International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-7, May 2020

Traffic Accidents Severity Prediction using Support Vector Machine Models

Zeinab Farhat, Ali Karouni, Bassam Daya, Pierre Chauvet, Nizar Hamadeh

Abstract: ***In recent years, road traffic accidents (RTA) have become one of the highest national health concerns worldwide. RTA have become the leading cause of losing lives among children and youth. Recent studies have proven that Data Mining Techniques can break down the complexity that prevails between RTA and corresponding factors. In this paper, Support Vector Machine (SVM) based on Radial basis function (RBF) and Linear Kernel Function is applied to predict fatal road accidents in Lebanon. The experimental results reveal that SVM using RBF give the highest accuracy (86%) and the best AUC (86.6%). The obtained decision-making model claims to tackle the fatal RTA phenomenon.***

Keywords: ***Data mining, Prediction, Road Traffic Accidents, SVM***

I. INTRODUCTION

Road traffic accident (RTA) is one of the most threatening phenomena four societies are facing. According to the world health organization (WHO), more than 1.25 million of victims and more than 50 million of injuries are reported every year [1]. There are several factors that can influence the likelihood of RTA such as recklessness, time, weather and visibility conditions, vehicle conditions, etc. These factors can vary from country to another.

According to the latest WHO report launched in 2018, RTA fatalities in Lebanon reached 1,090 in 2016. The estimated death rate for the same year is 18.1 per 100,000 of population [2]. According to the statistical study conducted by the Lebanese Internal Security Forces (ISF) and the Lebanese Red Cross (LRC), more than 8 people died weekly in 2019. Their report stated that the number of victims is increasing exponentially day after day which is quite alarming. Scientists have been working hard to decrease the numbers of RTA victims in the world. However, studies in the field of traffic safety indicate that the applied statistical data analysis modeling fails when dealing with nonlinear data.

Revised Manuscript Received on May 30, 2020.
***** Correspondence Author
Zeinab Farhart*, Computer Science, EDST, Lebanese University, Lebanon. Email: zeinabfarhat@live.com
Ali Karouni, Institute University of Technology, Lebanese University. Email: ali.karouni@gmail.com
Bassam Daya, Computer Science, Institute University of Technology, Lebanese University. Email: b_daya@ul.edu.lb
Pierre Chauvet, Computer Science, LARIS, Angers University, France. Email: pierre.chauvet@uco.fr
Nizar Hmadeh, Computer Science, Institute University of Technology, Lebanese University. Email: nizarhamadeh103@hotmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

This could suggest that the correlation between the influence factors and road accidents outcomes is more complicated than can be captured by a single statistical approach [3]. But recent studies showed that data mining techniques have proven to be promising approaches to classify and predict the important interactive causes of road accidents [4][5]. The rest of this paper is organized as follows: In the first part, it provides a literature review that includes recent initiative studies on traffic accident prediction, using different data mining techniques tools. In the second part, the overall model design is presented and the used Lebanese data set is described. In the third part, the proposed methodology of Support Vector machine (SVM) to predict RTA is explained at length. Finally, a discussion of results is provided and conclusion is drawn.

