International Journal of
Advances in Scientific Research and Engineering (ijasre)

E-ISSN : 2454-8006

DOI: 10.31695/IJASRE.2020.33771

Volume 6, Issue 3
March - 2020

Predicting On-road Traffic Congestion from Public Transport GPS Data

Thura Kyaw[1], Nyein Nyein Oo[2] and Win Zaw[3]

[1-3]Department of Computer Engineering and Information Technology

Yangon Technological University

Insein, Yangon

Myanmar

ABSTRACT

This paper predicts traffic congestion of urban road network by building machine learning models using public transport GPS data. The bus GPS data are collected over 18 months started in September 2017 and ended in January 2019. After the data cleaning and data processing are carried out, time series data analysis is performed on these data. Travel Speed Estimation and Traffic Jam Prediction Model are two major components of this work. Firstly, road network structure and GPS data sets are inputted to the Travel Speed Estimation Model to get estimated travel speed for every road segment in uniform time windows of a day. The next step is to set up Traffic Congestion Prediction Model from estimated average travel speed and current GPS data from the buses. Decision Trees, Random Forest Classifiers and ExtraTree Classifiers algorithms have been successfully applied and validated with K-Fold cross validation yielding high prediction accuracy to a specific bus route in Yangon, Myanmar.

Key Words: *GPS Data, Road Segments, Travel Speed Estimation Model, Traffic Jam Prediction Model, high prediction accuracy, Classifier Algorithms.*

1. INTRODUCTION

Traffic congestion is one of the most important problems to be solved in most mega cities around the world. People are getting jammed in long waiting queue of vehicles in their ways to workplaces and going back to home. The more vehicle population leads to not only the more congested roads within the cities also results more polluted air by the vast amount of CO_2 emission. Solving the traffic problem is becoming the active domain for the scholars today. There exists several models and approaches suggested and implemented by researchers to anticipate traffic related problems.

Traffic congestion usually occurs by two reasons. Firstly, when the number of on-road vehicle exceeds the capacity of roadway, the route becomes saturated and bottlenecks happen. This condition can be the most frequently and take place regularly when people go to work and back to home, as well as in holiday trips. Secondly, traffic congestion can be triggered by road incidents such as vehicle breakdowns, accidents, rain, snow and other environmental factors. These factors lead to a slower vehicle movementand high volume and compact density of vehicles on the road which ends up in a traffic jam [1]. As the population increased the number of vehicles will also increased, the more new roads and highways are built but it still can’t solve the traffic congestion problems. These improvements are both expensive and time consuming. Different solutions have been attempted in cities around the world, with varying degrees of success [2]. Vehicular traffic is an extremely complex dynamic process associated with the behavior of many-particle systems [3]. The complexity of vehicular traffic is due to nonlinear interactions between the three main dynamic processes , travel decision behavior which determines traffic demand, routing of vehicles in a traffic network and traffic congestion occurrence within the network.

Figure 1 clarifies the correlations between travel decision, route selection and traffic congestion occurrences in road network. The traffic related problems are very complicated to solve from a single point of view. The standardized traffic analysis can generally categorize into three different geographical scopes: Microscopic, Mesoscopic and Macroscopic area. Microscopic traffic analysis is focusing on individual vehicle, Mesoscopic is analyzing at single junction or road segment and macroscopic analysis is for metropolitan area. In this paper, the area of interest for traffic analysis is one of the busiest roads in Yangon City.

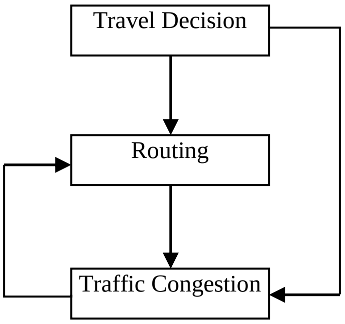


Figure.1 Explanation of Complexity of Vehicular Traffic

The congestion information is important for road users to make travel decision and routing pattern changes on their journeys to and from work on daily routines. Traffic congestion prediction needs traffic data from reliable sources, pre-processing of collected data for the consistency, integrity and clarity of data. There exist several approaches in data traffic collection process. The most popular road side traffic counters are pneumatic tubes, inductive loops, weigh-in-motion sensor types, micro-millimeter wave radar detectors and traffic counters and video recorders [4] to collect on-road vehicle data. These data collection methods need a high initial investment cost as well as maintenance procedures to get minimal tolerance values of traffic counting, speed measuring results. The alternative method for data collection is installing GPS sensors inside vehicles and logging the time-position data in a server machine. This type of data collection is more preferable than the first approach that uses road side traffic counters.

