Using hybrid Data Mining algorithm for Analysing road accidents Data Set

Dr.G.Parathasarathy
Department of CSE,
Jeppiaar Maamallan Engineering College,
Anna university ,Chennai India.
amburgrps@gmail.com

T.R Soumya
Department of CSE,
Jeppiaar Maamallan Engineering College,
Anna university ,Chennai India.
soumyatr.soumya@gmail.com

Y.Justin Das
Department of CSE,
Jeppiaar Maamallan Engineering College,
Anna university ,Chennai India.
justindhasy@gmail.com

J.Saravanakumar
Department of CSE,
Jeppiaar Maamallan Engineering College,
Anna university ,Chennai India..
saravanakumar2005@gmail.com

A.Anigo Merjora
Department of CSE,
Jeppiaar Maamallan Engineering College,
Anna university ,Chennai India.
anigomerjora@gmail.com

Abstract— Nowadays, road safety has become an important issue in the urban areas due to the high vehicle density. Road safety can be improved by reducing the accidents. Road accident causes traffic hindrance which has become intolerable especially in big-cities. Therefore, analyzing the road accidents accurately can help to solve the problem of traffic crashes. In our project, we propose a hybrid model that combines both K-Nearest Neighbor and Support Vector Machines algorithm for road accident analysis and prediction of accident type, which is based on the hierarchical-learning approach. The accident types are classified as crash, drunk & drive, fire and skid. Our proposed model uses the combination of both KNN and SVM algorithms with the historical datasets collected from UCI Repository. This analyzed data will be more useful to suggest better safety measures to avoid traffic crashes. We experimentally analyze the performance of both KNN and SVM algorithms using R programming with large accident datasets. Results show that our hybrid model enhances the accuracy of road accident analysis.

***Keywords:*clustering model,hierarchical learning,k nearest neighbour,machine learning,Road accidents,support vector machine**

I. .INTRODUCTION

Machine Learning is one of the applications of Artificial Intelligence technique where the machine learns the data implicitly rather than explicit programming. Nowadays, machine learning plays a crucial role in our day to day life. It is used almost in every field like transport, medical, banking etc. in which transport and medical field has more importance than others as they are related to lives. In the field of transportation, machine learning can be applied in many sectors like traffic flow prediction, accident prediction, tourist place suggestion etc. Machine learning is used not only for automation but also for safety. Road Safety is a big issue as traffic congestion increases in urban as well as rural areas [8]. It could be enhanced by decreasing the traffic crashes. Crashes occurs especially during peak hours which may leads to traffic jam as well as loss of social costs [4]. Therefore analyzing the traffic crashes accurately can help to solve the problem of traffic crashes thereby enhancing the road safety [3].

The prime objective of our work is to help the transportation system and the police to classify the accident types by accurately analyzing the factors responsible for road accidents and predicting the type of accidents. It also used to prevent accidents by creating awareness to the civilians. For the accident type prediction we are introducing a hybrid model that combines two classifiers i.e. K-Nearest Neighbor and Support Vector Machine. A hybrid model is most often used to enhance the accuracy of the process [9]. We also used it for the better performance of the classifiers. In our model the type of accident is the target variable which is to be predicted by the classifier. As it is a machine learning approach we initially train the classifier with some sample dataset contains the target variable. This sample dataset is collected from UCI Repository which contains lot of datasets in all domains for data analytics. The classifier learns the dataset by analyzing the factors responsible for accident and the type of accidents. The types of accidents are classified into four categories i.e. Crash, Drunk & Drive, Fire and Skid. Our target is to predict the type of accidents which belongs to any one of the above mentioned categories.

Lot of previous works focuses mainly on clustering and association rule mining for accident cause prediction[9];[8]; [10]. Accident cause prediction can help the civilians and transportation department to improve the safety measures thereby reducing the accidents [5]. As more researches were done in accident cause prediction, we concentrate on classification and regression for accident type prediction which will help the police and the transportation department to classify the accident type after the accident happens. In rural areas there may be no surveillance camera. So it is

difficult for the police to investigate on the accident to find its type. It is more complicated to identify the type of accident than other specifications like road condition, light condition and weather condition etc. By analyzing the existing accident dataset we can identify the accident type. To make this analysis more accurate we are using the hybrid model of hierarchical learning algorithm that combines both KNN and SVM classifier.

