Traffic Flow Forecasting in Intelligent Transportation Systems Prediction Using Machine Learning

Mohammad Naveed Hossain
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
naveedhossain99@gmail.com

Nafim Ahmed
Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
recentnafimahmed@gmail.com

S. M. Wazid Ullah
Information and Communication Technology
Mawlana Bhashani Science&Technology University
Dhaka, Bangladesh
wazidullahmurad@gmail.com

***Abstract*—Globally, intelligent transportation systems utilize traffic predictions. Traffic congestion, route planning, and vehicle dispatching all benefit from accurate traffic forecasts. The road system's changing geographical and temporal dependencies complicate the problem. In recent years, traffic forecasting has improved thanks to research, particularly deep learning. We investigate traffic predictions for Dhaka based on machine learning and deep learning techniques. The classification of existing traffic prediction methods comes first. To enable academics, we aggregate and arrange commonly used public datasets. We undertake comprehensive experiments on a publicly accessible real-world dataset to compare and contrast diverse methodologies. The contribution of the third section is automated approaches for traffic forecasting. In closing, we discuss some of the outstanding questions.**

***Index Terms*—Traffic Prediction, Machine Learning, Deep Learning, Intelligent Transport System (ITS).**

I. Introduction

Machine Learning (ML) is a key AI subfield. Machine learning is a growing topic in transportation engineering, especially for traffic prediction. Congestion hurts the economy indirectly or directly. Gridlock consumes time and gas. Traffic congestion is an issue for all socioeconomic strata. Hence a small-scale traffic forecast is needed. Economic growth requires road user convenience first. Lack of congestion allows this. Traffic forecasting estimates future traffic. Reduced pollution helps the economy. Government invests in ITS to address these concerns. This article explores ML and Python3 models. Traffic flow prediction provides quick traffic estimates. Drivers do not know what is causing today's gridlock. This study forecasts traffic. For this project, finding it easier to use python using a command prompt window. 10-section paper. Introduction, Traffic Prediction Purpose, Problem Statement, Related Work, [1] [3] Overview, Methodology, Software Implementation, and Conclusion.

Many traffic reports are real-time, but it is not advantageous and accessible to many users because we need to plan our route. During working days, we need daily or hourly traffic statistics, but when traffic congestion develops, we need real-time traffic prediction. Congestion has many causes. That can be predicted by comparing one year's data. If traffic is heavy, it can be predicted by studying the same period in last year's data set. Traffic gridlock worsens as fuel prices rise. This prediction provides real-time traffic congestion data. Complex and out-of-control metropolitan traffic makes such systems insufficient for prediction. ITS relies heavily on traffic flow forecasting. [4] [5]

Machine learning with pandas, os, NumPy, and matplotlib. pyplot is used to forecast traffic flows and thereby lessen gridlock. Congestion can be managed and reduced if this is implemented. One-hour data allows users to track traffic patterns and congestion from morning to night. Users can see the forecast for the routes they want to take. The accuracy of traffic predictions can be seen by comparing the mean square errors from one year to the current year. The traffic forecaster also provides information on the volume of traffic.

II. Previous Works

This research is based on previous research. Among them few are mentioned below:

Researchers focused on GPS Tracking for predicting traffic jams. The researcher tried to detect the vehicle's speed based on the speed. Tried to detect traffic jams. In a traffic jam, traffic moves at a low speed. And the researcher tried to predict the speed, and by the specific, the researcher tried to predict the traffic jam. [14]

The researcher tried to predict traffic jams by establishing communication between the vehicles. In this system, vehicles will communicate between them and detect traffic jams. [15]

The researcher tried to apply an Artificial neural network to predict road blockage. By predicting the road blockage, the researcher tried to reduce road blockage, and in this way, he wanted to predict traffic jams. [16]

The researcher tried to predict the traffic jam by establishing mobile communication between vehicles. In this system, every vehicle needs to contain sensors, and they will share the sensor data between them. Predicting Traffic Phases from Car Sensor Data using Machine Learning. [17]

The researcher predicts traffic jams by using the vehicle's

sensor. The vehicles collect sensor data in this system and use the controller area network to predict traffic jams. [18]