In our research, we collect the GPS trajectories of buses running along route 21 of Yangon Bus System (YBS). The route starts from the rural area of the city, goes pass through eleven townships and ends at the business district of Yangon city. Yangon Bus System (YBS) is the only public transportation system of Yangon operating with over 4000 running buses on 101 different routes. The selected bus route runs passing through the most congested area of Yangon City. We apply state of the art machine learning models on the collected data to build travel speed estimation model for Yangon City.

This paper is organized in the following sections. Section I introduces the traffic congestion problem and solving methods. Section II is the overview of related works. Section III explains the traffic congestion prediction model in two distinct procedures. Section IV exemplifies the results of conducted experiment and we discuss, conclude and plan our future work in Section V.

2. RELATED WORKS

Many researchers discovered and implemented several traffic congestion prediction models, traffic volume estimation models and proposed various traffic related applications. Xiaojun Shen [5] proposed a model using learn quantization vector (LQV) to predict traffic congestion using traffic parameters such as speed, traffic volume and traffic occupancy on the roads. Their model required to input traffic flow data and predicted the road traffic congestion situations. Another research related to GPS trajectory based drive path prediction was proposed by Ekim Yurtsever et.al. [6], their approach used collected GPS data to get history trajectory of drivers and drive path optimization compared to road curvatures.

Xianyuan et.al [7] built a model for prediction of traffic volume in citywide scale for Beijing from the collected GPS data. They used well implemented traffic flow theories in combination with machine learning models to predict travel speed of almost every road including all types of roads in Beijing the highways, artery roads and even the minor roads from the extracted features of GPS data. The relation of speed and traffic volume is used to classify traffic delay features and then used to predict travel speed for the road whole network. Finally they predicted city wide traffic volume using machine learning models.

Chinmaya Samal and others [8] proposed a multi-model approach for urban traffic speed prediction which combines the operations on historic data, the time-position information, the weather data of current time, and driver behaviors. They stated that the data sparse issue is a huge problem to get actual travel speed of probe vehicles and they used multi-model approach to improve prediction accuracy. We take this fact into account we extract the traffic delay related features caused nearby places along the route and used to predict travel speed.

In our model, we collect the GPS data from probe vehicles. Then we extract the features that can affect traffic delay from the nearby places that exist along every road segment. Using these features we estimate average travel speed of every road segment along the bus route. Then, we build a traffic history data for Yangon-Insein Road by computing the speed factor of road segments which is the ratio of the estimated speed and speed limit of the road segments. Finally, we predict the traffic congestion level of

DOI: 10.31695/IJASRE.2020.33771

the bus route by building machine learning models using Decision trees, Random Forest Classifiers and ExtraTree Classifiers algorithms.

3. TRAVEL SPEED ESTIMATION MODEL

The traffic congestion prediction model is depicted in Figure 2. The model consists of the Travel Speed Estimation Model (TSEM) and Traffic Congestion Prediction Model (TCPM). Firstly, raw GPS data collected from the running buses are preprocessed to get cleaned data. Road Network Data which includes location of bus stops along the bus route is the other data set for travel speed estimation model. The estimated travel speed for every road segments is stored in a matrix for further processing. The live GPS data from the buses and the output from the Speed Estimation Model are inputted into the Classification Algorithms to predict the congestion level of road segments. The classified results are validated with K-Fold cross validation scheme and the final result is displayed o the digital map of Yangon to provide the congestion information to the road users.