II. THE MATERIAL AND METHOD

In this article analyzed about data repository and challenges in databases [1]. In this article both classification and clustering algorithms in their work. They proposed a hybrid model that combines random forest and support vector machines algorithm for classification and also k-means for clustering. They used the dataset collected from the expressways on Shanghai and achieved the accuracy of 78% [4]. In this article used k-means clustering and association rule mining by apriori algorithm for accidents pattern analysis. Their accident dataset is collected from the Maharashtra road networks on the year of 2015-2016 [9]. In Proposed model that uses association rule mining by apriori algorithm, clustering by k-means algorithm and classification by Naïve Bayes algorithm. They applied this model on FARS fatal accident dataset to identify the factors that are closely related to fatal accidents [5].

In proposed system, uses apriori algorithm for association rule mining which helps to discover the patterns of road accident. They also used naïve bayes classifier to predict the accidents for new roads. They might use the sample dataset for their model [7]. Proposed a new model called Self Organizing Map (SOM) which is based on neural network to predict the reasons for accident. They compared their work with k-means clustering algorithm and concluded that their work is better than k-means. They used the dataset collect from US government accident data which is freely available [8].

In this paper used k-modes clustering model and association rule mining for analyzing different factors for different types of accident. They applied their work on the accident dataset collected from Dehradun roadways on the year of 2009-2014 [10]. By analyzing these related works, we conclude that most of the accident analysis is based on data mining approach. Each and every researches uses the same algorithm but with different implementation for better accuracy. Accuracy plays a major role in analysis as it helps for solving many problems with better solutions. To enhance this accuracy many researchers did a lot of work with different approaches. All the existing system are used to predict the causes for the accident i.e. the factors responsible for accident [4,5,8]. But in our model we are proposing a hybrid model to predict the type of the accident by analyzing the causes for accident. A brief summary of the related work is given in Table 1.

TABLE 1
SUMMARY OF RELATED WORK

S. no	Reference no.	Analysis	Algorithms used	Dataset used
1	[4]	Real time crash prediction on urban expressways: identification of key variables and a hybrid support vector machines.	Random Forest, K-Means, Support Vector Machines	Collected from expressways on Shanghai.
2	[9]	Analyzing road accident data using machine learning paradigms.	K-Means Clustering, Apriori rule mining.	Maharashtra road networks on 2015-2016
3	[5]	Analysis of road traffic fatal accidents using data mining techniques	Apriori, K-Means Clustering, Naïve Bayes Classifier.	FARS fatal accident dataset.
4	[7]	Analysis of road accidents using data mining techniques	Apriori algorithm, Naïve Bayes.	Sample dataset.
5	[8]	A review on road accident data analysis using data mining techniques	Self-Organizing Map (SOM)	US government dataset.
6	[10]	A data mining framework to analyze road accident data	K-Modes Clustering, Association rule mining	Dehradun Accident dataset on 2006-2014

III. PROPOSED SYSTEM

Our proposed system is a hybrid model of machine learning algorithm to analyze the road accident and predict the accident type. It combines K-Nearest Neighbor Classifier and Support Vector Machines Classifier. These two classifiers are used to enhance the accuracy of analysis. The working of our proposed system is clearly explained in this section with a neat architecture diagram shown in Fig. 1.

A. Problem Statement

Classification and regression of road accident analysis can be done by a single classifier. Either KNN or SVM is alone to classify and predict the target variable "accident type". But we are using both of them together in our approach to make it as hybrid. Classification using a hybrid model can give better accuracy than others. So we proposed a hybrid model for classification and regression of road accident analysis and accident type prediction.

B. Methodology

Classification and Regression comes under supervised machine learning where the machine learns the data implicitly with some training data containing the target variable. Here we use the KNN and SVM classifiers for this purpose.