II. LITERATURE REVIEW

Li et al., (2013) utilized SVM and Ordered Probit (OP) models for crash injury severity analysis in China. This study was applied on 1800 crash injuries.Several factors were used such as the length of the exit ramp and the shoulder width of the freeway mainline, etc. The percentage of accurate prediction for the SVM model recorded 48.8% while 44.0% for the OP model. Also the sensitivity and specificity of SVM model were found to be better than OP model [6]. Ibrahim and Far., (2014) developed a Real-Time Transportation Data Mining (RTransDmin) technique. This technique had the ability to examine real-time traffic data set and predict future information about RTA. The authors applied two types of decision tree, J48 and Active Directory tree, using 1385 of accidents records collected by the Department of Transport in United Kingdom. The accuracy results of prediction showed that j48 registered 87.2% while the Active Directory tree method registered 85.9% [7]. In Dubai city, Mohamed (2014) employed SVM using Gaussian Radial Basis function (RBF) to predict the causes of road traffic accidents based on 1000 real crashes using WEKA software. The accuracy of this multi-class SVM model was greater than 75% [8]. Effati et al., (2015) integrated data mining techniques between SVM, coactive Neuro-fuzzy and ANN inference system to discover the influential factors that involved severityprediction of crash dataset in Iran. This new integration method registered prettygoodprediction accuracy (85.49%) [9]. Perone, (2015) applied SVM , Logistic Regression , Random Forest , KNN and Naïve Bayes to create a new prediction model to evaluate injury severity in Brazil. The author used 20798 accident recordings of the city of Porto Alegrers. According to AUC records, the Logistic Regression and SVM satisfied the best scores with 94% followed by Random Forest with 93%andkNN with 90%;

while Naïve Bayes recorded the lowest AUC 83% [10]. Sharma et al., (2016) collected 300 real accident cases in India to apply SVM (Gaussian kernel) and MLP to find the best affecting factors on the majority of RTAs. Their data set were split into training (70%), cross- validation (20%) and testing (10%) using LIBSVM (library for support vector machines) integrated with octave. The results of the study revealed that SVM with Gaussian kernel function retrieved higher prediction accuracy (94%) as compared to traditional MLP (60%) [11].

Gu et al., (2017) applied Support Vector Machine (SVM) to predict fatal road traffic in China. This research aimed to apply comparative study between SVM, K Nearest Neighbor (KNN) and Bayesian network. The results showed that the prediction model of traffic fatalities based on particle swarm with mutation optimization-SVM obtained higher prediction precision (97%) and smaller errors (9%) in training and testing data [12].

Al-Radaideh and Daoud (2018) used Decision Tree (Random Forest C4.5/CART/J45), SVM (polynomial Kernel) and ANN back propagation to detect the influential environmental features of RTA in United Kingdom. The experimental results of this study showed that Decision tree (Random Forest) recorded the best accurate result (80.6%) in predicting the severity of the accidents in UK [13].

Farhat et al., (2019) applied different data mining techniques tools (Decision Trees and ANN) to predict traffic accidents in Lebanon. The results have shown that ANN using Multi Layers Perceptron (MLP) with 2 hidden layers and 42 neurons in each layer was the best algorithm with accuracy rate of prediction (94.6%) and AUC (95.71%) [14].

Karthik et al., (2019) applied different data mining techniques methods (J48, Random Forest and Naïve Bayesian) to predict the major causes for fatal accidents in Thanjavur district, India. 10 years accident data containing different RTA factors were collected (Accident Location, Road Bound, Accident Time, Surface Condition etc). J48 registered the highest accurate result (56.96%) followed by Naïve Bayesian (54%) and the Random forest method (49%) [15].

III. METHODOLOGY

The main objective of the proposed methodology is to apply Support Vector Machine (SVM) technique to predict the fatal RTA. This section explains the suggested research methodology to compare the accuracy of twodifferentSVM functions (Radial basis Function and Linear Kernel Function) and to use the best performing predictive method.

a. Overall Research Design

Fig.1 summarizes the phases of the methodology used in this paper. In the first phase, the balanced dataset is preprocessed to select the RTA attributes between Injured and death cases. Then SVM methods (RBF and Linear Kernel) are applied based on R-GUI software using SVM package (e1071) to build the predictive models. Finally, confusion matrices (Training and testing data) are interpreted based on five evaluation measures (precision, sensitivity, accuracy and specificity) and AUC. The latter evaluation helps us to find the best performing SVM model to be used to predict fatal RTA in Lebanon.

