Analysis and Predictive Modeling of Traffic Incidents in Karachi using Machine Learning

Syeda Batool
National Center of Big data and Cloud Computing
Ned University of Engineering and Technology
Karachi, Pakistan
syedabatool436@gmail.com

Muhammad Ali Ismail
Nationa National Center of Big data and Cloud Computing
Ned University of Engineering and Technology line 4:
Karachi, Pakistan
maismail@neduet.edu.pk

Mir Shabbar Ali
Dean Faculty of Civil Engineering and Architecture. name of organization
Sir Syed University line
4: Karachi, Pakistan
alimirshabbar@gmail.com

2021 IEEE 18th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET) | 978-1-6654-2386-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/HONET53078.2021.9615390

***Abstract*—Road traffic accidents have accounted to extremely dense road traffic and the relatively great freedom of movement given to drivers. Due to the increasing traffic accidents in Karachi, it is vital to investigate the major parameters that are causing these fatalities. For this purpose, machine learning techniques provide a greater advantage over other statistical methods. In this research, a novel approach that applies Random Forest and Support vector machine (SVM) algorithm out of many different machine learning algorithms for modeling traffic accidents prediction. Empirical results show that reasonable accuracy of the developed model. The results further showed the accuracy fluctuated according to the number of attributes in the output parameter. The results of SVM showed better predictions than that from Random Forest. The parameter with less attributes like Disposal has higher accuracy of prediction with Random Forest 83.12% whereas those with greater number of attribute have higher prediction accuracy with SVM e.g. Months with 64.98%.**

***Keywords*—Accuracy, attributes, fatalities, Machine Learning, parameters, Random Forest, Road Traffic Accidents, Support Vector Machine, traffic accidents prediction, Traffic safety.**

I. INTRODUCTION

Road safety has become a major concern affecting socioeconomics of a country. It has a massive impact on society as well as in the economy of our country as there is an immense cost of fatalities and injuries. According to World Health Organization (WHO, 2013) statistics, there are approximately 1.25 million fatalities due to road accidents and 20-50 million who are seriously injured and living with long-term disability every year around the world ranking the traffic accidents as the ninth most common cause of death among all age groups according to the data collected from 178 countries, It is very important to reduce the highway traffic accidents in the developing countries, and therefore the traffic safety should be improved by the analysis of accident characteristics.

Three major factors that play a major contribution in road accidents are human-related factors, vehicle-related factors, and roadway- related factors [1]. Highway Safety Manual published by American Association of State highway and Transportation (AASHTO) (2008) states that road accidents are occurred due to 3% roadway factors, while combined roadway-related factors and other factors are responsible for 34% contribution. is attributed by a Moreover 67 per cent of accidents occurred due to human errors whereas poor infrastructure, road designing and deteriorating road conditions were responsible for 28% of road accidents.

Road Traffic Accidents are common phenomena over the arteries of Karachi. With nearly 25,781 Road Traffic Collision (RTC) fatalities in year 2013 in five RTIRPC (Road Traffic

Injury Research and Prevention Centre) surveillance centers [2]., this problem not only denting the work force of the city but also highest proportion of people involved in Karachi belong to the age group 20-45 most of them are workers hurting the families[3].

Karachi is experiencing an increase in road accidents at an alarming rate, despite the city's many existing socio-political problems, confirming that the city ranks fourth in the world as the most accident conceiving city in the world. The city has recently experienced an increase in RTAs (Road Traffic Accidents) from 2007 to 2012 due to recent infrastructure development within the city regardless of considering the heterogeneous road users [4] However, according to a recent study conducted there was 35.6% decrease in number of fatalities reported during March-June 2020 than that reported in March-June 2019 (Figure 1), possibly due to the restrictions during lockdowns to control the COVID-19 pandemic during March-July 2020.

