Rule Induction = 75.54%  
**Precision**  
1. J48 = 0.983  
2. Random Forest = 0.975  
**Recall**  
1. J48 = 0.983  
2. Random Forest = 0.975  
**F-Measure**  
1. J48 = 0.983  
2. Random Forest = 0.975 |
| [9]    | Multi-class Support Vector Machine                                       | Identify the cause of road traffic accidents in the absence of eyewitnesses | Accuracy of the developed model is just acceptable | Accuracy = 75.395  
**Precision**  
1. J48 = 0.767  
2. Random Forest = 0.675  
**Recall**  
1. J48 = 0.767  
2. Random Forest = 0.675  
**F-Measure**  
1. J48 = 0.767  
2. Random Forest = 0.675 |
| [10]   | Temporal data mining                                                     | Highly effective              | Euclidean distance performs well only when the dataset include isolated or compact clusters | Nil |
| [11]   | Naïve Bayes, J48, Random Forest algorithm, Apriori association rule mining | High accuracy                 | Multiple scans of Apriori takes long time        | Accuracy  
1. J48 = 98.3%  
2. Random Forest = 97.5%  
**Precision**  
1. J48 = 0.983  
2. Random Forest = 0.975  
**Recall**  
1. J48 = 0.983  
2. Random Forest = 0.975  
**F-Measure**  
1. J48 = 0.983  
2. Random Forest = 0.975 |
| [12]   | Decision Trees, Neural Networks                                         | Low error rate, High accuracy rate | Low efficiency                                   | Accuracy  
1. Decision Tree = 0.777  
2. Neural Network = 0.547  
**Precision**  
1. Decision Tree = 0.78  
2. Neural Network = 0.523  
**Recall**  
1. Decision Tree = 0.777  
2. Neural Network = 0.547  
**F-Measure**  
1. Decision Tree = 0.769  
2. Neural Network = 0.524 |
| [13]   | Data Mining framework                                                   | Produces important information | Limited capacity to discover new and unanticipated patterns | Nil |
| [14]   | K-means clustering, Decision tree                                       | Better prediction             | Failed to determine the accident frequency       | Recall = 94.44%  
**Precision** = 73.91% |
III. PERFORMANCE EVALUATION

The performance of the efficient methodologies in the literature are analyzed and compared among them to determine the comparative performance efficiency. The methods considered for analysis are Long Short-Term Memory, Deep Belief Network, supervised Latent Dirichlet Allocation, Support Vector Machine [4], Hybrid Clustering, association rule mining [5], Random Forest, AdaBoostM1, Naïve Bayes, J48, PART [6], Preliminary real time autonomous accident detection system [7], Naïve Bayes, C4.5, C&RRT, RndTree, Decision List, rule induction, random tree [8], Multi-class Support Vector Machine [9], Naïve Bayes, J48, Random Forest algorithm, Apriori association rule mining [11], Decision Trees, Neural Networks [12], K means clustering, Decision tree [14] and Support Vector Machines [15]. The comparison is done by the experimental results of the methods in terms of accuracy, precision, recall and F-measure.

A. Accuracy

Accuracy is described as the closeness of a measurement to the true value. It is given as

\[
\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}}
\]

Fig. 1, shows the comparison of accuracy between Long Short-Term Memory, Deep Belief Network, supervised Latent Dirichlet Allocation, Support Vector Machine [4], Hybrid Clustering, association rule mining [5], Random Forest, AdaBoostM1, Naïve Bayes, J48, PART [6], Preliminary real time autonomous accident detection system [7], Naïve Bayes, C4.5, C&RRT, RndTree, Decision List, rule induction, random tree [8], Multi-class Support Vector Machine [9], Naïve Bayes, J48, Random Forest algorithm, Apriori association rule mining [11], Decision Trees, Neural Networks [12], K means clustering, Decision tree [14] and Support Vector Machines [15]. X axis denotes the methods and Y axis denotes the accuracy in %. The graph clearly shows that the preliminary real time autonomous accident detection system [7] has high accuracy than other methods.

B. Precision

Precision is the evaluated according to the road accident prediction at true positive and false positive prediction.

\[
\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}
\]

Fig. 2, shows the comparison of precision between Long Short-Term Memory, Deep Belief Network, supervised Latent Dirichlet Allocation, Support Vector Machine [4], Multi-class Support Vector Machine [9], Naïve Bayes, J48, Random Forest algorithm, Apriori association rule mining [11], Decision Trees, Neural Networks [12], K means clustering, Decision tree [14] and Support Vector Machines [15]. X axis denotes the methods and Y axis denotes the precision. The graph clearly shows the J48 [11.1] method has high precision than other methods.

C. Recall

Recall is evaluated according to the classification of data at true positive and false negative predictions.

\[
\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}
\]

Fig. 3, shows the comparison of recall between Multi-class Support Vector Machine [9], Naïve Bayes, J48, Random Forest algorithm, Apriori association rule mining [11], Decision Trees, Neural Networks [12], K means clustering, Decision tree [14] and Support Vector Machines [15]. X axis denotes the methods and Y axis denotes the recall. The graph clearly shows the J48 [11.1] method has high recall than other methods.
deriving the motivation for our future research work.