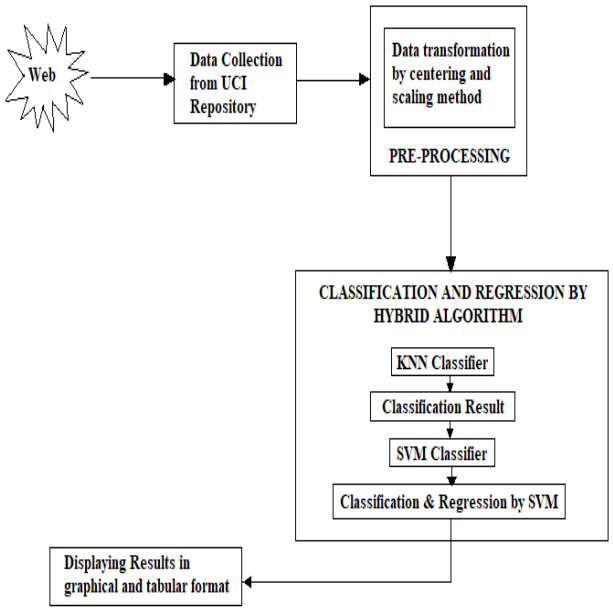

Fig. 1. Architecture Diagram

1) Need for Two Classifiers:

In our method we are using two classifiers i.e. KNN and SVM. The need for using both the classifiers for a single problem is to improve the accuracy. KNN will perform well in large datasets but it is a lazy learner. SVM's performance is better than KNN but the choice of kernel selection is a major problem while using SVM independently. Optimal kernel selection of SVM gives better accuracy. So that both of the classifiers have limitations when used independently. Therefore we decided to overcome the limitations of both classifiers by making it as a hybrid model.

2) K-Nearest Neighbor Classifier :

KNN classifier learns the training dataset and predicts the target variable for the test dataset. Parameter selection is the initial process for KNN i.e. the selection of k value. The k value should neither be larger nor be smaller. It is selected with the following formula

$$k = \sqrt[2]{n} \qquad (1)$$

After the parameter selection distance between the test data and each and every data in the training dataset is calculated. There are lot of distance calculating functions in which we are using Euclidian distance to calculate the distance between the training data and the test data. The Euclidian distance formula is

$$d(X,x) = \sqrt{\sum_{i=0}^{n}(X-x)^2} \qquad (2)$$

where X is the training data and x is the test data. Once the distance calculated for each training dataset it is sorted according to the k value. For instance, if the k value is 10, the first ten training dataset in the sorted list will be chosen for further process. These ten data are made into a category and the test data will be predicted as the type of the training data which most often occurs in the category.

3) Support Vector Machines Classifier:

SVM classifier considers the training data points as vectors in space and these vectors are classified into different categories. The classification process is carried out by the hyper plane. The hyper planes are chosen in such a way that it should have a maximum distance from each data points and this hyper plane is called as maximum margin hyper plane. Choosing optimum hyper plane is the main process of SVM classifier which has the following notation.

$$w.x + b = 0 \qquad (3)$$

where w is the vector for the data point x and b is a constant. As we have linearly separable training datasets we chose the above notation to find the maximal margin hyper plane. There are lot of SVM methods available for better classification like multi-kernel SVM, C-SVM etc. The kernel function has the following notation.

$$k(x,z) = \emptyset(x)^T \emptyset(z) \qquad (4)$$

where $\phi(x)$ and $\phi(z)$ are decision functions. This function helps us to choose an optimal separating hyper plane in the space without explicitly performing calculations in this space [16]. In this paper we are using simple SVM algorithm without any kernel functions as we already use KNN for classification.

4) Hybrid model of KNN-SVM classifier:

Hybrid model of KNN and SVM combines the process of both KNN and SVM. Initially KNN algorithm is used to classify the datasets and predict the type of accident. The predicted accident type along with other factors are used to train SVM algorithm. Finally the SVM classifier predicts the accurate accident type by analyzing the factors and the predicted result of KNN.

