



**IJIRCCCE**

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH


IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 6, June 2023

**ISSN** INTERNATIONAL  
STANDARD  
SERIAL  
NUMBER  
INDIA

**Impact Factor: 8.379**

 9940 572 462

 6381 907 438

 [ijircce@gmail.com](mailto:ijircce@gmail.com)

 [www.ijircce.com](http://www.ijircce.com)



# Bus Travel Time Prediction by Using Machine Learning Algorithm

Pani Regis Livina J<sup>1</sup>, Valli Rathi I<sup>2</sup>

PG student, Department of Computer Science and Engineering, PET Engineering College, Tirunelveli, India<sup>1</sup>

Assistant Professor, Department of CSE, PET Engineering College, Tirunelveli, India<sup>2</sup>

**ABSTRACT** - The notion of smart cities is being adapted globally to provide a better quality of living. A smart city's smart mobility component focuses on providing smooth and safe commuting for its residents and promotes eco-friendly and sustainable alternatives such as public transit (bus). Among several smart applications, a system that provides up-to-the-minute information like bus arrival, travel duration, schedule, etc., improves the reliability of public transit services. Still, this application needs live information on traffic flow, accidents, events, and the location of the buses. Most cities lack the infrastructure to provide these data. In this context, a bus arrival prediction model is proposed for forecasting the arrival time using limited data sets. The location data of public transit buses and spatial characteristics are used for the study. One of the routes of, India, is selected and divided into two spatial patterns: sections with intersections and sections without intersections. The machine learning model KNN and Random Forest modeled for both spatial patterns individually. A model to dynamically predict bus arrival time is developed using the preceding trip information and the machine learning model to estimate the arrival time at a downstream bus stop. The performance of models is compared based on the R-squared values of the predictions made, and the proposed model established superior results. It is suggested to predict bus arrival in the study area. The proposed model can also be extended to other similar cities with limited traffic-related infrastructure.

**KEYWORDS:** Real Time; prediction; Machine learning; Bus

## I. INTRODUCTION

Over recent years, traffic congestion has increased at an alarming rate and has become a global phenomenon. This surge is attributed to increases in motorization, urbanization and population growth. Congestion creates burdens on transportation infrastructure, increases travel time, fuel consumption and pollution, and reduces accessibility and mobility [1]. A way to mitigate this issue is by increasing the capacity of transport infrastructure by building more roads, highways, separate lanes etc. This option is not viable because of its own limitations. A more economical option is to encourage the use of public transport by the public and efficiently manage the existing resources using Intelligent Transportation Systems (ITS). Buses are considered one of the important means of public transport owing to their coverage and accessibility by mass people. Buses are available all year long, are more economical and eco-friendly than private vehicles. Moreover, in some cities buses have their own dedicated lanes which makes them faster than cars and help reduce travel times in heavy traffic conditions during peak hours. But to motivate and encourage usage of buses, providing commuters with reliable bus travel time and arrival information is essential. Advanced Public Transportation System (APTS) is an integral component of intelligent transportation systems. With the advent of intelligent transport systems in cities, buses are fitted with GPS enabled in-vehicle navigation systems as part of urban planning. This tracking system integrated in buses generates automatic vehicle location (AVL) data which can be used to provide accurate bus arrival information to passengers waiting at the bus stops, leading to a decrease in waiting times. It increases the satisfaction of commuters by enabling them to plan their travel ahead of time and also attracts additional ridership. Predicting accurate bus travel and arrival time is not an easy task because it depends on many external factors such as passenger load, passenger boarding/alighting time, number of signalized intersections, traffic congestion and weather. Moreover, it is not guaranteed to have



information about all the above factors to predict bus journey times. Hence there is a need to develop intelligent models which can estimate reliable journey and arrival times using minimal features for situations where all features are not available.

**Objectives** - Bus transit planning has an important role to play as part of urban transportation planning and providing passengers with accurate and reliable travel information is very important to increase additional bus ridership. In cities, bus transit times are difficult to estimate because travel times on links, dwell times at stops, and delays at intersections fluctuate spatially and temporally. It becomes even more challenging if do not have much information/data about the external factors which influence the travel time [2]. The main objective of the thesis is to develop and compare models to predict bus journey and arrival times using archived/historical GPS-based AVL data along with prior bus routes information and stops information.