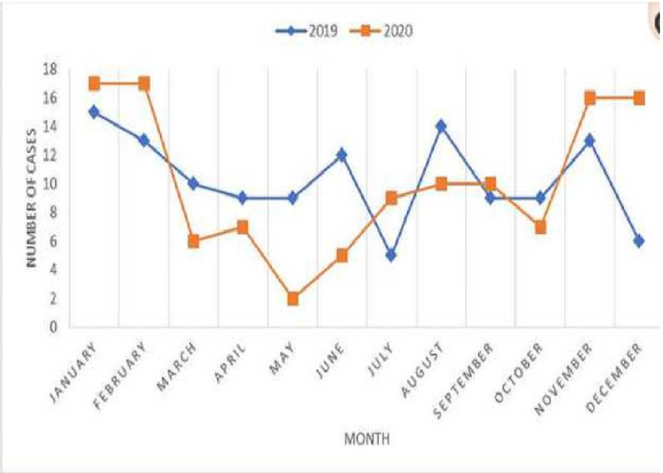

Fig. 1. Monthly distribution of RTA fatalities.RTA, road traffic accident

II. RELATED WORK

Data analytic methods are employed by researchers so as to induce helpful data concerning variables characterization of the road accident, in order to establish rules and inferences from hidden patterns, profile behaviors which are useful to profile drivers or drivers' behavior on the road, to delimitate unsafe areas for driving, by generating classification rules related to road accident data for variable selection useful in real-time model of accidents and to pick relevant variables for training different strategies, like artificial neural networks and deep learning algorithms [5].

Due to a combination of several factors causing road accidents, the identification of the certain cause of road accidents becomes complex due to excess interrelated road accidents factors.

First the data set is divided into training data and a test data set. The training data comprises of observations called attributes and output variable (binary in the case of a classification model). The algorithm is run over the training data and a tree is generated that can be read like a series of rules. These rules are then run over the test data set to determine how good this model is on "new data." Accuracy measures are provided for the model in which a popular technique is the confusion matrix which is a table that gives information about how many cases were correctly versus incorrectly classified. If the model looks good, it can be deployed on other data available. Based on the model, the company takes decisions.

A. Heterogeneous nature of Traffic accident data and Problems with Statistical models

Several Statistical models were developed, including linear regression, Poisson models, Poisson-gamma models, negative binomial models, generalized estimating equation, random parameters, and bivariate and multivariate models to quantify the relationship between road geometric elements and road accidents. However unsatisfactory results showed by conventional linear regression models due to normally distributed properties generated negative or non-discrete values of accident rates. As a result, Poisson and negative binomial models were prompt and used instead thanks to their random and random and sporadic data consideration and easy estimation of variables relationships in spite of overestimated or underestimated results [6].

B. Modern Applications of Big Data Analysis and Prediction of Traffic Accidents

On the aspect of the algorithms and computational methods employed in road accident research to analyze and forecast road accident data include clustering algorithms; decision trees and classifiers; association rules and natural language processing algorithms, SVM, Naïve Baise, Logistic regression, and artificial neural network.

Clustering algorithms were used to study abrupt braking events in real time, considering the time and location to determine sectors where driving is dangerous. As a result of the batch clustering analysis results, correlations were obtained that indicate potentially dangerous places for driving, according to the time of day [7]

Other prediction models such as K-means, decision tree classifier, a hierarchical tree approach to identify the main predictors of accident risk, prediction of number of accidents on any road or intersection and determination of the most significant variables suitable for the prediction of the traffic accident severity.

A multivariate logistic regression model, for calendar years 1999–2008, was developed including the input parameters crash direction (front, left, right, and rear), change in velocity (delta-V), multiple vs. single impacts, belt use, presence of minimum one older occupant (≥55 years old), presence of at least one female in the vehicle, and vehicle type (car, pickup truck, van, and sport utility).to predict the probability that one or more occupants with serious or incapacitating injuries will accompany a crash-involved vehicle. The model was established using predictor variables that may be readily available, post-crash using telematics systems. [8]

SVM models were employed by [9] to investigate driver injury severity patterns in rollover crashes using two-year crash data collected in New Mexico and found reasonable predictions delivered by SVM models and outperformance of the polynomial kernel over the Gaussian RBF kernel.