IV THE MATERIAL AND METHOD

A. Algorithm: Hybrid KNN-SVM

Input: Historical Accident Dataset

Output: Predicted Accident Type

Step 1: Partition dataset into training and test datasets
Step 2: Select k value.
Step 3:Calculate Euclidian Distance between the training a test data with the above mentioned formula.
Step 4: Sort the distance and choose first k data to make it as a group.
Step 5: Find the data which most often occurs in the group and predict the test data as the type of this data.
Step 6: Train SVM with the predicted accident type from KNN.
Step 7: Choose hyper plane to classify the training dataset.
Step 8: Predict accident type and summarize the result.

In this section we will explain the flow of our process with a clear architecture diagram used for accident type prediction. Fig II shows the process flow diagram.

B. Data Collection

The accident dataset required for our analysis is downloaded from UCI Repository which is an open source platform for dataset collection where large amount of datasets are available for different domains. The datasets are downloaded in the form of comma separated value (.csv) files which contains 1000 datasets with 12 variables. The variables are the reasons for the accident like road type, light condition, weather condition and also it contains the target variable "accident type".

C. Data Preprocessing

Once the dataset is downloaded it should be undergone through the preprocessing technique. It is an unavoidable process in data analytics as raw data is unsuitable for analysis. There are five steps in preprocessing technique and in our paper we are using data transformation to make the data suitable for further process. Centering and scaling are the data transformation methods used in our system. Centering is the process of subtracting each and every data points from the mean value of the data while scaling is the process of dividing each data points from the standard deviation of the data points. These two methods are used to transform the raw numerical data into a range which is suitable for classification and regression.

D. Classification and Regression

Classification is the process of classifying the data into certain classes based on our analysis. After this process the classified data is used for regression i.e. prediction. In our process the factors responsible for accidents are analyzed and the target variable is classified into four categories (Crash, Drunk & Drive, Fire, Skid). Initially the dataset is partitioned into two datasets i.e. the training datasets and the test datasets. The training data are used to train the classifier and the test data are used to test the trained classifier and also for predicting the target variable. As it is a machine learning approach we initially train the classifier with some sample dataset which contains the target variable. The classifier learns the dataset by analyzing the factors responsible for accident and the type of accidents. Our target is to predict the type of accidents which belongs to any one of the above mentioned categories. For this classification and regression we are using a hybrid model of hierarchical learning algorithm which combines KNN and SVM algorithms. Initially KNN classifier is trained with the training datasets. It predicts the target variable of the test dataset (i.e. the accident type) by analyzing the factors responsible for accidents given in the training datasets. These predicted accident type is given as input to the SVM classifier to train it. Finally it predicts the accurate accident type.

D. Data Visualization

After the classification and regression process the predicted results are visualized in graphical or tabular format for better understanding of the users. This process is called as Data Visualization. We can also get the summary of the results in numerical format.

V RESULTS AND DISCUSSION

We implemented our hybrid algorithm with the dataset which contains 1000 accident data with 13 variables including the target variable "Accident Type".

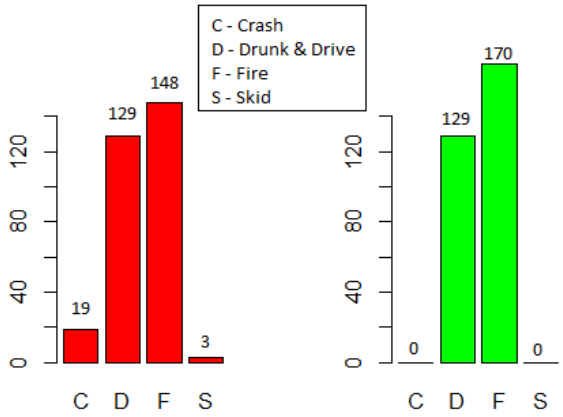

Fig. 2. Comparison of original test dataset and predicted dataset.

We used 700 data as training dataset and the remaining 300 data as test dataset and we came up with the following results. In the test data, it is calculated that 19 accidents are of the type crash, 129 are drunk & drive, 148 are fire and 3 are skid. But by the prediction of our hybrid algorithm there are 129 drunk & drive, 170 fire and 0 crash and skid. Fig. 2 shows the graph of this calculation. Fig. 3 shows the accident severity which is one of the variables used for our prediction. The accident severity is classified into three i.e.

low severity, moderate severity and high severity. Low severity comes under the range of 1 & 2. Moderate severity comes under the range of 2 & 3 and high severity is between 3 and 4.