III. Data Source

Traffic is either static or dynamic. Camera and sensor data. Vehicle GPS data. Real-time monitoring of activities based on data from traffic sensors Faulty sensors. Examine the sensor to see if there is a short in it. Sensors detect places. The PMS system monitors vehicular traffic movement in California's most populated cities. Extensive exploitation of I-5 5-minute San Diego data. blufaxcloud GBTSE (TIGER). Network-wide probe data. [10] Roadways. Data probes network-wide. 20,000 Beijing taxis. Passenger status, latitude, longitude. It is updated anywhere from once every 10 seconds and once every five minutes. We obtained data not just from the PVD but also from the GPS in the bus. The investigation was done on the variations. Probe-matching. Data-driven. The data obtained from the city probe are unsuitable for reproducing network configurations. There is a set of coordinates available for Beijing. Conduct research on the information using the city model. The numbers that are related to tolling and TA are considered to be more dependable. The statistics on toll roads are very seldom brought up to date. Monitoring people's mobile phones is not an invasive practice by any stretch of the imagination. This dataset makes it seem like only one kind of car. Moving about on foot generates irregularities in the network. Driver with a poor track record; the findings of a recent survey. [11]

IV. Applied Methods

A. KNN

KNN has proved useful for regression classification and prediction. It's used to categorize business difficulties. KNN requires all training and classification data. Hence it's sluggish. KNN is non-parametric since it makes no data-specific assumptions. KNN makes intelligent estimates about new data by comparing it to the original "training" set. Algorithms need data. KNN loads training and test data. Then, pick K. K is an integer. Use Euclidean, Manhattan, or Hamming distance to compare live data to model training data. [12] [7] Euclidean distance is commonly utilized. Order them from farthest to closest. Finally, the first K rows are randomly chosen. The test point's category will be the one with the most rows.

B. Bagged Tree

Bagging is a meta-algorithm for increasing the consistency and accuracy of machine learning algorithms for various statistical tasks, such as classification and regression. In addition, reducing the number of observations helps reduce variance and prevent overfitting. As a result of its capacity to reduce data size, bagging enables the use of other machine learning algorithms for categorization. [9] Figure 3 depicts the procedures required for classifying. Other approaches may also be employed. However, decision tree methods are the most common.

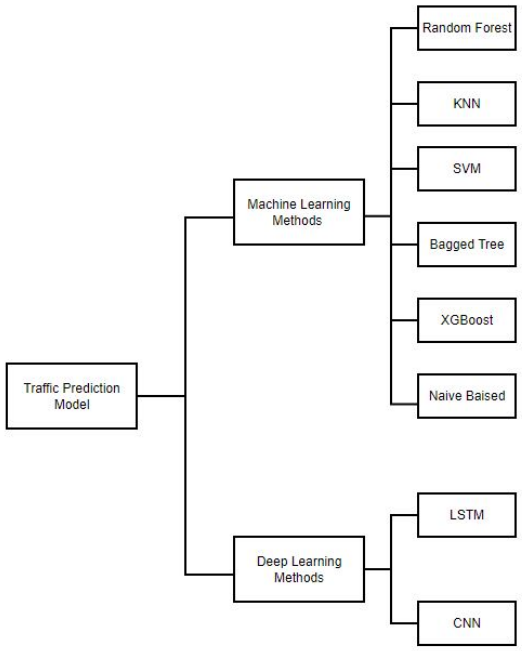

Fig. 1. Applied Models

C. Random Forest

Random forest employs decision trees. Bootstrap samples are randomly chosen training set data used to create trees. Training and test sets overlap by 50%. We'll discuss OOB samples again. Introducing a second random event via feature bagging helps identify decision trees. Problem scope determines forecast. In categorization, the predicted class is the most popular. In regression, the mean of the decision trees is obtained. Cross-validation evaluates forecast dependability.