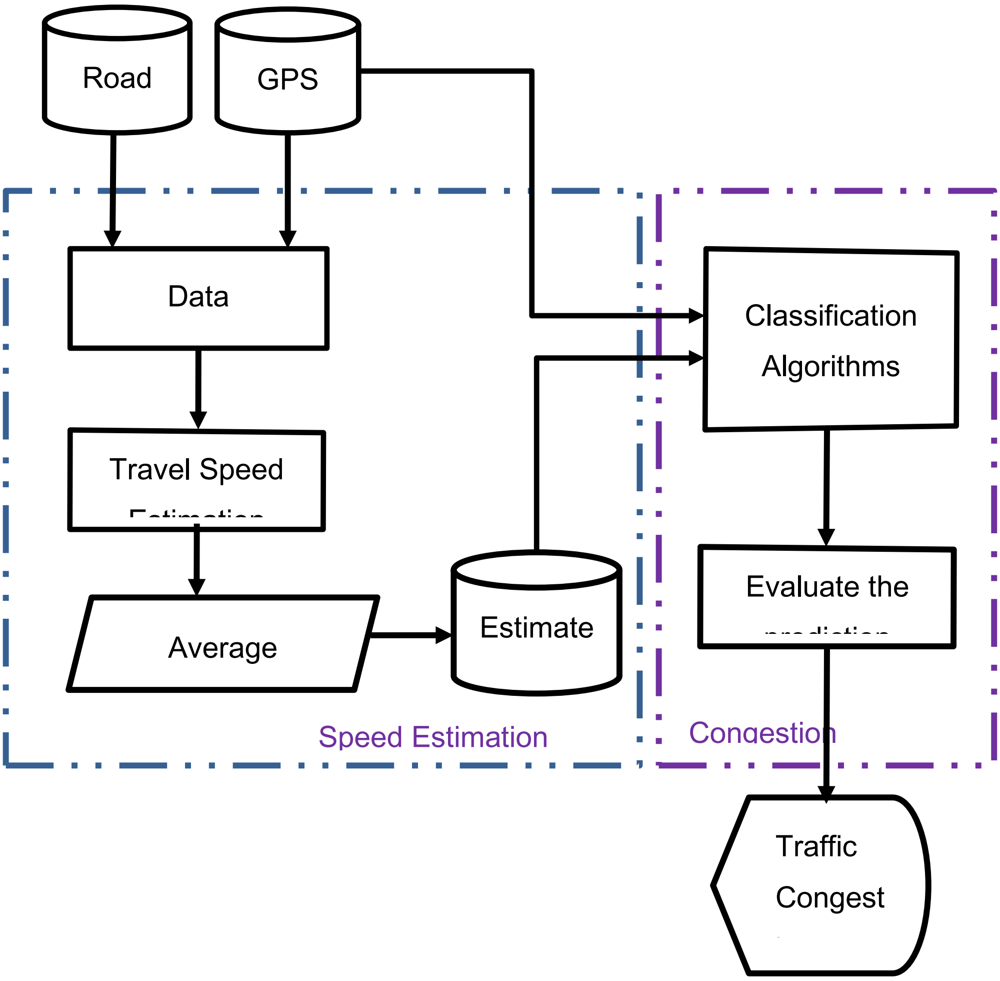


Figure.2 Traffic Congestion Prediction Model

3.1. GPS Data Collection and Preprocessing

Data collection plays in an important role in every data analysis model. The integrity and clarity of collected data are essential properties to build a good data set. YBS bus no. 21 provides their bus tracking data for our research. They collect GPS data by installing Sino Track GPS trackers in the buses and trace the travel patterns, fuel consumption rate, mileage, engine on/off status and location of buses in every 20 seconds for 24 hours a day.

GPS data by installing Sino Track GPS trackers in the buses and trace the travel patterns, fuel consumption rate, mileage, engine on/off status and location of buses in every 20 seconds for 24 hours a day. We can access the live GPS data for every bus that services along the bus No.21 route as well as history data for last three months. The next step is to build a well-structured data set from the collected trajectories. The preprocessing of GPS data includes removing outlier points, dropping unwanted features by dimension reduction methods and selecting important features for Travel Speed Estimation Model (TSEM). Preprocessed trajectory can be expressed as:

$$T_i \in \{ t_i, lat_i, lon_i , dir_i, v_i\} \tag{1}$$

DOI: 10.31695/IJASRE.2020.33771

where, T_i is the specific trajectory from a bus, t stands for time of the day in local time zone, position data (latitude, longitude) information is represented by (lat_i, lon_i), dir_i is the direction of the bus and v_i represents current speed of the bus. After preprocessing of raw GPS data, the data set is ready for analysis.

3.2. Road Network Data

The map is a collection of data that represents the detail of a particular location with detailed major roads and other point of interests. The map data can be divided into smaller data set. Suppose that each road R is a set of road segments e: $R \epsilon \{e_1, e_2, \ldots, e_n\}$ and each road segment consists of start and end points: $e \epsilon \{ lat_1, lon_1, lat_2, lon_2 \}$where lat_1,lon_1 represents start point and lat_2,lon_2 is end point of the road segment.

Table1. Bus Stop Name and Location along the Route

BUS_STOP_NAME	Latitude	Longitude
A Nout Paing Takke Tho	16.948604	96.013681
Awine	16.922083	96.05449
Mile Hnat-sae	16.922576	96.06975
Ayar Myay	16.930438	96.082753
Tit-Koe-Koe	16.931731	96.085049
Sae-Lay Lan Sone	16.933209	96.087753
Butar Zay	16.933733	96.093795
Danyin Gone Lan Sone	16.933805	96.101891
Bo Chan	16.929699	96.100649

Table 1 is the sample of bus stop location and names along the bus route. The interested bus route consists of 18 signalized junctions and 48 bus stops. We have to segment the bus route by the collected position (*lat, lon*) of every bus stop along the way.