## II. REVIEW OF LITERATURE

Travel time is a fundamental measure in transportation that is defined as the time it takes for a vehicle to navigate between two points of interest. It is used by planners and engineers to help to schedule transit bus arrivals at each bus stop along a route. Accurate prediction of bus arrival time can help improve the quality of bus-arrival-time information service, and attract more ridership. Travel time reliability is one of the major measurements of effectiveness that affect mode choice for transportation between two locations within a network. Mobility in urban areas impacts urban livelihood to a great extent. To enhance urban mobility, several research studies on predicting travel time have been conducted to provide passengers (or commuters) with estimations (within a margin of error) of how long a particular trip will take time to reach the destination.

Feng, Wei, "Analyses of Bus Travel Time Reliability and Transit Signal Priority at the Stop-To-Stop Segment Level" (2014). In 2014, Feng analyzed bus travel times and the factors that affect its reliability. And included a review of several articles that studied impacting factors on travel time. One of the most influential factors that is associated with travel time is travel distance. Other studies considered the number of signalized intersections to be an impacting factor; however, these factors' impacts varied due to the different geometric characteristics and signal timings of the arterials used in the study. Another important factor that impacts transit travel time is traffic congestion. The author reviewed the impact of congestion on travel time using "time of day" and/or "travel direction" as the independent variable. The number and spacing of bus stops is also a variable that had a positive impact on bus travel time and reliability. Several studies use the number of actual stops made as an independent variable. Other variables such as bus departure delays and dwell time impact bus travel time [3]. Lastly, passenger load, number of passengers boarding and alighting had a significant influence on bus travel time and reliability. However, nearside and far side bus stop types did not have any significant impact on travel time.

Xinghao et. al predicting bus real-time travel time basing on both GPS and RFID data. 13th COTA International Conference of Transportation Professionals (CICTP 2013). A study was conducted in 2013 by Xinghao et. al to develop a short-term prediction model using real-time bus location and radio-frequency identification (RFID) data. The proposed model were based on an augmented self-adapting smoothing algorithm that is used to predict the running speed of transit buses using short-term sample speeds of taxis and buses. In the development of the model, the researchers took into consideration the variation of bus speeds due to traffic controls and other impacting factors. The proposed model, which integrated AVL and RFID data, was tested against the historical data-based model which used only historical AVL data. The results indicated that the relationship between speeds of transit buses and taxis on the same link during the same time period is linear which was determined to be statistically significant with R2 values ranging from 0.72 to 0.83. Also, the results showed that the combined data model out-performed the AVL-only data model [4].

Yetiskul and Senbil public bus transit travel-time variability in Ankara (Turkey) in September 2012  
A study in Ankara, Turkey by Yetiskul and Senbil was conducted to determine which factors influence the variability of bus transit travel time. The causes of inconsistent travel times were identified as both external and internal factors. Re-occurring traffic congestion during peak hours and non-recurring factors such as traffic accidents or roadway maintenance were classified as external factors; whereas, fare collection process, passenger capacity, and number of stops along a route were classified as internal factors. To account for variation caused by service





region, highways, and individual bus lines, three models were developed and tested. The outcome indicated that travel time variability in transit systems were caused by temporal dimension (time of day and day of week), spatial dimension (operation system's physical characteristics), and service characteristics (number of stops on a route, dwell time, maximum passenger load, etc.) [5].

Zhang and Xiong approach an agent-based model (ABM) that performs multi-step travel time predictions by using historic and real-time traffic data. Zhang and Xiong employed an agent-based model (ABM) approach that performs multi-step travel time predictions by using historic and real-time traffic data. Each agent in the model represented a domain in a decision making system that predicts travel time for each time interval based on a historical database and real-time data. A combination of each agent's prediction results in an output that presents the predicted travel time distribution of the proposed model. The instantaneous travel time method, historical average method, and the k-Nearest Neighbor (k-NN) prediction method were all compared with the proposed model to evaluate its performance. The instantaneous travel time method was used to predict future travel times with the assumption that the current speed of traffic along a segment will remain constant throughout the trip.