D. Support Vector Machine

SVM is a supervised learning model to differentiate classes accurately. Input data is converted into many feature spaces. Complexity depends on the algorithm's feature space. SVM has three uses. Details: Kernel linearization 2.) PolRBF Kernel (RBF). Using the RBF kernel in this investigation improved accuracy. Nonlinear data partitioning is significant. [4] The following equation explains RBF.

$$f(X1, X2) = (a + X1T * X2)b$$

Simple kernel polynomial formula. f is their polynomial decision limit (X1, X2). Data are in X1 and X2 formats. This vast category includes classification concerns. This function determines the linear kernel:

$$f(X) = wT * X + b$$

If you have grouping data (X) and a linear coefficient of prediction (b), you may compute the lowest relevant weight vector (w) (obtained from the training data). This equation shows SVM's decision threshold. [4]

E. Naive Bayes

It's a cliche for a reason, and Nave Bayes's work proves it: sometimes the simplest answers are the best. Despite the recent progress in the field, Machine Learning still has its traditional strengths of being quick, accurate, and simple to use. While it has shown success in several settings, natural language processing (NLP) problems are where it shines. Naive Bayes is a popular machine learning technique for a wide range of classification tasks because of its theoretical foundation in Bayes' Theorem. This article deeply dives into the Naive Bayes algorithm and its underlying concepts.

F. LSTM

LSTM uses deep learning to find long-term dependencies. Because they can be used in many ways, they are a popular solution to many problems. LSTMs stop habits from forming. An LSTM RNN has four layers that are connected. Change or get rid of the LSTM's gated cell state. Lines carry both the outputs and inputs of nodes. The layers of a neural network are shown by the orange boxes, while the green circles show the activities. Forking makes a copy of the text, while concatenation joins lines together. The sigmoid layer gives each input element a weight between 0 and 1. This weight tells the next layer how much of each input element to send. [12] Three gates set the state of LSTM cells. At each cycle in LSTM, cell input is ignored by using a sigmoid function. Sigmoid and tanh functions give variables weights (from -1 to 1) based on whether they pass or fail the test (0 or 1). In the last step, the output% is calculated with the help of the sigmoid and tanh functions. Each stage of an RNN will choose its data from a very large database.

V. Methodology

Research has been conducted using a variety of methodologies from start to finish. This research uses the statistical packages Pandas, Numpy, Open Source, and Matplotlib.pyplot, Keras, and Sklearn in order to create regression models that are then applied to the process of traffic forecasting.

A. Data Set

Over the past several years, there has been a discernible rise in the amount of traffic congestion. Several factors are involved, including the rise in the number of people living in metropolitan areas, the absence of coordinated scheduling for traffic signals, and the dearth of real-time data. The current state of traffic congestion has highly significant and far-reaching consequences. These outputs of machine learning algorithms written in Python 3 are illustrated using data collected from the Kaggle website. The use of these algorithms is necessary in order to get an accurate estimate of the flow of traffic. There are two collections of data that are collected: the first is traffic data from 2015, which includes the date, the hour, the number of cars, and the intersection; the second is traffic data from 2017, which includes the same information so that comparisons can be made quickly and without ambiguity. In order to generate a traffic flow prediction for each 1-hour interval, it was necessary to pre-process the data acquired from 1 to 24 hours of a time interval. This pre-processing resulted in the elimination of data that was not required.

B. Regression Model

Regressor model analysis may be a mathematical method for resolving the link between one dependent (criterion) variable and one or more independent (predictor) variables. In this case, the criterion variable would be the dependent variable. The evaluation comes up with a predicted value for the benchmark, which results from adding the scalar vectors that make up the predictors. The mean square error is used to determine

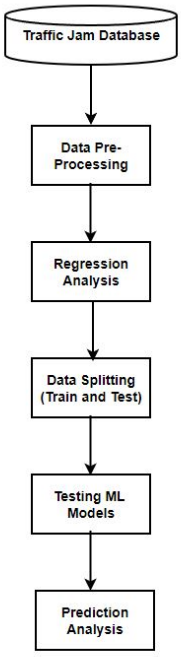

Fig. 2. Data prepossessing flow

how accurate something is. Therefore, to acquire the predicted error from the observed value and the true value compared to the standard deviation utilized in statistical techniques, it is necessary to first determine the true value. In Figure 2, you can see the Regression model used for the Traffic Prediction. JupyterLab is a collaborative programming environment that runs in a web browser. JupyterLab is a flexible platform that can design and display the user interface to accommodate a wide variety of metadata in machine learning. The code is now being implemented in the Jupyter notebook, which uses Python3, the familiar environment. Through the use of the command prompt, this may be accessed and installed. This is to gain access from the local disk, which is done in this way. The Jupyter notebook is installed using the command prompt, and a local host is developed after that. The file is accessible through this host, and the prediction is carried out in the python environment utilizing various library packages and mathematical models.