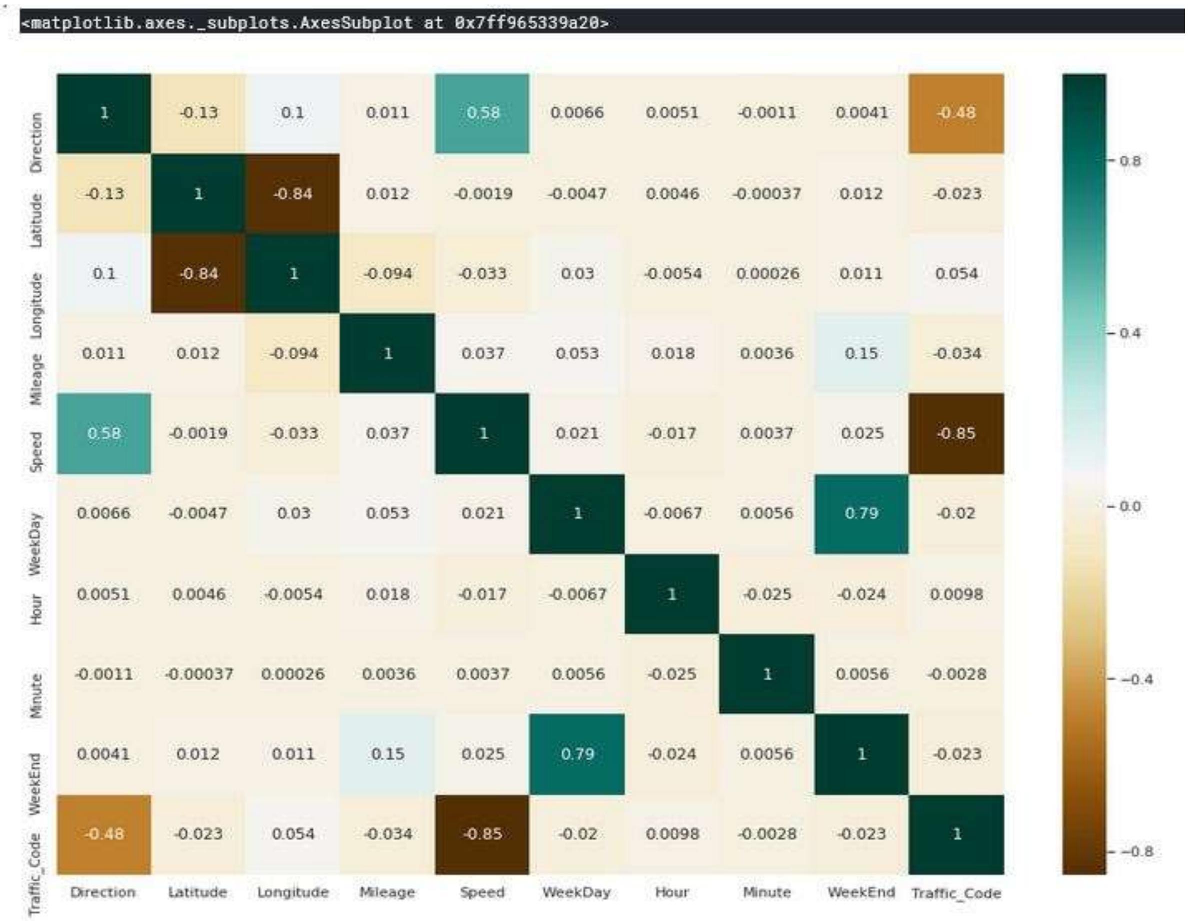
3.3. Travel Speed Estimation Model

The Speed Estimation Model is built using GPS data from buses and the road network data from the bus stop location data. The very first step is placing the bus stops of YBS 21 onto the Open Street Map of Yangon. In order to calculate traffic trajectory history, the bus routes has been divided into 48 segments according to Greenwich coordinates transformed into Cartesian coordinates and a piecewise polynomial function representing each group has been fitted with least squares method. Due to limited number of running buses that can track in a particular time window (e.g. 8:00 – 8:05), there exist some road segments with no buses running on them. In order to solve this problem, we have to overlap the historical GPS trajectories to get the average travel speed of each road segment. The historical data such as the average speed, standard deviation, minimum speed and maximum speed of a road segment e are calculated at every 5 minutes time interval and recorded as statistical data. The speed is calculated and updated in regular time interval from the statistical data, live data and traffic delay factors of each segment.

