Road Accident Prediction and Feature Analysis By Using Deep Learning

[1]Aradhana Behura, [2]Ashutosh Behura
[1,2]Veer Surendra Sai University Of Technology,Sambalpur,Odisha,India
[1]aradhanabehura@gmail.com, [2]btechitvssut@gmail.com

***Abstract*--With the growing number of road accidents in the city Kerala, it is authoritative to carry out road accident mitigation processes to prevent damage of many lives. Handling of information mining has usually encountered issues like noise, high dimensionality and imbalance data. High dimensionality denotes to the large quantity of attribute among the data set. On the off chance that the quantity of attributes in a datasets is huge, at that point it is in all likelihood that a large portion of these qualities are repetitive or futile for building an prediction model and better result can be accomplished if the redundant attributes from the dataset are removed. Feature selection techniques are used to select the subset of the features which is used in the learning process. The SMOTE technique is used to handle the imbalance of the Kerala road data set then δ – agree based boosting algorithm for stacked autoencoder techniques are used for feature selection as well as data classification analysis. From this research, we conclude about various causes of the accidents, then find out total accidents per annual in separate state in particular time interval and helps to find accident trend in each state in each time interval over the year in each state.**

Keywords: *δ* – agree based AdaBoost and Auto encoder, Heterogeneous Traffic, Deep learning, prediction, Classification, Urbanization.

I. INTRODUCTION

Detection of road accidents are very important for safety communication. In developing countries enormous growth of traffic, improper traffic movement with no lane discipline invited accident and delay especially at intersection. Delay at signalized intersection depends on so many parameters like signal timing and number of phases, vehicle composition, space availability, vehicle headways, queue length etc. The most frequently used terms of delay that are used by traffic engineers and researchers- stopped time delay, approach delay, travel time delay, time in queue delay, control delay. Travel demands vary throughout the day and congestion occurs frequently most of the time, especially during morning and evening peak hours. When demand exceeds capacity, queuing of vehicles occurred at the intersection and continues until the demand decrease below the capacity level allowing the dissipation of queue. Such condition is known as oversaturated condition. The most effective way that can be accepted to avoid oversaturation is to restrict demand. At the intersection, capacity can be increased by proper channelization, restriction of on-street parking during peak hours and other traffic control measure. Delay has two major components known as uniform delay which is determined based on signal timings and traffic volumes. In current society, regular day to day existence is personally worried about transportation making many issues on it. Surrounded by this problems, one of the significant problems are regarding road accidents, then it is essential to maintain a strategic distance from those accidents or diminish harm from them. Foreseeing conceivable auto collisions can be a procedure for those objectives. To foresee road accidents, we can utilize video picture from cameras of the different traffic information to examine[1,2,3,4]. We are worried about breaking down traffic information so that we can anticipate conceivable accident. The expectation through information analysis is made out of for the most part the order examination through gaining from past information in data mining procedure[5,8]. The grouping examination learns the preparation informational index and makes a norm foreseeing model for an expectation result. In view of these, the model predicts the outcome to emulate the current substance. Making another foreseeing model includes various issues, the first is unevenness information. Awkwardness information implies information in which there is an extensive distinction between the watched sizes from one informational index. To take care of this issue, testing strategies can be utilized, of which there are two sorts: over sampling and under sampling [4]. Under-sampling includes utilizing all the observation esteems in a small class and utilizing some portion of the results in a observed large class. The over-sampling includes utilizing all observations in a huge class and expanding the size of the observed value in the small class and afterward utilizing this worth. Then under-sampling computes the utilization of lost information While such testing methods can accelerate handling information, losing the dependability of information can't be kept away from. Then again, over-sampling uses all information, yet demands

more assets for handling extra information [5]. The subsequent issue includes information handling to make a preparation in the training dataset. This data set has a lot of numerous features that can influence the expectation result. In this manner, preparing the dataset for the a wide range of sorts of information is time consuming, more resources are required and depending upon the information size. We mean to take care of these two issues and do effectively forecast the accident rate in city kerala by using deep learning model named as auto encoder procedure. In this way, this paper describes about the data processing as well as classification steps which provides better accuracy . Our projected model involves in the four steps. When the preprocessing of the target data sets are finished, then these are combined and the training data sets are generated. The sampling techniques are carried out a balanced classification based on the training accident data set of Kerala. If the positive classes are outnumbered by negative classes then the dataset is called as imbalance dataset which is described in Fig.1. The false negative result can be harmed the important minority class. Noise refers to missing and incorrect values in a dataset. The sampling techniques convert the dataset which is imbalance to a new balanced dataset by removing or adding instances of the data until an anticipated class proportion is reached.