VI. IMPLEMENTATION AND RESULTS

Several machine algorithms were implemented and assessed to increase productivity and precision. Utilizing a Bagged Tree, we determined classification and regression. Algorithm. This method aims to predict the value of the target variables. Decision tree learning is an example of a formula whose input is a vector of values and attributes and whose output is a single value termed "Decision." [8] It belongs to the class of supervised learning algorithms. It may be used to tackle classification and regression issues. BT recognizes by evaluating the results of a series of tests on the training data set. [13]

This study examines our results based on accuracy, precision, recall, and F1-Score performance parameters. The data employed for training and testing is separated into 80-20 unique sets using all available methods. All major parameters of the classification model were fine-tuned to get the best accurate metrics possible. The accuracy, recall, and F1-score performance of the six classification models evaluated in this study are shown in Table (1).

A software that can produce GPS coordinates. Follow the procedure outlined. Verify if the matrix of the data set is enough. The data is separated into a training set and a test set. Investigate many techniques for machine learning. Foretell the interval's characteristics over the following half hour using a machine learning technique. Conclusions Congestion Issues By employing the procedures above as a blueprint, this strategy might be utilized to construct a more precise machine learning model that is now accessible. Coaching a deep network is much easier when BP and the radiant-based improvement mechanism are used. These findings demonstrate the poor performance of deep neural networks trained in this way. The reason is that we have delayed the implementation of deep learning models in my team. Nonetheless, deep learning and genetic algorithms are not recommended due to the low quality of the selected data set. To prevent the model from overfitting, we reduced the data set's massive dimensions and overcame several obstacles, including those associated with Big Data.

Method name	**A**ccuracy	**P**recision	**R**ecall
Bagged Tree	0.92	92%	89%
SVM	0.94	94%	92%
KNN	0.96	95%	94%
Random Forest	0.92	91%	91%
XGBoost	0.93	94%	93%
Naive Biased	0.78	077%	76%
LSTM	0.98	0.97%	95%

TABLE I
ACCURACY TABLE

VII. CONCLUSION

In this paper, we used different deep learning and machine learning methods to determine which one can be used to determine traffic in Dhaka City accurately. Using these methods, we were able to determine classification and regression, increasing the accuracy of the results. In this paper, our data set was created using two main methods, GPS AND PVD, from different types of road transport vehicles to create a more reliable and precise data set. The results show that using the LTSM method will give a higher accuracy in determining traffic hot spots, while the bagged tree method increases accuracy because it reduces variance. Our research confirmed that the LTSM method would be the best method to know traffic patterns since its high recall percentage and ability to stop patterns from forming. The combination and usage of these methods create opportunities to explore how to best understand traffic patterns and to foretell areas that are liable to be congested within a particular time range.

VIII. FUTURE WORK

Deep learning, artificial neural networks, and even big data are some of the methods that may be used to improve future versions of the system so that they include extra features linked to traffic management. The users may then use this method to search for the route that will make getting to their place the least difficult possible. Users may get assistance from the algorithm in search recommendations and the location of the choice that is easiest to utilize in a region with low foot traffic. In the past, a wide range of different methods of forecasting was utilized to predict road traffic congestion. Even though there is more potential for improvement in terms of the accuracy of the congestion forecast, there are more methods that produce credible forecasts. In addition, at this time, making use of traffic data that is more easily available in conjunction with newly created forecasting algorithms could increase the accuracy of the predictions. These days, accurate traffic forecasting is essential for almost every part of the nation and for every part of the world. Therefore, this kind of prediction might be effective for predicting traffic ahead of time. When it comes to improving the ability to foresee congestion, the grade and accuracy of the traffic forecast are crucial aspects. In the future, established order accuracy prediction will be calculated using simpler and more accessible methodologies. This is done in the hopes that users would find the prediction model useful and not spend their time attempting to forecast the data. Other elements are easily available, such as weather predictions, GPS that displays

the route, and accident-prone areas that will be marked to deter users from choosing to utilize dangerous roads and anticipate traffic. This can be accomplished using artificial neural networks, massive data, and deep learning.