4. TRAFFIC CONGESTION PREDICTION MODEL

Once the average travel speed for the whole bus route is stored in a matrix on segment by segment basis, the congestion prediction for the bus route is almost ready. The raw GPS data are simplified by a preliminary analysis in section 3.1 and estimated travel speed for the bus route is calculated in section 3.3. This section is about the prediction of traffic congestion using tree based classifiers namely Decision Trees, Random Forest Classifiers and ExtraTree Classifiers algorithms. GPS data from buses collected for 24 months and estimated travel speed from Travel Speed Estimation Model are used as train and test data to fit into the models. The training data includes Time, Speed, Direction, Mileage and Location (Latitude, Longitude) information of buses along with the level of congestion. Since the correlation of predictor variables to the target variable is very important for the classification, we have conducted exploratory analysis of training data to ensure only the important features are selected to apply the machine learning models.

DOI: 10.31695/IJASRE.2020.33771



```
<matplotlib.axes._subplots.AxesSubplot at 0x7ff965339a20>
```

	Direction	Latitude	Longitude	Mileage	Speed	WeekDay	Hour	Minute	WeekEnd	Traffic_Code
Direction	1	-0.13	0.1	0.011	0.58	0.0066	0.0051	-0.0011	0.0041	-0.48
Latitude	-0.13	1	-0.84	0.012	-0.0019	-0.0047	0.0046	-0.00037	0.012	-0.023
Longitude	0.1	-0.84	1	-0.094	-0.033	0.03	-0.0054	0.00026	0.011	0.054
Mileage	0.011	0.012	-0.094	1	0.037	0.053	0.018	0.0036	0.15	-0.034
Speed	0.58	-0.0019	-0.033	0.037	1	0.021	-0.017	0.0037	0.025	-0.85
WeekDay	0.0066	-0.0047	0.03	0.053	0.021	1	-0.0067	0.0056	0.79	-0.02
Hour	0.0051	0.0046	-0.0054	0.018	-0.017	-0.0067	1	-0.025	-0.024	0.0098
Minute	-0.0011	-0.00037	0.00026	0.0036	0.0037	0.0056	-0.025	1	0.0056	-0.0028
WeekEnd	0.0041	0.012	0.011	0.15	0.025	0.79	-0.024	0.0056	1	-0.023
Traffic_Code	-0.48	-0.023	0.054	-0.034	-0.85	-0.02	0.0098	-0.0028	-0.023	1

Figure.3 Correlation Matrix of Predictor Variables to the Target Variable

According to Figure 3, traffic code is the target variable and the highly correlated predictor variables are Speed of the buses, Direction of the buses and weekday which was generated from time stamp variable.

Figure 4 explores the histogram patterns of Speed Vs Traffic Code, Mileage Vs Traffic Code and Direction Vs Traffic Code. The traffic variable is coded into four different colors representing heavy traffic by red, moderate traffic by orange, light traffic by Yellow and finally free flow traffic by green colors. We can clearly make a note of heavy traffic (red) is the most dominant one to the other traffic colors in all histograms.