A. Create preparing dataset and over-inspecting

To assemble a learning information, the preprocessed information[16] were consolidated, and a unit learning information model at that point was formed. The information was then altered to actualize the final learning information. The features that influenced the objective factors were constantly viewed as when the learning information were made. In this way, recognizable proof (ID) was applied to utilize the factors of information adequately when the learning information were made. Unfortunately, the training information contained the information imbalance. To handle this issue, an over-sampling activity was processed to repair the information. Past research [20] proposed[18,2,4] a strategy for over-sampling the minority class, which is called synthetic minority over-sampling method (SMOTE)[11,14,15]. It [12,13]accomplishes its objective by making synthetic models utilizing real information. It makes synthetic models utilizing each models k nearest neighbor models. To create this synthetic pattern, many procedures are accompanied: take the modification among the attributes and that of a nearest neighbour of its. Then the differences are multiplied with a number among 0 and 1 randomly and this is added to the attributes of real time examples. This real time designs are done for the attributes . Here chapter 2 describes about the related work then chapter 3 introduces about the proposed model. The chapter 4 & 5 depicted about the performance analysis and result of the proposed scheme.

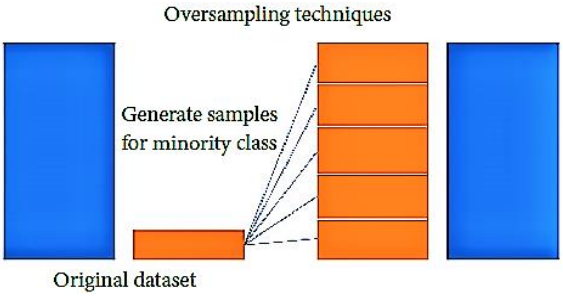

Fig.1. SMOTE Technique

II. RELATED WORK

MapReduce[17] and hive a programming model for handling enormous informational data with an equal, appropriated algorithm on a bunch, which forms large information effectively [6]. Along these lines, in numerous investigations on enormous content information preparing, MapReduce stages have been utilized [7]. The over-sampling strategy is utilized to take care of the imbalance information issue in the preparation dataset. Very few studies have been conducted for oversaturated traffic conditions in developing countries like India. The conventional method for analysis of traffic delay in India is the U.S. Highway Capacity Manual (HCM)[18]. The methodologies and procedure of this manual evolved from a broad range of empirical research conducted in U.S.A. since 1950. Methodologies that have been developed in HCM are for homogeneous traffic flow condition where vehicles follow proper lane discipline. The traffic situation in India is highly heterogeneous with non-lane discipline. Also occurrence of oversaturation is common at urban arterial streets especially during peak periods. As a result, distorted values have been observed for delay estimation when compared to observed field delay. To overcome this situation for mixed traffic conditions, an attempt has been approached to estimate delay under oversaturated

traffic system with initial queue length at signalized intersection. The modified Webster's delay model and developed an adjustment factor using artificial neural network (ANN) for heterogeneous traffic condition in India. It has been observed that the adjustment factor has impact on volume-capacity ratio, number of lanes, proportion of heavy vehicles, approach width and signal controller type. Satisfactory results have been found from the comparison of estimated delay by proposed model with observed field delay [9,10].

III. METHODOLOGY

In this paper, SAE and the hybrid algorithm will achieve a discriminant one by adjusting through the labelled information. The deep learning technique known as SAE (Stacked Autoencoder) is used to build many models for forecasting of road accidents in Kerala. The δ – agree based AdaBoost algorithm is used to enhance the learned scheme.

A. Preprocessing

The information preprocess was required to select the variables of information for preparing also, every perfect estimation of the factors. The factors of information were chosen agreeing to every characteristic of each variables and excluded variable in the absolute information. The perfect worth process was supplanted with overlooked or introductory qualities by considering each estimation of the variables. Moreover, the formatting was important to coordinate the information.

B. Stacked Autoencoder is a unsupervised learning procedure and it's layers are trained by building a deep learning model (stacked autoencoder). This architecture have three layers (input layer, hidden layer and output layer). The output section is expected to reproduce the input section, thus the hidden section can be viewed as a type of data encodings of the input section. The weights value of the each layers can be set to functional local suboptimal.