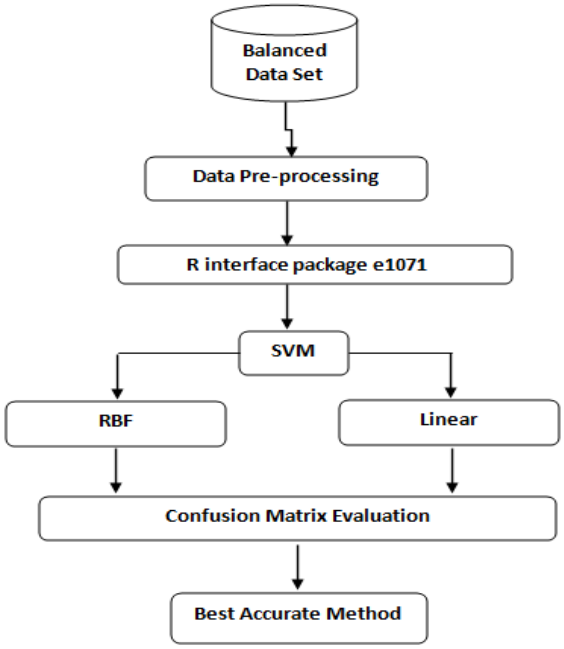

Fig. 1. Methodology Flowchart

b. The Dataset

The dataset containing 11014 traffic accident records with 12 different attributes was collected from the Lebanese Internal Security Forces (ISF) for the years 2016-2017. Each accident record has its own class output: dead or injured. The dataset under study comprises 1100 dead and 9914 injured.

c. Data Pre-Processing

Data pre-processing is a sensitive and important step for handling the data before being usable by the data mining technique tools. In this section, Data is arranged, normalized and nominal attributes are transformed to numeric values. It is not weird that RTA dataset is imbalanced between the two classes (Injured and Dead) which leads to inaccurate classification. The imbalance is a serious problem in the Data Mining Methods. It is caused by the skewed distribution of data between classes [16]. To resolve this matter and improve the prediction accuracy of imbalanced data between dead and injured classes (1100 dead and 9914 injured), 1100 dead is presented 9 times.

Table 1 below underpins the nominal values for each attribute.

Table- I: Nominal Attributes

Attributes	Values	Attributes	Values	Attributes	Values
Month	1:January 2: February 3. March 4. April 5:May 6: June 7:July 8:August 9:September 10: October 11 November 12:December	Official Holidays	1:Yes 2: No	Causes	1:Defects 2: Distracted Driving 3:Drunk or Drugs Driving 4: Design 5:Night Driving 6:Reckless Driving 7:Running Red Lights 8: Sliding 9:Speeding 10:Tailgating 11:Teenage Drivers 12:Wrong-way Driving 13:Tire Blowouts
		Time	1:Morning 2:Afternoon 3: Evening 4:Night		
Day	1:Monday 2: Tuesday 3:Wednesday 4:Thursday 5: Friday 6:Saturday 7:Sunday	Casualty	1:Driver 2: Passenger in front 3:Rear passenger 4: Pedestrian 5: bicyclist		
		Weather	1:Clear 2: Cloudy 3: Sunny 4:Rainy 5: Snowy.	Accident Type	1:Head-on collisions 2:Rear-end-collisions 3:Side-impact collisions 4:Sideswipe collisions 5:Single-car accidents 6:Vehicle rollover
Road shape	1:Straight 2: Bridge 3:Cross 4 :Curve 5: Junction 6: Slope 7:Tunnel 8: Turn	Road	1:Main 2:International 3: Internal	Road Type	1:One way in one direction 2:One road in two directions 3:Two ways in two directions
		Road Status	1:Dry 2:Wet 3:Ice	Class	1:Injured 2: Dead

d. Support Vector Machine (SVM)

Support vector machine (SVM) was founded by Vladimir Vapnik in 1992 [17]. It is an algorithm for prediction and classification of linear and non-linear data.
The aim of SVM is to find the maximum margins of hyperplane. The maximum margin hyperplane gives the maximum distance between the separation decision classes. The training examples that are closest to the largest margin hyperplane are called support vectors [18][19]. A brief mathematical description of SVM algorithm is provided as follows. Assume a training set $Q = \{xi, yi\}_{i=1}^{N}$ with input vector $xi = \{xi^1, \dots, xi^n)^T \in R$ and target labels $yi \in (-1, +1)$, according to Vapnic Formula, satisfies the following conditions:

$$\begin{cases} W^T \emptyset(xi) + b \geq +1, & if yi = +1 \\ W^T \emptyset(xi) + b \leq -1, & if yi = -1 \end{cases} \quad (1)$$