### III. PROPOSED SYSTEM

For predicting the bus delays and to bus the models, the data assemble by the organization of Transportation Statistics of all the domestic bus taken in 2015 is collected and used. This Model is capable of filling the absent values which is crucial for refining data for model.

**Problem overview:** Deciding of the problem of the thesis was one of the main difficulties of the project. Indeed, at the beginning of the project, the framework was not specified. The first objective was to get ideas of various projects and determine their potential in term of machine-learning and their feasibility. As the service quality is one of the main issues for public transportation planners and because GIRO is one of the main actors in this sector, it was decided that the project would aim to improve service quality. An analysis of the available data led to the first outline of a problem: forecasting the delay at each stop using real-time data. However, the literature review showed that this problem was addressed by several before. Moreover, GIRO acts upstream in the public transportation system, and no real application could be found. Working with off-line data was essential for the project. Then, the idea of working with the layover appeared. The analysis of the data shows that the scheduled time is not always respected, and the delay could be propagated inside a block. The layover, which is a buffer time, could prevent the propagation of delay. Thus it was thought that a new method to find adequate layover could be developed. The layover is the main leverage a public transport planner could use in order to prevent delay propagation. Predicting the departure status of the trip was then the new objective of the thesis. However, once the situation is modeled, it just described the adequacy of the layover in this particular situation, and the importance of the feature layover was too significant. It was too difficult to transpose the models to other periods. The final idea was to forecast the delay of the end-trip, in order for public transport to have an idea of the necessary buffer time to prevent delay propagation.

**Machine Learning:** In the statistical context, Machine Learning is defined as an application of artificial intelligence where available information is used through algorithms to process or assist the processing of statistical data [6]. While Machine Learning involves concepts of automation, it requires human guidance. Machine Learning involves a high level of generalization in order to get a system that performs well on yet unseen data instances.

**Classes Of Machine Learning -** There are two main classes of machine learning techniques:

- supervised machine learning and
- unsupervised machine learning

Logistic regression, when used for prediction purposes, is an example of supervised machine learning. In logistic regression, the values of a binary response variable (with values 0 or 1, say) as well as a number of predictor variables (covariates) are observed for a number of observation units. These are called training data in machine learning terminology. The main hypotheses are that the response variable follows a Bernoulli distribution (a class of probabilistic models), and the link between the response and predictor variables is the relation that the logarithm of the posterior odds of the response is a linear function of the predictors. The response variables of the units are assumed to be independent of each other, and the method of maximum likelihood is applied to their joint probability



distribution to find the optimal values for the coefficients (these parameterize the aforementioned joint distribution) in this linear function. The particular model with these optimal coefficient values is called the “fitted model,” and can be used to “predict” the value of the response variable for a new unit (or, “classify” the new unit as 0 or 1) for which only the predictor values are known. Support Vector Machines (SVM) are an example of a non-statistical supervised machine learning technique; it has the same goal as the logistic regression classifier just described: Given training data, find the best-fitting SVM model, and then use the fitted SVM model to classify new units [7]. The difference is that the underlying models for SVM are the collection of hyper planes in the space of the predictor variables. The optimization problem that needs to be solved is finding the hyper plane that best separates, in the predictor space, the units with response value 0 from those with response value 1. The logistic regression optimization problem comes from probability theory whereas that of SVM comes from geometry.

Other supervised machine learning techniques mentioned later in this briefing include decision trees, neural networks, and Bayesian networks. The main example of an unsupervised machine learning technique that comes from classical statistics is principal component analysis, which seeks to “summarize” a set of data points in high-dimensional space by finding orthogonal one-dimensional subspaces along which most of the variation in the data points is captured. The term “unsupervised” simply refers to the fact that there is no longer a response variable in the current setting. Cluster analysis and association analysis are examples of non-statistical unsupervised machine learning techniques [8]. The former seeks to determine inherent grouping structure in given data, whereas the latter seeks to determine co-occurrence patterns of items.