C. δ – agree based Boosting algorithm:

The aforementioned autoencoder model aims to learn the hypothesis which described as G(.), it forecasts the rate of accident by viewing the recent road accident rate of the city kerala. Here we describe about accident information x as $\{V_{i,j}\}_{ii = 1....M,\ j = t.....N+1,}$ $V_{i,j}$ is denoted as rate of accident information on the i^{th} position at j^{th} time series interval. The accident rate prediction in the kerala of the upcoming interval by taking K^{th} SAE architecture on the i^{th} position can be introduced as $G_i^{(k)}(x)$. Without considering the loss factor, researchers simplify the $G_i^{(k)}(x)$ as $G^{(k)}(x)$ ignoring the factor (i). The prediction of s^{th} number of the samples denoted as

$$\hat{y_s} = G^{(k)}(x_s), \text{ Here } T = \{(x_s , y_s)\}_{s = 1....s}$$

Y_s is denoted as groundtruth of $\hat{y_s}$, s introduces about the training instances. The adaboost procedure is focused on a particular place, then this is easy to be drawn-out to all positions. When accidents are more in the kerala, to get better accuracy, SAE is used with huge prediction error. A δ – agree method is used in the boosting phase and this discriminative function is derived from equation (1).

$$\zeta(|\ G^{(k)}(x_s) - y_s| - \delta) \qquad (1)$$

$$\text{Where } \zeta(x) = \begin{cases} 1, & x > 0 \\ -1, & otherwise \end{cases}$$

In this weighting scheme, δ value will take positive effects, if the estimation miscalculation exceeds than δ. If the expectation error is huge, then autoencoder will be trained by compelling more consideration of those test cases. Then the discriminative miscalculation of the autoencoder(AE) is described as:

$$\epsilon^{(k)} = \frac{1}{2}\sum_{s=1}^{s} w_s^{(k)}[\zeta(|G^{(k)}(x_s) - y_s| - \delta)+1] \qquad (2)$$

A weight notation $W_S^{(k)}$ for the each instance for the k^{th} number of stacked autoencoder(SAE). In the beginning, the weight value of the instances is equal to the $W_s^{(1)} = \frac{1}{s}$ for the first stacked autoencoder. The significance of this architecture is decided by this discriminative error shown as:

$$\alpha^{(k)} = \frac{1}{2}\log\frac{1-\epsilon^{(k)}}{\epsilon^{(k)}} \qquad (3)$$

In equation 3, conclude that if discriminative error is smaller then it achieves additional importance this one gains. The updated new weights value of the instances can be used as per the discriminative miscalculation as well as the importance of the stacked autoencoder.

$$W_s^{(k+1)} = \frac{w_s^{(k)}}{\zeta^{(k)}} e^{\alpha(k)} \zeta(|G^{(k)}(x_s) - y_s| - \delta) \quad (4)$$

Where

$\zeta^{(k)} = \sum_{s=1}^{s} w_s^{(k)} \; e^{(k)} \; \zeta \; (|\; G^{(k)} \; (x_s) - y_s| - \delta)$ is known as normalization issue. The training data set can be expanded by taking the help of a replication factor known as $r_s^{(k)} = CW_s^{(k)} s$. The notation C is known as a constant value showing the average times of the replication factor, here C value is set to 99. We repeat the s^{th} training instance $r_s^{(k)}$ number of times to build a new accident data set. By using stacked autoencoder, road accident testing instance(x) can be trained and the accident rate prediction is illustrated as:

$$\hat{y} = \text{argmin } \hat{y} \epsilon_{[0, v_{max}]} \sum_{k=1}^{k} \alpha^{(k)} \; \zeta(|G^{(k)}{}_{(x)} - \hat{y}| - \delta) \quad (5)$$

Where v_{max} is depicted as maximum rate of accident on the particular area, $\hat{y}$ is known as accident rate of kerala to forecast which is expected as a number. The k trained simulations are active to create k number of the estimation.