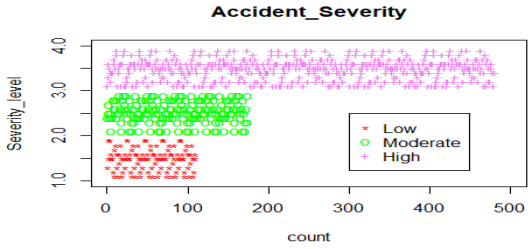

Fig. 3. Accident Severity

Performance measure of our work is carried out by determining the metrics like overall accuracy, sensitivity (Recall), specificity and positive prediction value (Precision), negative prediction value, prevalence, detection rate, detection prevalence and balanced accuracy. These values are calculated for each of the four classes (crash, drunk & drive, fire and skid) used in our dataset. The overall accuracy of our hybrid model is compared with KNN and SVM and is shown in Table 2.

TABLE 2
PERFORMANCE OF DIFFERENT ALGORITHMS FOR THE SAME DATASET.

Model	**Overall Accuracy**	**Kappa**
Hybrid	0.9264	0.8618
KNN	0.602	0.2663
SVM	0.602	0.255

TABLE 3
PERFORMANCE OF KNN (STATISTICS BY CLASS)

Metrics	**C**	**D**	**F**	**S**
Sensitivity	0.1052	0.5504	0.7230	0.0000
Specificity	0.9750	0.7412	0.5497	1.0000
Pos Pred Val	0.2222	0.6174	0.6114	NaN
Neg Pred Val.	0.9413	0.6848	0.6694	0.9899
Prevalence	0.0635	0.4314	0.4950	0.0100
Det. Rate	0.0066	0.2375	0.3579	0.0000
Det. Prev.	0.0301	0.3846	0.5843	0.0000
Bal. Accuracy	0.5401	0.6458	0.6363	0.5000

Table 3 shows the performance of KNN algorithm for the same dataset that we used for our hybrid model. This performance measure proves that KNN is a lazy learner as it gives low accuracy. Performance of SVM for the same dataset is given in Table 4. SVM has better performance compared to KNN. It is proved by each of metrics.

TABLE 4
PERFORMANCE OF SVM (STATISTICS BY CLASS)

Metrics	C	D	F	S
Sensitivity	0.0000	0.5659	0.7095	0.6666
Specificity	1.0000	0.7118	0.5364	1.0000
Pos Pred Val	NaN	0.5984	0.6000	1.0000
Neg Pred Val.	0.9364	0.6836	0.6532	0.9966
Prevalence	0.0635	0.4314	0.4950	0.0100
Det. Rate	0.0000	0.2441	0.3512	0.0066
Det. Prev.	0.0000	0.4080	0.5853	0.0066
Bal. Accuracy	0.5000	0.6388	0.6229	0.8333

Finally the performance measure of our hybrid model is analyzed and given in table 5. By this analysis we proved that our hybrid model is better than both of the above mentioned algorithm.

TABLE 5
PERFORMANCE OF HYBRID KNN-SVM (STATISTICS BY CLASS)

Metrics	C	D	F	S
Sensitivity	0.0000	1.0000	1.0000	0.0000
Specificity	1.0000	1.0000	0.8543	1.0000
Pos Pred Val	NaN	1.0000	0.8706	NaN
Neg Pred Val.	0.9364	1.0000	1.0000	0.9899
Prevalence	0.0635	0.4314	0.4950	0.0100
Det. Rate	0.0000	0.4314	0.4950	0.0000
Det. Prev.	0.0000	0.4314	0.5686	0.0000
Bal. Accuracy	0.5000	1.0000	0.9272	0.5000