Which is equivalent to:

$$yi[w^T \emptyset(xi) + b] \geq 1, \quad i = 1 \quad (2)$$

Where w the weight vector (maximum margin) and b is the bias

e. Kernel functions

SVM techniques use a set of mathematical functions that are known as the kernel. The aim of kernel functions is to take dataset as input and transform it into the needed form. Kernels in SVM classification refer to the function that is responsible for defining the decision boundaries among the classes.
In recent years, many studies applied SVM based on different functions such as linear, RBF, sigmoid and polynomial, but showed that RBF is the most usable function. Apart from the classical linear kernel function which proposes that the different classes can be separated by straight lines, RBF is used when the boundaries are hypothesized to be curve-shaped [20] [21].

Table- II: Data Splitting

Data	percentage	Total numbers of accidents	Numbers of Injured	Numbers of Dead
Training data	90%	9913	8922	990
Testing data	10%	1100	991	110
Total		11013	9913	1100
Training (90%)		17832	8922	8910 (1100 dead presented 9 times)
Testing (10%)		1981	991	990

On the other hand, researchers found that linear functions based on geometrical separate line can register good classification results if and only if we have linearity of the data set [22][23].

f. *Data splitting*

After data preprocessing, we need to split our dataset into training and testing sets. In this paper, we have used 10-fold Cross Validation and Holdout (training data 90% and testing data 10%) method during our experiments. Table.II ummarizes the dataset splitting in the two phases (Training and Testing).

g. Tools and implementation

For the implementation phase, R-Gui software is used to apply SVM algorithm to predict fatal RTA in Lebanon.

SVM is applied with different training functions (linear kernel function and RBF) using R interface package e1071.

IV. EXPERIMENTS AND RESULTS EVALUATION

In this section, we discuss the experiments and results of SVM application. The evaluation of SVM functions will be based on confusion matrix analysis (precision, specificity, accuracy and sensitivity) and AUC (Area under Rock) that is to see which of the two functions provides best accuracy in predicting traffic accident severity.

Confusion matrix, a performance measurement for machine learning, is based on True Positive (TP) , False positive (FP), True Negative (TN) and false Negative rate (FN) (See Fig.2) . While AUC inform us how much the model is capable of assorting into classes?

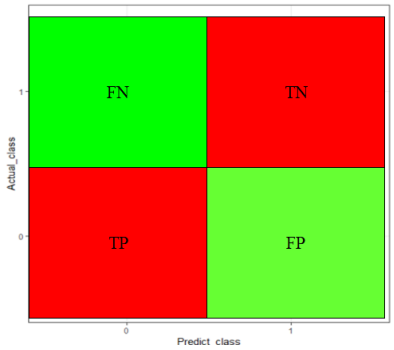

Fig. 2. confusion matrix

2.1. Linear Kernel Function

In this section, SVM linear kernel function is used. 10-fold cross validation is applied to find the best cost (The large margin) in training algorithm over the testing data set.

Fig 3 analysis shows that the highest training accuracy during the 10 folds cross validation has recorded 0.733 at cost = 0.01.

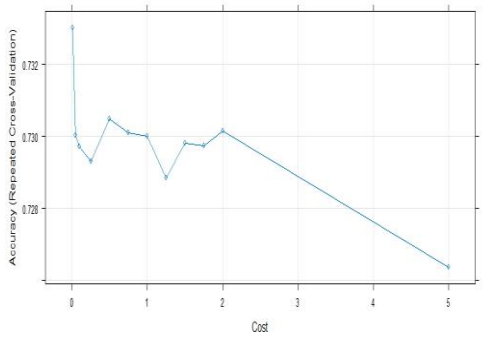

Fig. 3. Linear Kernel function Training accuracy in function of Cost

After finding the best cost of linear function we can simply find the training and testing records of confusion matrices (TP, FP, TN and FN) as shown in Fig.4 and Fig.5. After finding the best cost of linear function we can simply find the training and testing records of confusion matrices (TP, FP, TN and FN) as shown in Fig.4 and Fig.5.