Algorithm 1: Training the boosting algorithm for stacked autoencoder (SAE)

Require: $T = \{(x_s, y_s)\}_{s=1...s}$, δ, C.
Ensure: $\alpha^k, G^k(.), k = 1....k$

1. $K = 1$
2. $W_s^{(1)} = \frac{1}{s}$
3. While $k \leq K$ do
4. Replicate T according to $r_s^{(k)} = CW_s^{(k)} s$
5. Train and cross validate to choose the best deep architecture for the SAE $G^{(k)}(.)$ with the replicated samples.
6. Calculate the discriminative error $\epsilon^{(k)} = \frac{1}{2} \sum_{s=1}^{s} w_s^{(k)} [\zeta(|\; G^{(k)}(x_s) - y_s| - \delta) + 1]$
7. If $\epsilon^{(k)} \geq \frac{1}{2}$ then
8. Continue
9. End if
10. Calculate $\alpha^{(k)} = \frac{1}{2} \log \frac{1 - \epsilon^{(k)}}{\epsilon^{(k)}}$
11. Update the weight $W_s^{(k+1)} = \frac{w_s^{(k)}}{\zeta^{(k)}} e^{\alpha(k)} \zeta(|\; G^{(k)}(x_s) - y_s| - \delta)$
12. $k = k + 1$
13. end while

Then the researcher systematically itemize all the feasible accident rate starting from the value 0 to v_{max} to examine an optimal value $\hat{y}$. In this equation 5, the notation α^k is the prominence of the kth number of autoencoder, which is explained by using equation 3 as stated by the discriminative miscalculation of the kth number of AE. If discriminative error is large, then less significance the kth number stacked autoencoder achieves. If error between $\hat{y}$ and the estimation by the SAE is not exceed than value of δ, then the value of $\zeta(.)$ is (-1). So to minimize the equation 5, the optimal value $\hat{y}$ is estimated to encounter the predictions prepared by the stacked autoencoder of high significance as soon as possible. The process can be précised as follows:

IV. PERFORMANCE ANALYSIS

For safety designation and specially, identification of dangerous zones in network by ranking the sites by their accident rates, the model is also very helpful. Here iteration , batch size and learning rate are 500,32 and 0.9.

Table I. Experiment Result

	First Experiment	Second Experiment	Third Experiment
Sampling	Under	Over	Proposed
Classification	SVM	Linear Regression	Auto encoder
Positive Samples	145(7.8%)	1557(3.9%)	1421(0.27%)
Negative Samples	1702(92.2%)	38,443(96%)	522.710(99.73%)
Accuracy	73.63%	84.77%	99
True Positive rate		39.4%	40.83%

V. EXPERIMENT

Kerala has high accident rate because of high population density and use of high density of vehicles. Fig.2. helps to figure out the number of road accidents in state kerala and Fig.2. introduces about the accident rate in each month as well as Fig.4 throws light towards seasonal decomposition of the data set which is collected from UCI repository.

Fig.2. Total accident in state Kerala

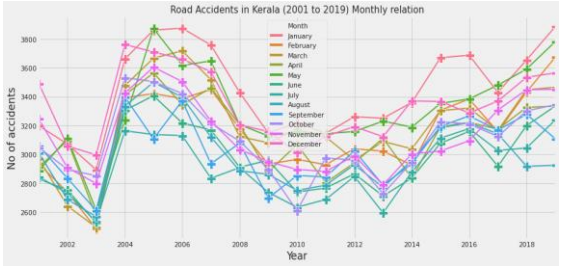

Fig.3. Accident rate in each month at kerala

Fig.5. describes about the performance analysis of the data set. With the analytical above figure we can able to picture out significant data as the circulation and as well as Auto correlation function (correlogram). The range upward the "0" has certain correlation in excess of the time series information and the range nearer to "1" establishes strongest correlation. Accident forecast in 2020 at kerala can be visualise from the Fig.7 and validation of forecasting model can be picture out from the Fig.6.