SVM model to handle multidimensional spatial data in crash prediction was performed by [10]whose results showed that the SVM models outperform the non-spatial models in terms of model fitting and predictive performance in addition to providing better goodness-of-fit by the SVM models compared with Bayesian spatial model with conditional autoregressive prior when utilizing the whole dataset as the samples.

Against this background, the principal goal of the study is to develop a systemic approach to identify and estimate the key parameters responsible for the road traffic accidents in a developing country like Karachi. The research concerns the accidents occurring in the black spot locations of Karachi city. The Data is collected from five trauma receiving hospitals of Karachi namely JPMC, CIVIL, ASH, LNH AND AKH.

III. METHODOLOGY

This research explores the application of machine learning techniques on four years' accident data in Karachi for modeling traffic accident data. The problems of sparsity of data are addressed and methodology to clean data for model testing and training is discussed. The selection of appropriate algorithm among a list of techniques for the prediction of accidents is identified through feature selection method of the parameters that are to be incorporated in the model and the accuracy is compared based on the importance of parameters and features of attributes.

A. Data Descriptions

The Road Accident data provided from RTIRPC (Road Traffic Injury Research and Prevention Center) had many data flaws. Most of the values were missing or not readable. The dataset used in this research is the four years record of accidents in Karachi, Pakistan from year 2009 to 2012.The dataset comprised of 59 attributes and 130,584 tuples, initially. The provided dataset was very much ambiguous and needed a lot of preprocessing efforts. The data was first prepared in order to train the system. Our task was to develop a machine learning based intelligent model that could accurately classify the target variable. This can in turn lead to greater understanding of the relationship between different factors of accidents such as driver, vehicle, roadway, and environment and driver injury severity. Accurate data analysis results could provide crucial information for the road accident prevention policy. A supervised learning algorithm is tried to map an input vector to the desired output class.

B. Data Preprocessing

Initially, it is observed that out of 59 attributes, 37 attributes contain approximately more than 95% missing data values. So these 37 attributes had to be eliminated for data analysis and training, as such a large number of missing values cannot be imputed by any value. Remaining 22 attributes still had some missing values which were worked out so that the selected attributes must contain 100% of data

After observing, some of the selected attributes found unimportant for data analytics such as patient ID, name, address, occupation, name of data collector and contact details. Since these fields are unnecessary so they were eliminated in order to reduce the training overhead.

After reduction, only 15 attributes were remained that are represented in the Table I. All these attributes contained categorical data. In order to deal with machine learning classifiers, all these attributes were converted into numeric digits.

C. Exploratory Data Analysis

After the preparation and processing of data, Exploratory Data Analysis (EDA) was conducted to explore data and understand the behavior of data. Exploratory data analysis or EDA is an approach of data visualization whose purpose is to explore data and understand the behavior of data. Mostly, graphs are used for exploratory data analysis. For data visualization, we performed graphical analysis of each attribute and some fruitful outcomes were generated. For example, in Karachi, the most vulnerable vehicle is motor bike. The month in which most of the accidents occur is August and so on.

D. Feature Selection

Feature selection is basically selecting the important fields that are directly contributing in predicting the model's behavior and abandoning the unimportant ones. By considering the contributing attributes only, we can not only improve our processing speed but also get more accurate results.

Out of three types of feature selection: filter method, wrapper method and embedded method, the wrapper method was used which consists of forward selection, Backward selection and Boruta algorithm. In forward selection, minimum attributes were taken initially for designing a model and its accuracy is checked. One by one attributes were added until the optimal accuracy was obtained. Backward elimination is opposite of forward selection, in which all attributes are taken at the beginning and attributes are eliminated one by one till the maximum accuracy is achieved. Boruta algorithm is a wrapper built around a random forest classifier for feature selection. Similar results were obtained from these three methods of feature selection and it was found that all the attributes are significant other than age. We discarded the field of age to get better results. Now all the attributes were significant and were ready to train using different machine learning classifiers.