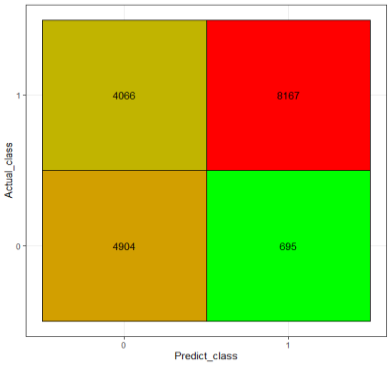

Fig. 4. Training confusion matrix

.

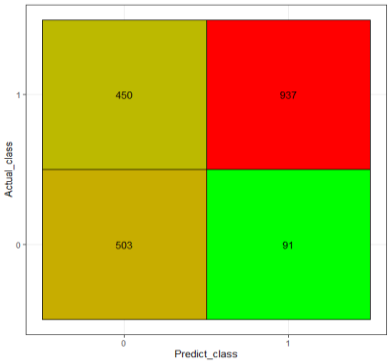

Fig. 5. Testing confusion matrix

The training and testing records of precision, specificity, sensitivity accuracy and AUC of Linear Kernel function are reported in Table.III.

Table- III: SVM (Linear Basis Function) Results

SVM (With Linear Basis Function)	Precision %	Specificity %	Sensitivity %	Accuracy %	AUC %
Training	87	92	87	73.3	72.8
Testing	84.6	91	78.9	72.69	71.2

a. Radial Basis Function (RBF)

SVM is applied using RBF function based over 10-folds cross validation training method. To find the best accurate testing results, we shall find the best cost and the best sigma (smaller sigma tends to make local strict and sharp classifiers) in training data.

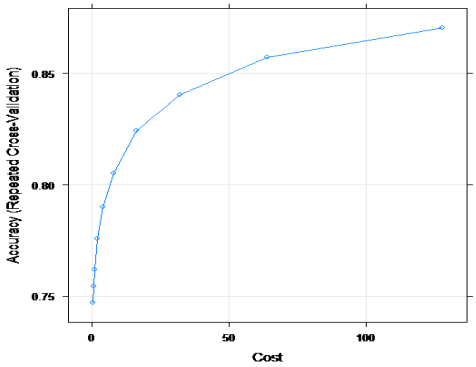

Fig. 6. RBF Training accuracy in function of Cost

Fig.6 has clearly showed that the highest accuracy of RBF registered 0.87 at cost = 128 in training algorithm.

In addition, sigma is found to be 0.057.
The training and testing confusion matrices of SVM using RBF method are shown in Fig.7 and Fig.8.

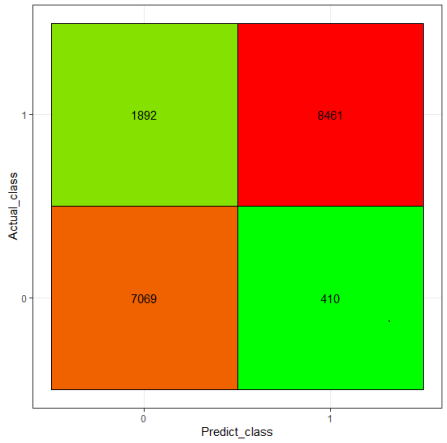

Fig. 7. Training confusion matrix

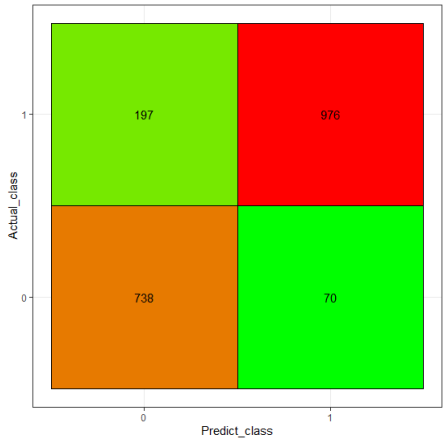

Fig. 8. Testing confusion matrix

Table.IV views a brief conclusion of precision, specificity, sensitivity and accuracy and AUC results in training and testing data:

Table.IV: SVM (RBF) Results

SVM (With RBF)	Precision %	Specificity %	Sensitivity %	Accuracy %	AUC %
Training	94	95	78	87	7.4
Testing	91	91	79	86	86.6

b. Model Evaluation

As noticed from the above training results, SVM with RBF has recorded the best precision (94%), specificity (95%), sensitivity (78%), accuracy (87%) and AUC (87.4%). However, SVM using Linear Kernel Function has registered lower rates ofprecision(87%), specificity(92%), sensitivity (87%), accuracy (73.3%) and AUC (72.8%).
As well, the SVM testing results using RBF has recorded the highest precision (91%), specificity (91%), sensitivity (79%), accuracy (86%) and AUC (86.6%). While SVM using Linear Kernel Function has recorded lower rates ofprecision (84.6%), specificity (91%), sensitivity (78.9%), accuracy (72.69 %)and AUC (71.2%).
The performance of a chosen classifier (RBF) is validated based on the best accuracy, precision and AUC. The precision inform how many of the positives (fatal road accidents occurrences) the model depicting (91%). The AUC (86.6%)based on RBF has a superior ability to classify fatal road accident correctly relying on the analysis between sensitivity and specificity. In addition, the model accuracy(86.4%)has showed a good record of overall correct predictions between Fatal and Injured classes in the entire dataset.

V. CONCLUSION

Road traffic accident is a public health issue which needs to be addressed. Data mining techniques are found to be promising approaches in prediction of RTA severity. SVM is used and adopted in different places after recording high accuracy in that field.
In this research, SVM is implemented using two different functions (Linear Kernel and RBF) to predict fatal road accidents in Lebanon. SVM models are trained and tested using the Lebanese Internal Security forces RTA data set. R-Gui software is used to apply SVM model based on the package e1071. 10 folds cross validation technique is used after data splitting (90% training and 10% testing).
The experiment results show that SVM model with RBF has recorded the best accuracy (86%) and highest AUC (86.6%). While SVM model with Linear Kernel function has recorded lowervalues of accuracy (72.69 %)and AUC (71.2%). But both SVM models with different functions perform very well on nonlinear data, which has been proven in our experiment.
Due to the high accuracy of Data mining techniques in predicting RTA, we will apply in future work different data mining techniques methods such as ANN, Random Forest Tree, KNN and Naïve Bayes. Perhaps these studies reduce the amount of traffic accidents in Lebanon and the world at large.

REFERENCES:

1. G. W. H. Organization, Ed., Global Status Report on Road Safety 2018. Geneva: World Health Organization, 2018
2. World Health Organization. Global status report on road safety 2018.
3. M. G. Karlaftis and E. I. Vlahogianni, “Statistical methods versus neural networks in transportation research: differences, similarities and some insights,” Transportation Research Part C: Emerging Technologies, Vol.19, no. 3, pp. 387–399, 2011.
4. Z. Farhat, Ali. Karouni, B .Daya and P .Chauvet, “An Improved Approach to Analyze Accidents and Promote Road Safety using Association Rule Mining and Multi-Criteria Decision Analysis Methods”, Recent Advances in Computer Science and Communications. Vol. 13 Issue: 1, 2020, pp: 119-128
5. G. Cuenca, Laura, S. Enrique, A. Nourdine, A. Javier, “Traffic Accidents Classification and Injury Severity Prediction”, 3rd IEEE International Conference on Intelligent Transportation Engineering (ICITE), September 2018, pp: 52-57
6. Li. Z, Liu. P, Wang. W, Xu. C, “Using support vector machine models for crash injury severity analysis”. Accid Anal Prev. 2012 Mar; vol: 45, pp: 478-86
7. H. Ibrahim, and B.H. Far, “Data-oriented intelligent transportation systems”, (IRI), IEEE 15th International Conference on In Information Reuse and Integration, pp. 322-329 , 2014
8. E.A. Mohamed, “Predicting Causes of Traffic Road Accidents Using Multi-class Support Vector Machines”. Proceeding of the 10th International Conference on Data Mining, Las Vegas, Nevada, USA, Page 37-42, 2014
9. M. Effati, J.C. Thill and S. Shabani, “Geospatial and machine learning techniques for wicked social science problems: analysis of crash severity on a regional highway corridor”, Journal of Geographical Systems, 17(2), pp.107-135, 2015