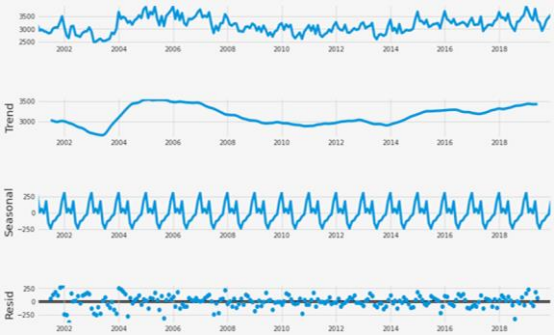

Fig 4. Seasonal Decomposition

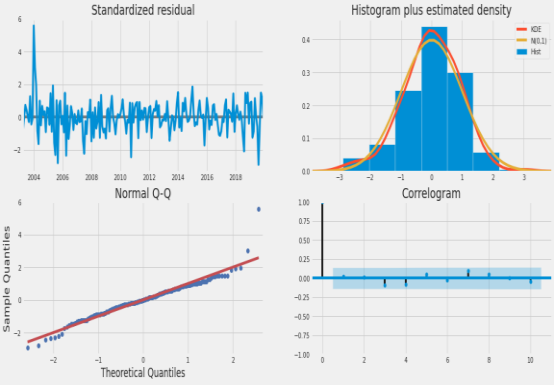

Fig.5. Performance Analysis of data set

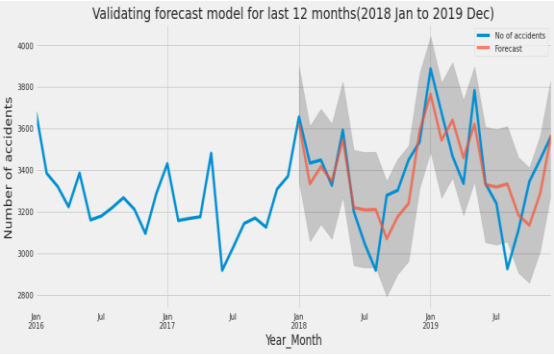

Fig.6.Validating forecast model of kerala

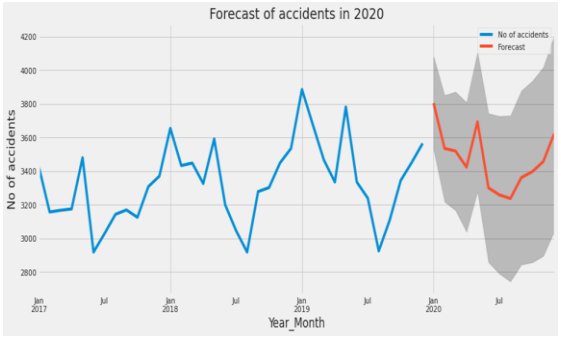

Fig.7. Forecast of accident in 2020 at kerala

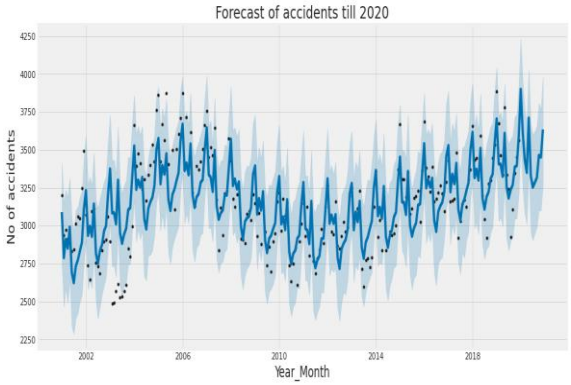

Fig.8. Predict the accidents till 2020

From Fig.8 and Fig. 9, we conclude that the deep blue line is predicting number of road accidents and black dotted lines are real number of road accidents then the blue shade(which is light) is providing 96% of sangfroid interval nearby the prediction. From 2020, the black dotted lines are not perceptible as it illustrates only about the future accident prediction in the state Kerala of country India.