E. Implementation

This step consists of applying different machine learning classifiers one by one to check the accuracy of each algorithm through training, testing and validation. For machine learning, data was split into training and test dataset. 70% of the dataset was used for training and the remaining 30% was used to test the models. Town was taken as the target variable with 19 entities. The results obtained from each algorithm were compared and the best model was selected.

A number of machine learning classifiers and regressions were tested including decision Tree, Ada boost, K Nearest Furthermore, random forest also avoids over- fitting of data achieving appropriate results. The more randomness is injected in forests, the more accurate results will be generated. Random inputs and random features are more responsible for accuracy and produce good results in classification than in regression. [11]

F. Model Development

The work was conducted in four steps: (1) studying and processing the data to be made ready for training (2a) selecting the most important fields contributing to model by feature selection. (2b) Eliminating attributes one by one by backward elimination to achieve maximum accuracy (3) Increasing training set volume (4) Training with different groups of features and testing model performance. For last performance we trained the data from year 2009-2011 and test the data in 2012.

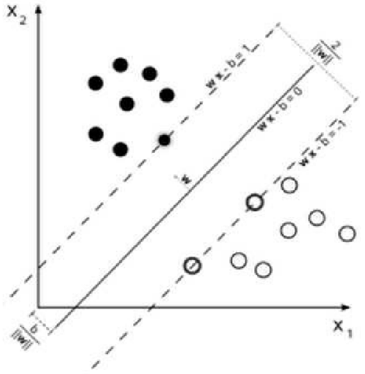

Fig. 2. SVM Model Presentattion

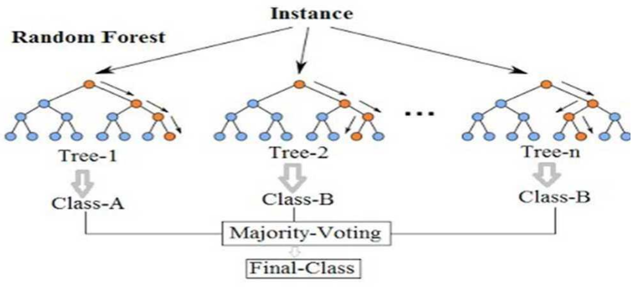

Fig. 3. Random Forest Model

Table I NUMBER OF CLASSES IN EACH ATTRIBUTE

Fields	Attributes
Respiratory rate, rr	1 – 5
Systolic Blood Pressure, sbp	1 to 5
Glasgow Coma Score, gcs	1 to 5
Vehicle (modes of transfer)	1 to 9
Gender	2
Location Detail, ld	2
Center (representing four main hospitals of Karachi)	4
Arrive by, arby	1 – 5
Reason of accident, bd1	1 – 19
Town	1-18, 19=out of city
Time (hrs)	1 to 24
Age (according to age groups)	1 to 5
Months	1 to 12
Year	2009-2012
Disposal, disp	1 to 6

The final dataset contained 15 attributes, each of which implemented for both supervised machine learning algorithms for a single variable prediction. Based on these, we reviewed the finally selected variables in terms of the methods used, performance measures as well as the accuracy.

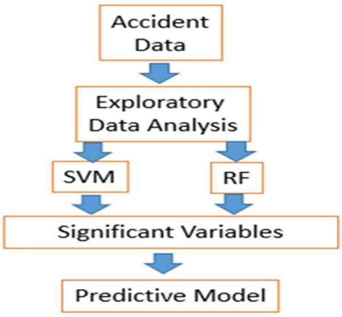

Fig. 4. Model Development Framework

IV. RESULTS AND DISCUSSIONS

The applications of SVM and RF stages provided the most appropriate combination of model variables and resulting accuracy of these selected variables was compared as it is shown in Fig. 5. We evaluated the models by using accuracy of each model. Random Forest and SVM were selected as the final algorithms due to their increased classification characteristics. In this way each attribute was taken as target variable and tested to determine the algorithm with which it achieved the best results.

The performance of both predictive models is summarized in Table II. Names of classes, their number of attributes, and the corresponding supervised machine learning algorithms used to predict them are discussed. For each of the class,, the better performing algorithm is also described in this table.