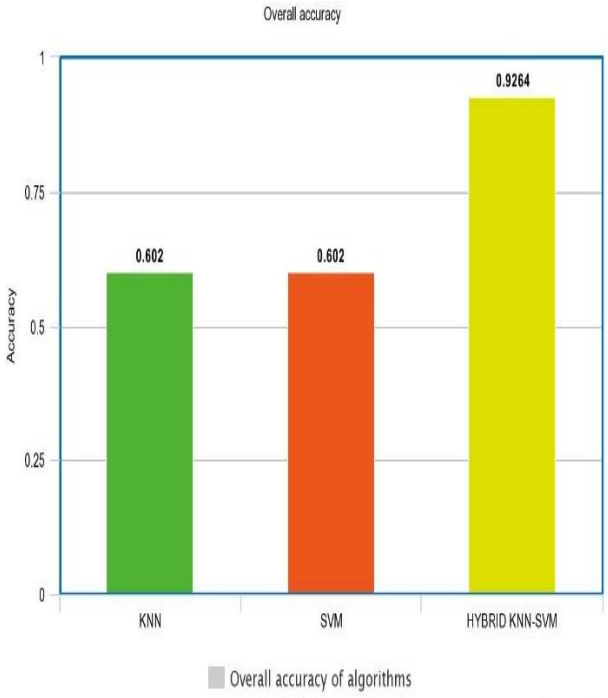

Fig. 4. Overall Accuracy Graph

There are three algorithm are used such as KNN,SVM,HYBRID KNN-SVM .In these algorithm accuracy of output are same for KNN and SVM.In case of SVM and KNN the overall accuracy is far better than these separate algorithm .Overall analysis for these three algorithm above 50%.For 100% accuracy Both KNN and SVM must be used.Thus We used hybrid KNN-SVM is used which gives nearest to 100%.It gives accuracy of 0.9264.

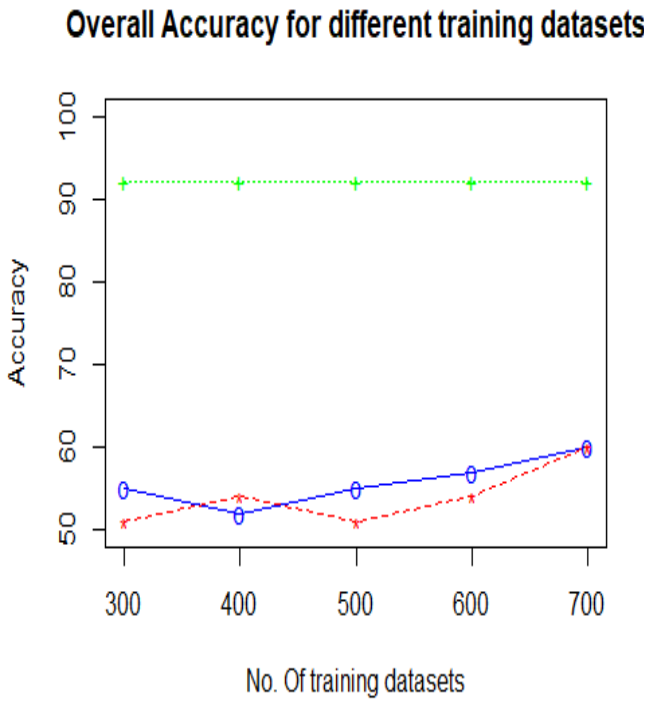

Fig. 5. Precision changes for different no. of training sets

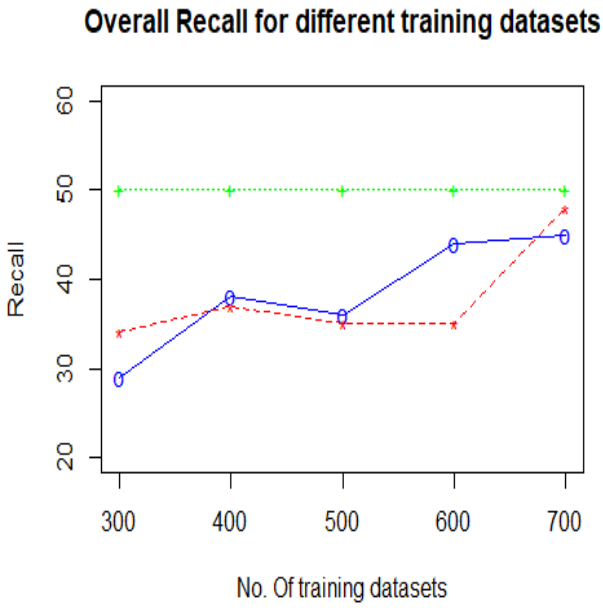

Fig. 6. Recall changes for different no. of training sets

From these graphs we conclude that our hybrid model performs well when compared to other algorithms and it remains same for different training datasets.