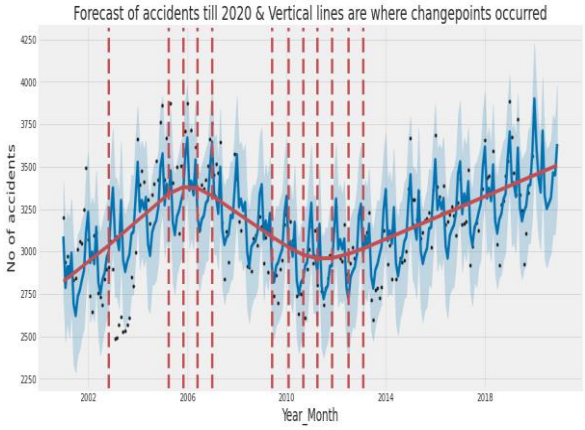

Fig.9. Predict the accidents till 2020

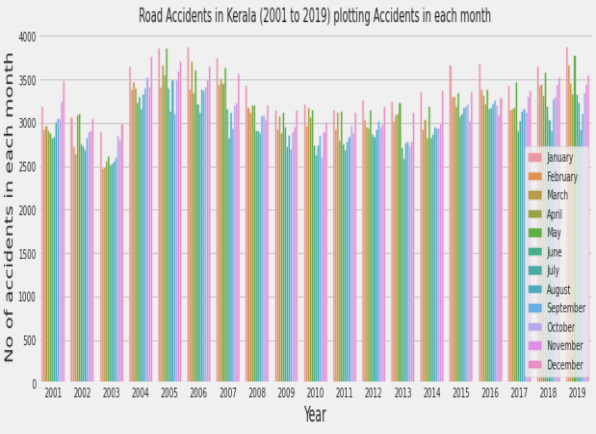

Fig.10. Plotting of accidents in each month

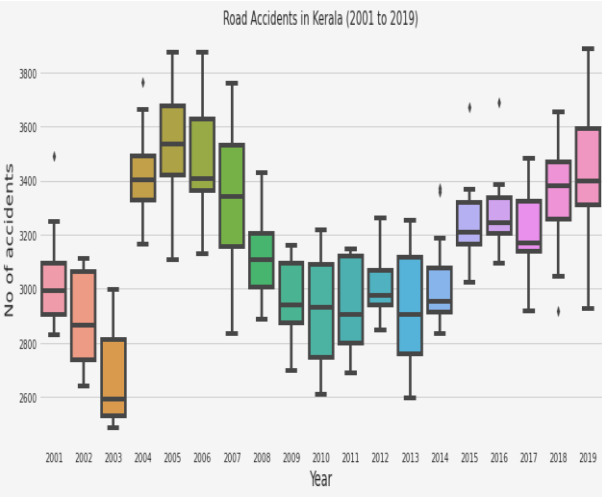

Fig.11.Box-plot analysis of kerala

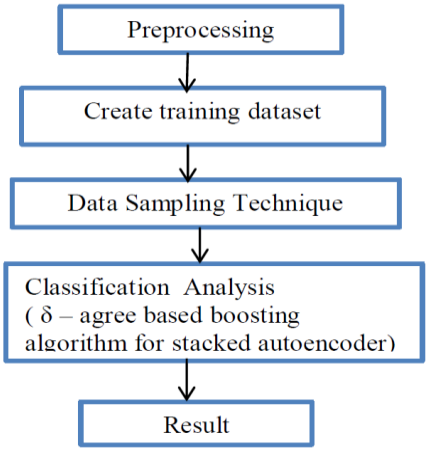

Fig.12. Steps of the algorithm

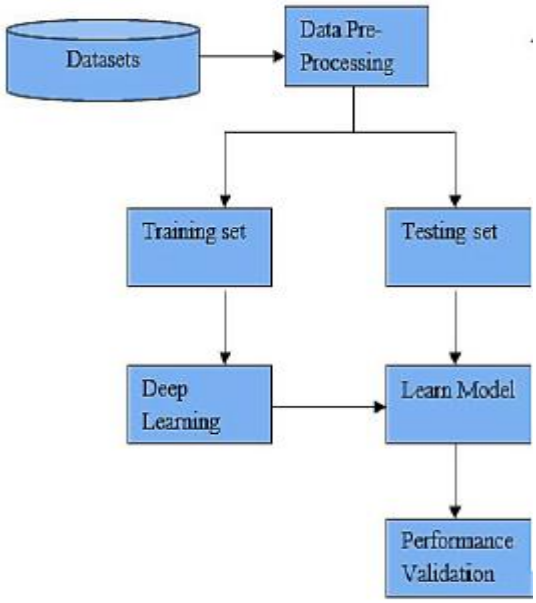

Fig.13. Fundamental Procedure of the road accident dataset (of city Kerala) handling

VI. CONCLUSION AND FUTURE WORK

Road accident is a very serious problem in metropolitan cities, effecting people every single day, stressing to go early for their work. Improper signal timing increases delay at intersection mainly during peak hours of the working day. Delay and accident rate are crucial parameter to measure the performance level of signalized intersection. Proper provision of cycle length is required according to the capacity of the approach to reduce the delay. This will help not only to lower excessive fuel consumption and air pollution, but also improve the performance of intersection and loss of time due to delay. Extensive studies have been done on delay at signalized road intersection under mixed and under-saturated traffic conditions. By using multichannel scheme, accident rate can be reduced which is very essential. Many improvements should be tended to address in the future works. Initial, a proficient over-sampling technique ought to be built up. Second, this procedure time and precision of investigated results ought to be contrasted with other existing work. Next, many testing strategies will be expected to assess the analysed outcomes. For future work, other efficient research is essential to provide a smart algorithm to solve the problem of real time data processing and different researches are essential to detect another new technique to solve the issues of real-time data.