precision, recall and F-measure. This survey also helps in

association rule mining [11] method was efficient in terms of

and Naïve Bayes, J48, Random Forest algorithm, Apriori
detection system [7] method was efficient in terms of accuracy

concludes that the preliminary real time autonomous accident

accuracy, precision, recall and f-measure. The survey

comparison of the efficient techniques is carried out in terms of

novel ideas. The analysis of these methods provides better

understanding of the steps involved in each process in a way of

increasing the scope for finding the efficient

Road accident detection is considered to be the

contemporary ever growing process focused primarily to

reduce death. Here this paper provides the recent developments

in the road accident detection techniques by analyzing the

novel ideas. The analysis of these methods provides better

understanding of the steps involved in each process in a way of
connsequently increasing the scope for finding the efficient

techniques to achieve maximum accurate performance. The

comparison of the efficient techniques is carried out in terms of

accuracy, precision, recall and f-measure. The survey concludes that the preliminary real time autonomous accident
detection system [7] method was efficient in terms of accuracy and Naïve Bayes, J48, Random Forest algorithm, Apriori
association rule mining [11] method was efficient in terms of precision, recall and F-measure. This survey also helps in
deriving the motivation for our future research work.

D. F-measure

F-measure is an accuracy testing score considering both the

precision and recall.

\[
F - \text{measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

Road accident detection is considered to be the

recent developments in the road accident detection techniques by analyzing the

novel ideas. The analysis of these methods provides better

understanding of the steps involved in each process in a way of

consequently increasing the scope for finding the efficient

techniques to achieve maximum accurate performance. The

comparison of the efficient techniques is carried out in terms of

accuracy, precision, recall and f-measure. The survey concludes that the preliminary real time autonomous accident
detection system [7] method was efficient in terms of accuracy and Naïve Bayes, J48, Random Forest algorithm, Apriori
association rule mining [11] method was efficient in terms of precision, recall and F-measure. This survey also helps in
deriving the motivation for our future research work.

IV. CONCLUSION

Fig. 4, shows the comparison of F-measure between Multi-
class Support Vector Machine [9], Naïve Bayes, J48, Random
Forest algorithm, Apriori association rule mining [11] and
Decision Trees, Neural Networks [12]. X axis denotes the
methods and Y axis denotes the F-measure. The graph clearly
shows the J48 [11.1] has high f-measure than other methods.

V. REFERENCES

[1] D. T. Akomolafe and A. Olatayo, “Using Data Mining
Technique to Predict Cause of Accident and Accident Prone
Locations on Highways”, American Journal of Database Theory
and Application, vol. 1, no. 3, 2012, pp. 26-38.

[2] B. Khatri and H. Patidar, “Road Traffic Accidents with Data
Mining Techniques”, International Journal of Information
Engineering and Technology (IJIET), vol. 2, no. 1, 2016, pp. 1-
6.

[3] S. Kumar and D. Toshniwal, “A data mining approach to
characterize road accident locations”, Journal of Modern
Transportation, vol. 24, no. 1, 2016, pp. 62-72.

[4] Z. Zhang, Q. He, J. Gao, and M. Ni, “A deep learning approach for
detecting traffic accidents from social media data” Transportation Research Part C: Emerging Technologies, vol. 86, 2018, pp. 580-596.

[5] G. Karur and G. Gandhi, “A framework for analyzing the road
accidents in Data Mining using Rule Mining”, International
Journal of Innovative Research in Computer Science and
Communication Engineering, vol. 5, no. 4, 2017, pp. 6931-6939.

[6] S. Krishnaveni and M. Hemalatha, “A perspective analysis of
traffic accident using data mining techniques”, International
Journal of Computer Applications, vol. 23, no. 7, 2011, pp. 40-
48.

[7] M. Ozbayoglu, G. Kucukayan, and E. Dogdu, “A real-time
autonomous highway accident detection model based on big data
processing and computational intelligence”, Big Data (Big Data), 2016 IEEE International Conference on IEEE, 2016, pp. 1807-1813.

[8] S. Shanthi and R. G. Ramani, “Classification of vehicle collision
patterns in road accidents using data mining algorithms”,
International Journal of Computer Applications, vol. 35, no. 12,
2011, pp. 30-37.

[9] E. A. Mohamed, “Predicting Causes of Traffic Road Accidents
Using Multi-class Support Vector Machines”, In Steering
Committee of The World Congress in Computer Science,
Computer Engineering and Applied Computing (WorldComp),
2014, pp. 441-447.

[10] S. An, T. Zhang, X. Zhang, and J. Wang, “Unrecorded accidents
detection on highways based on temporal data mining” Mathematical Problems in Engineering, 2014.

[11] B. Atnafu and G. Kaur, “Analysis and predict the nature of road
traffic accident using data mining techniques in Maharashtra
India”, International Journal of Engineering Technology Science
and Research (IJETSR), vol. 4, no. 10, 2017, pp. 1153-1162.

[12] V. A. Olutayo and A. A. Eludire, “Traffic accident analysis
decision trees and neural networks”, International Journal
of Information Technology and Computer Science, vol. 2, 2014,
pp. 22-28.

[13] S. Kumar and D. Toshniwal, “A data mining framework to
analyze road accident data” Journal of Big Data, vol. 2, no. 1,
2015, pp. 1-18.

[14] A. Jain, G. Ahuja, and D. Mehrotra, “Data mining approach to
analyse the road accidents in India”, In Reliability, Infocom
Technologies and Optimization (Trends and Future
Directions)(ICRITO), 2016 5th International Conference on IEEE, 2016, pp. 1153-1162.

[15] V. Ravindran, L. Viswanathan, and S. Rangaswamy, “A Novel
Approach to Automatic Road-Accident Detection using Machine
Vision Techniques”, International Journal Of Advanced
Computer Science And Applications, vol. 7, no. 11, 2016, pp. 235-242.