DOI: 10.31695/IJASRE.2020.33771

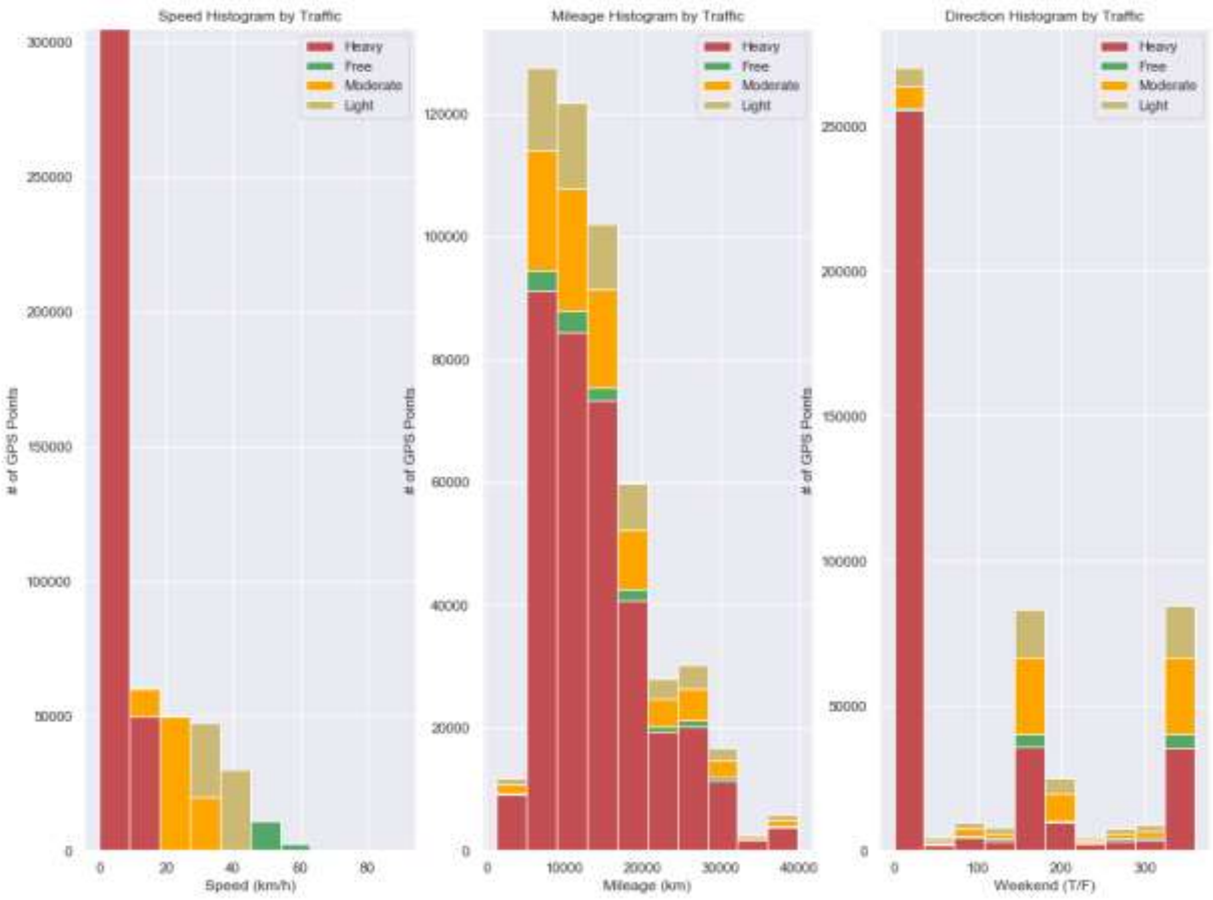


Figure.4 Histograms of Traffic Code by Speed, Mileage and Direction Variables

5. MODELS TUNINGS AND RESULTS

Since the art of data science relies on the performance of the machine learning algorithms, it is important to get best fitted model for the cleaned and preprocessed data. In our model, all the training and testing data are spitted with 10-fold cross validation scheme and then fitted into classification algorithms.

5.1. Parameter Tuning

It is very important to tune the machine learning models with hyper parameters to get the best results of prediction. In our models we adjusted the parameters of tree based algorithms to a variety of values to get the best fitted model with the highest prediction accuracy.

Table2. Classification Algorithms and Tuned Parameters

Algorithms	Parameters
DecisionTreeClassifier	{'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features': None, 'max_leaf_nodes': None, 'min_impurity_decrease': 0.0, 'min_impurity_split': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}
RandomForestClassifier	{'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features': 'auto', 'max_leaf_nodes': None, 'min_impurity_decrease': 0.0, 'min_impurity_split': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 'warn', 'n_jobs': None, 'oob_score': False, 'random_state': None, 'verbose': 0, 'warm_start': False}
ExtraTreesClassifier	{'bootstrap': False, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features': 'auto', 'max_leaf_nodes': None, 'min_impurity_decrease': 0.0, 'min_impurity_split': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 'warn', 'n_jobs': None, 'oob_score': False, 'random_state': None, 'verbose': 0, 'warm_start': False}

DOI: 10.31695/IJASRE.2020.33771

The algorithms used for traffic jam classification and their best accuracy yielding parameters are expressed in Table 2.

5.2. Results of Classification

After setting up the models, fitting training and testing data for the several computing hours and observing the final results, the prediction accuracies of all algorithms are stored in a csv file.

Table3. Results of Prediction Models

Algorithms	Train Accuracy Mean	Test Accuracy Mean	Test Accuracy 3*STD	Running Time
DecisionTreeClassifier	0.993659486	0.982540843	0.000480309	1.776602
RandomForestClassifier	0.993427049	0.980264893	0.000941144	7.204553
ExtraTreesClassifier	0.993659486	0.972943519	0.003633083	4.784269

Table 3 compares the train accuracy, test accuracy, standard deviations and running time of three different traffic congestion prediction models.