REFERENCE

[1] Formosa, N., Quddus, M., Ison, S., Abdel-Aty, M., & Yuan, J. (2020). Predicting real-time traffic conflicts using deep learning. Accident Analysis & Prevention,136,105429. doi:10.1016/j.aap.2019.105429

[2] Wang, J., Gu, Q., Wu, J., Liu, G., & Xiong, Z. (2016). Traffic Speed Prediction and Congestion Source Exploration: A Deep Learning Method. 2016 IEEE 16th International Conference on Data Mining (ICDM). doi:10.1109/icdm.2016.0061.

[3] Pradhan, B., & Ibrahim Sameen, M. (2019). Predicting Injury Severity of Road Traffic Accidents Using a Hybrid Extreme Gradient Boosting and Deep Neural Network Approach. Urbanization and Its Impact in Contemporary China, 119–127. doi:10.1007/978-3-030-10374-3_10 .

[4] Singh, G., Pal, M., Yadav, Y., & Singla, T. (2020). Deep neural network-based predictive modeling of road accidents. Neural Computing and Applications. doi:10.1007/s00521-019-04695-8

[5] Nallaperuma, D., Nawaratne, R., Bandaragoda, T., Adikari, A., Nguyen, S., Kempitiya, T. Pothuhera, D. (2019). Online Incremental Machine Learning Platform for Big Data-Driven Smart Traffic Management. IEEE Transactions on Intelligent Transportation Systems, 1–12. doi:10.1109/tits.2019.2924883 .

[6] Zhu, L., Yu, F. R., Wang, Y., Ning, B., & Tang, T. (2018). Big Data Analytics in Intelligent Transportation Systems: A Survey. IEEE Transactions on Intelligent Transportation Systems, 1–16. doi:10.1109/tits.2018.2815678 .

[7] P. S. Praveen, R. Ashalatha. "Passenger Car Equivalency Factors Under Platooning Conditions", Transportation in Developing Economies, 2019.

[8] Pradhan, B., & Ibrahim Sameen, M. (2019). Forecasting Severity of Motorcycle Crashes Using Transfer Learning. Urbanization and Its Impact in Contemporary China, 141–157. doi:10.1007/978-3-030-10374-3_12.

[9] Lv, Y., Duan, Y., Kang, W., Li, Z., & Wang, F.-Y. (2014). Traffic Flow Prediction With Big Data: A Deep Learning Approach. IEEE Transactions on Intelligent Transportation Systems, 1–9. doi:10.1109/tits.2014.2345663 .

[10] Yash R. Dasani, Monicaba Vala, Bindiya Patel."Chapter 37 Estimation of Dynamic Equivalency Factor Under Heterogeneous Traffic Condition on Urban Arterial Road—A Case Study of Porbandar City", Springer Science and Business Media LLC, 2020.

[11] Haibo He, & Garcia, E. A. (2009). Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. doi:10.1109/tkde.2008.239.

[12] Alkouz, B., & Al Aghbari, Z. (2020). SNSJam: Road traffic analysis and prediction by fusing data from multiple social networks. Information Processing & Management, 57(1), 102139. doi:10.1016/j.ipm.2019.102139.

[13] Santhosh A., Sam E., Bindhu B.K. (2020) Pedestrian Accident Prediction Modelling—A Case Study in Thiruvananthapuram City. In: Mathew T., Joshi G., Velaga N., Arkatkar S. (eds) Transportation Research. Lecture Notes in Civil Engineering, vol 45. Springer, Singapore.

[14] F. Barboza, H. Kimura, and E. Altman, "Machine learning models and bankruptcy prediction," Expert Systems with Applications, vol. 83, pp. 405–417, 2017.

[15] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 42, no. 4, pp. 463–484, 2012.

[16] Li, P., Abdel-Aty, M., & Yuan, J. (2020). Real-time crash risk prediction on arterials based on LSTM-CNN. Accident Analysis & Prevention, 135, 105371.

[17] Seong-hun Park, Sung-min Kim, Young-guk Ha."Highway traffic accident prediction using VDS big data analysis", The Journal of Supercomputing, 2016.

[18] Official webportal of kerala police. http://www.keralapolice.org/publicinformation/crime-statistics/road-accident.