The conducted study attempted to originate a better understanding of the analysis of big data related to Road Traffic Accidents (RTA) through machine learning in Karachi.

From the results obtained, it can be presented that both Random Forest and SVM can be used for the prediction of a target variable in road traffic accident study. However the accuracy of both models differentiated based on the amount of values in the target variable.

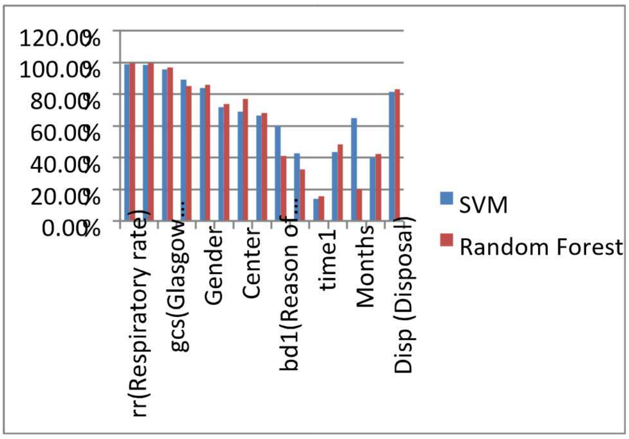

Fig. 5. Graphical presentation of results

TABLE II ACCURACIES OF CLASSIFIERS BASED ON ATTRIBUTES

S. No	Attributes	Number of classes	SVM	Random Forest
1	rr(Respiratory rate)	1-5	98.78%	99.84%
2	sbp(Systolic Blood Pressure)	1 to 5	98.67%	99.59%
3	gcs(Glasgow Coma Score)	1 to 5	95.57%	96.66%
4	Vehicle	1 to 9	88.92%	85.04%
5	Gender	2	83.80%	85.84%
6	ld(Location Detail)	2	71.63%	73.85%
7	Center	4	68.85%	77.10%
8	arby (Arrive by)	1- 5	66.61%	67.99%
9	bd1(Reason of Accident)	1 – 19	59.68%	40.99%
10	Town	1-18, 19=out of city	42.64%	32.59%
11	time1	1 to 24	13.80%	15.81%
12	Age	1 - 5	43.57%	48.49%
13	Months	1 - 12	64.98%	20.26%
14	Year	2009-2012	39.98%	42.20%
15	Disp (Disposal)	1 - 6	81.24%	83.12%

The accuracy of both the types of classifiers was compared. The results (Fig. 6)showed the accuracy fluctuated according to the number of attributes in the output parameter. It is found that Random Forest is good for the target variable where we have lesser entities (less than 5). But for those target variables, where the possible outcomes are more than 5 (approximately 15 to 20) like town, Support Vector Machine with radial kernel is better to use.

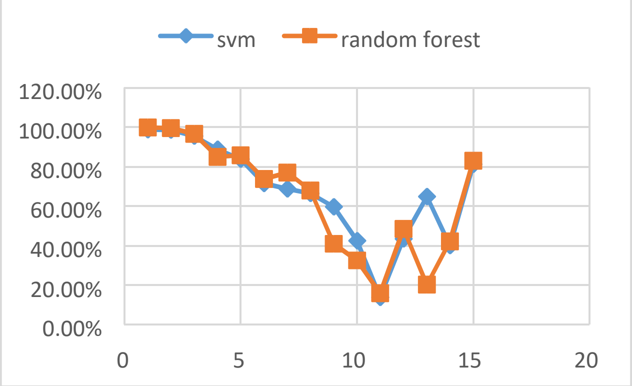

Fig. 6. Comparison of accuracies based on number of attributes

V. CONCLUSION AND RECOMMENDATIONS

This paper investigated the application of machine learning techniques for developing traffic accident prediction model using heterogeneous urban data which is a vital problem to transportation and public safety which is also very challenging due to its spatial heterogeneity and its non-linear separable nature. The comparison of different machine learning algorithms is carried out in order to determine which algorithm gives the highest accuracy when tested with target variable of specific number of entities. Results show that our proposed approach significantly improved Random Forest and SVM accuracy to maximum 99.84% and 98.78% respectively. Furthermore, the prediction accuracy of Random Forest method is superior for parameters of less attributes, fast computation speed, and good interpretability whereas Support Vector Machine (SVM) showed better results in case of parameters with greater attributes.