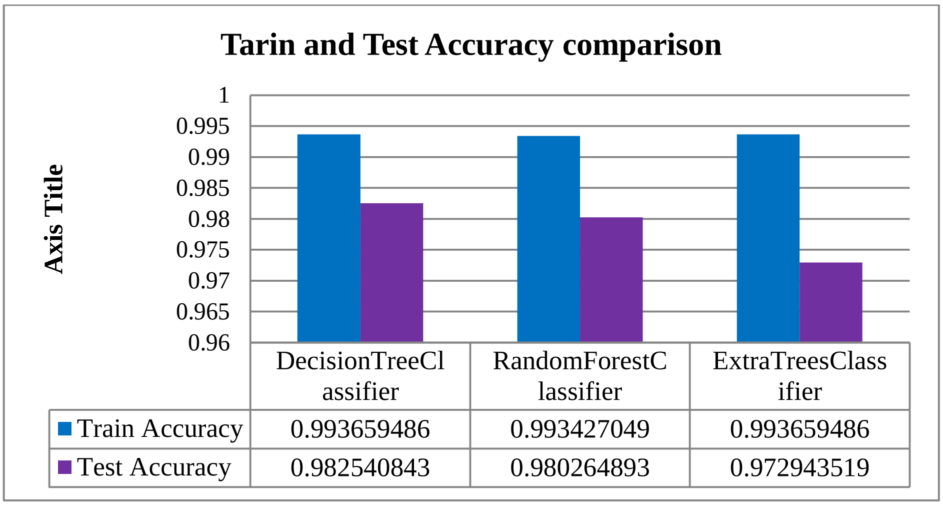


Figure.5 Comparison of Train and Test Accuracies in Mean Values

In accordance with figure 5, all the classifier algorithms give high accuracies in training with the value of nearly 99.4 %, but in the testing phase, the accuracy of Decision tree is the best with 98.25 % while Random Forest gets 98 % and ExtraTreesClassifier with the lowest accuracy of 97.3%.

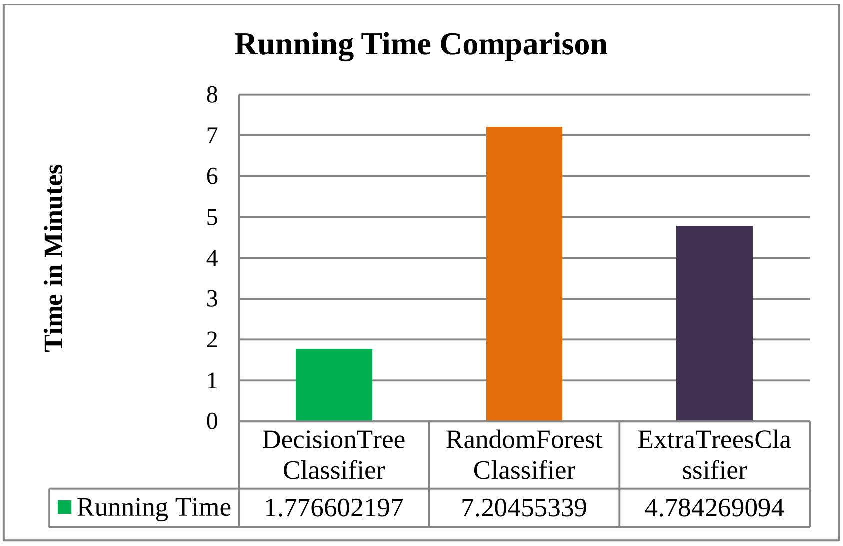


Figure.6 Comparison of Running Time in Minutes

DOI: 10.31695/IJASRE.2020.33771

As illustrated in Figure 6, the most efficient algorithm is Decision Tree with lowest running time with highest accuracy among all the three classifiers. The random Forest is slowest one with over 7.2 minutes with to get the high accuracy and the Extra Trees Algorithm takes 4.78 minutes which only yield lowest accuracy score compared to the other algorithms.

5.3. Displaying Predicted Results

After the models are tested and validated with K-fold cross validation with k value of 10, the resulting predicted traffic code values are to be displayed on the digital map of Yangon.