10. C.S. Perone, “Injury risk prediction for traffic accidents in Porto Alegre/RS, Brazil”, arXiv preprint arXiv:1502.00245, 2015
11. B. Sharma, V.K. Katiyar, K. Kumar, “Traffic Accident Prediction Model Using Support Vector Machines with Gaussian Kernel”, Proceedings of Fifth International Conference on Soft Computing for Problem Solving.Advances in Intelligent Systems and Computing, Vol .437. Springer, Singapore, 2016
12. Gu. Xiaoning, Li. Ting , W. Yonghui , Z. Liu, W. Yitian , Y. Jinbao , “Traffic fatalities prediction using support vector machine with hybrid particle swarm optimization”, Journal of Algorithms & Computational Technology, Volume: 12 issue: 1, pp: 20-29, 2017
13. A. Qasem Al-Radaideh and J. Esraa Daoud, “Data Mining Methods for Traffic Accident Severity Prediction”, International journal of neural networks and advanced applications, Vol 5, pp: 1-12, 2018
14. Z. Farhat ,A. Karouni, B.Daya and P.Chauvet, “Comparative Study Between Decision Trees and Neural Networks to Predict fatal Road Accidents in Lebanon”, Computer Science & Information Technology (CS & IT), Vol: 9, Number: 11, pp: 01-14, 2019
15. D. Karthik, P. Karthikeyan, S.Kalaivani, K.Vijayarekha, “Identifying Efficient Road Safety Prediction Model Using Data Mining Classifiers”, International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8 Issue-10, 2019, pp: 1472-1474
16. R. Longadge and S. Dongre, “Class Imbalance Problem in Data Mining Review”, arXiv preprint arXiv:1305.1707, 2013
17. V.Vapnik, “The Nature of Statistical Learning Theory”, Springer-Verlag, New York, 1995
18. N. Cristianini and J. Shawe-Taylor, “An introduction to support vector machines”, Cambridge, England: Cambridge University Press, 2000
19. S.R. Gunn, “Support vector machines for classification and regression” Technical Report, University of Southampton, 1998
20. E. Zanaty, and A. Afifi, “Support Vector Machines (SVMs) with Universal Kernels”, Applied Artificial Intelligence, Vol: 25, PP: 575-589, 2011
21. DeepikaKancherla, Jyostna Devi Bodapati, Veeranjaneyulu N,” Effect of Different Kernels on the Performance of an SVM Based Classification” , International Journal of Recent Technology and Engineering (IJRTE) , Volume-7, Issue-5S4, February 2019
22. A. Singh D, PA. Hsiung, K. Hawari, P. Lingras, P. Singh, Advanced Informatics for Computing Research. ICAICR 2018. Communications in Computer and Information Science, vol: 955. Springer, Singapore
23. S.Huang, N.Cai, P.Penzuti Pacheco, S.Narandes, Y. Wang, And W. Xu, “Applications of Support Vector Machine (SVM) Learning in Cancer Genomics”. Cancer Genomics Proteomics 15(1): PP: 41–51, 2018