REFERENCES

[1] do Vale Saraiva, Tiago, and Carlos Alberto Vieira Campos. "An approach using machine learning and public data to detect traffic jams." 2021 International Wireless Communications and Mobile Computing (IWCMC). IEEE, 2021.

[2] Mugion, Roberta Guglielmetti, et al. "Does the service quality of urban public transport enhance sustainable mobility?." Journal of cleaner production 174 (2018): 1566-1587.

[3] Mandhare, Pallavi A., Vilas Kharat, and C. Y. Patil. "Intelligent road traffic control system for traffic congestion: a perspective." International Journal of Computer Sciences and Engineering 6.07 (2018): 2018.

[4] Yoon, Jungkeun, Brian Noble, and Mingyan Liu. "Surface street traffic estimation." Proceedings of the 5th international conference on Mobile systems, applications, and services. 2007.

[5] Le, Luong-Vy, et al. "A practical model for traffic forecasting based on big data, machine learning, and network KPIs." 2018 15th IEEE Annual Consumer Communications Networking Conference (CCNC). IEEE, 2018.

[6] M. N. Hossain, S. F. U. Zaman, T. Z. Khan, S. A. Katha, M. T. Anwar, and M. I. Hossain, "Implementing Biometric or Graphical Password Authentication in a Universal Three-Factor Authentication System," 2022 4th International Conference on Computer Communication and the Internet (ICCCI), 2022, pp. 72-77, DOI: 10.1109/ICCCI55554.2022.9850264.

[7] Hashemi, Sirous, et al. "Seasonal variations of the surface urban heat island in a semi-arid city." Remote Sensing 8.4 (2016): 352.

[8] Liu, Xue, et al. "A system dynamics approach to scenario analysis for urban passenger transport energy consumption and CO2 emissions: A case study of Beijing." Energy Policy 85 (2015): 253-270.

[9] Tišljarić, Leo, et al. "Traffic state estimation and classification on the citywide scale using speed transition matrices." Sustainability 12.18 (2020): 7278.

[10] Ran, Xiaojuan, et al. "A novel k-means clustering algorithm with a noise algorithm for capturing urban hotspots." Applied Sciences 11.23 (2021): 11202.

[11] Priambodo, Bagus, Azlina Ahmad, and Rabiah Abdul Kadir. "Predicting Traffic Flow Propagation Based on Congestion at Neighbouring Roads Using Hidden Markov Model." IEEE Access 9 (2021): 85933-85946.

[12] Gupta, Anunay, et al. "Advances of UAVs toward future transportation: The State-of-the-Art, challenges, and Opportunities." Future Transportation 1.2 (2021): 326-350.

[13] Afrin, Tanzina, and Nita Yodo. "A Long Short-Term Memory-based correlated traffic data prediction framework." Knowledge-Based Systems 237 (2022): 107755.

[14] Goggin, Gerard. 2012. "Driving the Internet: Mobile Internets, Cars, and the Social" Future Internet 4, no. 1: 306-321. https://doi.org/10.3390/fi4010306

[15] Ata, Ayesha Khan, Muhammad Abbas, Sagheer Ahmad, Gulzar Fatima, Areej. (2019). MODELLING SMART ROAD TRAFFIC CONGESTION CONTROL SYSTEM USING MACHINE LEARNING TECHNIQUES. Neural Network World. 2019. 99. 10.14311/NNW.2019.29.008.

[16] Elfar A, Talebpour A, Mahmassani HS. Machine Learning Approach to Short-Term Traffic Congestion Prediction in a Connected Environment. Transportation Research Record. 2018;2672(45):185-195. doi:10.1177/0361198118795010

[17] Polish-Japanese Institute of Information Technology, Koszykowa 86, 02-008 Warsaw, Polska, jrzeszotko@gmail.com; hoa@mimuw.edu.pl.

[18] E. Heyns, S. Uniyal, E. Dugundji, F. Tillema, C. Huijboom, Predicting Traffic Phases from Car Sensor Data using Machine Learning, Procedia Computer Science, Volume 151, 2019, Pages 92-99, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2019.04.016.