Despite the fact that human factors play a significant role in road accidents, controlling and predicting them is difficult. However, investigation of roadway factors, especially roadway geometric design, may aid in their indirect prediction and control. The accuracy of the same models with attributes related to traffic engineering, such as vehicle speed, can be compared to the results obtained.

The built model application can be maximized by making it available to traffic regulatory committees for public awareness on local and state levels, as well as safety commissions and/or safety teams, emergency-based analysis services, regular or month-by-month inspection programs to improve safety education and training at the local and/or state levels.

ACKNOWLEDGMENT *(Heading 5)*

The work of this paper is supported by National Center for Big Data and Cloud Computing (NCBC), NED University, Karachi. We thank the Road Traffic Injury Research and Prevention Center (RTIRPC) for providing motor vehicle crash data as primary data in the early stage of this work.

REFERENCES

[1] W. Haddon, “Advances in the Epidemiology of Injuries as a Basis for Public Policy,” *Public Health Rep. 1974-*, vol. 95, no. 5, pp. 411–421, 1980.

[2] R. Jooma and M. A. Shaikh, “Descriptive epidemiology of Karachi road traffic crash mortality from 2007 to 2014,” *J Pak Med Assoc*, vol. 66, no. 11, p. 6, 2007.

[3] S. Zubair and J. Kazmi, “Spatial Framework for the Assessment of Road Traffic Accidents in Karachi,” *J. Basic Appl. Sci.*, vol. 9, pp. 525–532, Jan. 2013, doi: 10.6000/1927-5129.2013.09.67. J. H. Kazmi and S. Zubair, “Estimation of Vehicle Damage Cost Involved in Road Traffic Accidents in Karachi, Pakistan: A Geospatial Perspective,” *Procedia Eng.*, vol. 77, pp. 70–78, Jan. 2014, doi: 10.1016/j.proeng.2014.07.008.

[4] C. Gutierrez-Osorio and C. Pedraza, “Modern data sources and techniques for analysis and forecast of road accidents: A review,” *J. Traffic Transp. Eng. Engl. Ed.*, vol. 7, no. 4, pp. 432–446, Aug. 2020, doi: 10.1016/j.jtte.2020.05.002.

[5] M. H. Islam, L. T. Hua, H. Hamid, and A. Azarkerdar, “Relationship of Accident Rates and Road Geometric Design,” *IOP Conf. Ser. Earth Environ. Sci.*, vol. 357, p. 012040, Nov. 2019, doi: 10.1088/17551315/357/1/012040. [6] G. Cao, J. Michelini, K. Grigoriadis, B. Ebrahimi, and M. A. Franchek, “Cluster-based correlation of severe braking events with time and location,” in *2015 10th System of Systems Engineering Conference (SoSE)*, May 2015, pp. 187–192, doi: 10.1109/SYSOSE.2015.7151986.

[7] D. W. Kononen, C. A. C. Flannagan, and S. C. Wang,

"Identification and validation of a logistic regression model for predicting serious injuries associated with motor vehicle crashes," *Accid. Anal. Prev.*, vol. 43, no. 1, pp. 112–122, Jan. 2011, doi: 10.1016/j.aap.2010.07.018.

[8] C. Chen, G. Zhang, Z. Qian, R. A. Tarefder, and Z. Tian, "Investigating driver injury severity patterns in rollover crashes using support vector machine models," *Accid. Anal. Prev.*, vol. 90, pp. 128–139, May 2016, doi: 10.1016/j.aap.2016.02.011. [9] N. Dong, H. Huang, and L. Zheng, "Support vector machine in crash prediction at the level of traffic analysis zones: Assessing the spatial proximity effects," *Accid. Anal. Prev.*, vol. 82, pp. 192–198, Sep. 2015, doi: 10.1016/j.aap.2015.05.018.

[10] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.