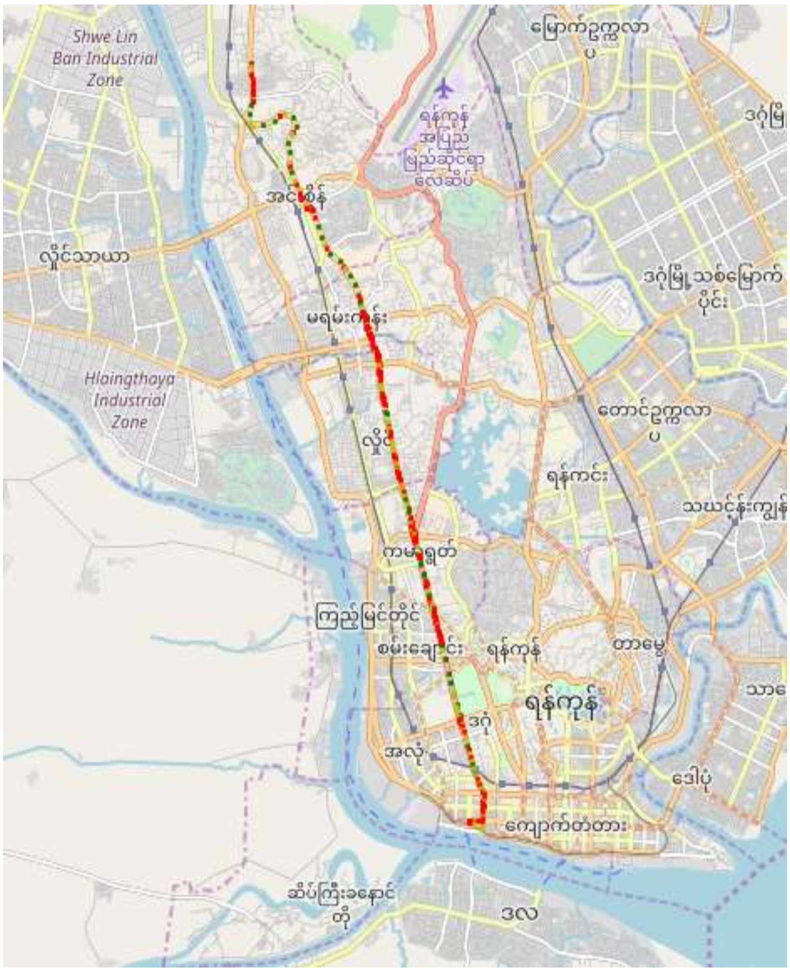


Figure.7 Predicted Results Displayed on the Color Coded Traffic Map.

According to Figure 7, the entire bus route is coded with red, orange, yellow and green colors representing heavy traffic, moderate traffic, low traffic and free flow of traffic degrees respectively. The road users can decide which route to choose in their journeys within city boundary.

6. DISCUSSION AND CONCLUSION

In this paper travel speed and traffic congestion level have been predicted using data collected by YBS 21, especially at the Yangon- Insein road and Pyay road of Yangon, Myanamr. The extraction of these traffic parameters has been used to predict traffic jams early enough to avoid retention at that point, minimizing response time to this event and giving alternatives to the circulation. To achieve this, it has be successfully applied the well known tree based algorithms and obtained results with an error ranging from 97% to 98%, depending on the anticipation time of the prediction made. Moreover, the models have been successfully validated with K-Fold cross validation.

ACKNOWLEDGMENT

The authors wish to thank Power 11 Co. Ltd., for supporting GPS data of YBS bus route 21 and the support needed to conduct this study.

REFERENCES

1. K. Mahmud, K. Gope, and S. M. R. Chowdhury. Possible causes and solutions of traffic jam and their impact on the economy of Dhaka city. Journal of Management and Sustainability, 2(2):p112, 2012.
2. R. C. Carlson, I. Papamichail, M. Papageorgiou, and A. Messmer. Optimal main stream traffic flow control of large scale motorway networks. Transportation Research Part C: Emerging Technologies, 18:193–212, 2010.
3. B.S. Kerner, *Introduction to Modern Traffic Flow Theory and Control*, DOI 10.1007/978-3-642-02605-8 1, c Springer-Verlag Berlin Heidelberg, 2009, p 45.
4. Guideline No.4 Axle Load Surveys, Botswana, *Traffic Data Collection and Analysis*, Roads Department, Ministry of Works and Transport, Private Bag 0026 Gaborone, Botswana, Feb. 2004, pp. 15–18.
5. Xiaojun Shenand Jun Chen, *"Study on Prediction of Traffic Congestion Based on LVQ Neural Network*", in *Proc*. 2009 International Conference on Measuring Technology and Mechatronics Automation, 2009, pp. 318–321.
6. Ekim Yurtsever, Kazuya Takeda and Chiyomi Miyajima, *"Traffic Trajectory History and Drive Path Generation using GPS Data Cloud"*, in *Proc.* IEEE Intelligent Vehicles Symposium (IV),2015, pp.229–232
7. Xianyuan Zhan, Yu Zheng, Senior Member, IEEE, Xiuwen Yi, and Sa tish V. Ukkusuri, *"Citywide Traffic Volume Estimation Using Trajectory Data"*, IEEE Transactions on Knowledge and Data Engineering, Vol. 29, No. 2, February 2017,pp.229–238
8. *Chinmaya Samal, Fangzhou Sun, Abhishek Dubey "SpeedPro: A Predictive Multi-Model Approach for Urban Traffic Speed Estimation", IEEE International Conference on Smart Computing (SMARTCOMP), 2017, pp.110–114*
9. *Xianyuan Zhan, Yu Zheng, Senior Member, IEEE, Xiuwen Yi, and Sa tish V. Ukkusuri.: Citywide Traffic Volume Estimation Using Trajectory Data, IEEE Transactions on Knowledge and Data Engineering, Vol. 29, No. 2, February, (2017).*

DOI: 10.31695/IJASRE.2020.33771