VII. CONCLUSIONS

Analyzing the road accidents with more accuracy can relieve traffic crashes in urban areas. In this paper, we proposed a hybrid model that applies both KNN and SVM algorithms to analyze road accidents accurately and predict the type of accidents. These methods and techniques have been analyzed and integrated for road accident analysis. Our method has taken the large historical data of road accidents and predicting the accident type. We have also evaluated the performance of KNN and SVM individually. Thus, we experimentally demonstrated that our proposed model for road accident analysis has superior performance and achieved the accuracy of 92%.

RECOMMENDATION

The study is deals with by taking the kind of strategy as an essential dimension besides others, such as the source of various informational collections used and the relative performance of methods as far as prediction exactness, wherever required, As necessity of future bearings, a lot of expansion still exists in growing new hybrid frameworks in various structures for the road accident data sets.

REFERENCES

[1] J.Cooper and A.James, "Challenges for database management in the internet of things", IETE Technical Review, Vol. 26, no. 5, pp. 320-329, 2009.

[2] G.Kaur, G.Gandhi, "A Framework for Analyzing the RoadAccidents in Data Mining using Rule Mining", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, 2017.

[3] J.Xi, Z.Gao, S.Niu, T.Ding, and G.Ning , "A Hybrid Algorithm of Traffic Accident Data Mining on Cause Analysis", Mathematical Problems in Engineering, pp. 1-8, 2013.

[4] J.Sun and J.Sun, "Real-time crash prediction on urban expressways: identification of key variables and a hybrid support vector machine model", IET Intelligent Transport System, Vol. 10, no. 5 ,pp. 331-337, June 2016.

[5] L.Li, S.Shrestha and G.Hu, "Analysis of Road Traffic Fatal Accidents Using Data Mining Techniques", SERA 2017 - IEEE, 2017.

[6] L.Martin, L.Baena, L.Garach and G.Lopez, "Using data mining techniques to road safety improvement in Spanish roads", Procedia - Social and Behavioral Sciences, Vol.160 , pp. 607 – 614, 2014

.

[7] P.Shetty, P.C.Sachin, S.V.Kashyap and V.Madi, "Analysis of road accidents using data mining techniques", International Research Journal of Engineering and Technology (IRJET) Vol. 04, no.04, 2017.

[8] A.V.Sakhare and P.S.kasbe, "A Review On Road Accident Data Analysis Using Data Mining Techniques", International Conference on Innovations in information Embedded and Communication Systems, 2017.

[9] P.A.Nandurge and N.V.Dharwadkar, "Analyzing Road Accident Data using Machine Learning Paradigms", International conference on I-SMAC IoT in Social, Mobile, Analytics and Cloud, 2017.

[10] S.Kumar and D.Toshniwal, "A data mining framework to analyze road accident data", Journal of Big Data,Nov 2015 .

[11] S.Kumar and D.Toshniwal, "A data mining approach to characterize road accident locations", Journal of Modern Transportation, Springer,Vol. 24, no. 1, pp. 62-72, 2016.

[12] T.Beshah, D.Ejigu, A.Abraham, V.Snasel and P.Kromer, "Pattern Recognition and Knowledge Discovery from Road Traffic Accident Data in Ethiopia:Implications for improving road safety", World Congress on Information and Communication Technologies, 2011.

[13] C.Xu, W.Wang, P.Liu, "A genetic programming model for Real Time crash prediction on freeways", IEEE Transaction Intelligent Transport System, Vol. 14, No. 2, June 2013.

[14] Y.Lv, S.Tang and H.Zhao, "Real-time Highway Traffic Accident Prediction Based on the *k*-Nearest Neighbor Method", International Conference on Measuring Technology and Mechatronics Automation, 2009.