#### IV. METHODOLOGY

KNN Machine Learning-The KNN algorithm means K-NEAREST NEIGHBOURS. This algorithm often used in classification have some classified data and have new data item, but not sure which is the class of that new data, use KNN machine learning algorithm. KNN is supervised and pattern classification learning algorithm. KNN can be used in classification as well as regression. The KNN algorithm is the most accurate model because it makes highly accurate prediction, so KNN algorithm want highly accuracy. This algorithm has some drawback which is the outcomes accuracy is depend on the quality of the available data. So, if have good quality of data then outcomes accuracy is higher else, won't get the higher accuracy. The KNN algorithm is easy to implement there are two parameters is required to implement. First is the value of the K, and second one is the distance function. KNN is one of the most use data mining and classification algorithms. And KNN is used in cancer diagnosis, pattern recognition, text classification, email spam detection, fraud detection, and in regression it used to risk assessment, score prediction etc. KNN algorithm store all the available cases and classify new data of K in similarity measure [9]. It suggests that if you are similar to your neighbor then you are one of them. For example, apple is more similar to orange, banana, and mango rather than dog, monkey, lion then most likely apple is belonging to group of fruits. KNN used is search application when you want to similar items then you called the search as a KNN search. K is the number of neighbors near to the new object have to assign. If  $k=3$  then the most common three nearest neighbors are checked and the most common neighbors class are assigning to the testing data item. So, this is the K in KNN algorithm. The biggest use of KNN algorithm is the recommendation system. The recommended system is like the shop counter when you asking for a product it not shows only that product, it displayed you and also suggest your relevant sets of products and related to the item you are already too interested to buying it. The KNN algorithm is used in recommended product like in amazon and recommended media in case of NETFLIX. More than 35% of amazon revenue is generated by its recommendation engine. For this kind of purpose, use KNN algorithm and more advanced example may like handwriting detection, image recognition or even video recognition, and it is used to get missing value, used in pattern recognition, and it used in gene expression this is also the example of KNN machine learning algorithm [10].

KNN Machine Learning Algorithm -The k nearest neighbor algorithm find the nearest neighbor of new data item, if  $k=3$ , (k is the nearest neighbor) then 3 closest neighbors has been checked and most common of cases data item class has been assign to new data item. This is about KNN algorithm. Measure distance between k and new data point through Euclidean distance. Calculate distance through hamming distance, Manhattan distance

formula for KNN algorithm The Euclidean distance is based on the Pythagoras theorem if to calculate the distance of  $AC^2(AC \text{ SQUAR})$  then in Pythagoras to calculate the  $AB^2 + BC^2(AB \text{ SQUAR} + BC \text{ AQUAR})$  this method can be used in Pythagoras[11]. To calculate the distance between two point  $(x1,x2)$  and  $(y1,y2)$  in two dimensional space then the Euclidean distance between two points is  $D = \sqrt{(x2-x1)^2 + (y2-y1)^2}$ .(^2 is denoted the square) And in three-dimensional space, for points  $(x1,y1,z1)$  and  $(x2,y2,z2)$  then in this case the Euclidean distance of this point is  $D = \sqrt{(x2-x1)^2 + (y2-y1)^2+(z2-z1)^2}$ . This is how Euclidean distance is calculated. Now will see how dose KNN algorithm work.

Random Forest-Random Forest is an ensemble supervised machine learning technique. Machine learning techniques have applications in the area of Data mining. Random Forest has tremendous potential of becoming a popular technique for future classifiers because its performance has been found to be comparable with ensemble techniques bagging and boosting. Hence, an in-depth study of existing work related to Random Forest will help to accelerate research in the field of Machine Learning. A systematic survey of work done in Random Forest area. Derived Taxonomy of Random Forest Classifier which is presented in this paper. Prepared a Comparison chart of existing Random Forest classifiers on the basis of relevant parameters [13]. The survey results show that there is scope for improvement in accuracy by using different split measures and combining functions; and in performance by dynamically pruning a forest and estimating optimal subset of the forest. There is also scope for evolving other novel ideas for stream data and imbalanced data classification, and for semi-supervised learning. Finally presented a few future research directions related to Random Forest classifier.

### V.MODULES

Data Collection : This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data to get, the better our model will perform. There are several techniques to collect the data, like web scraping, manual interventions and etc.

Dataset : The dataset consists of 100 individual data. There are 10 columns in the dataset.

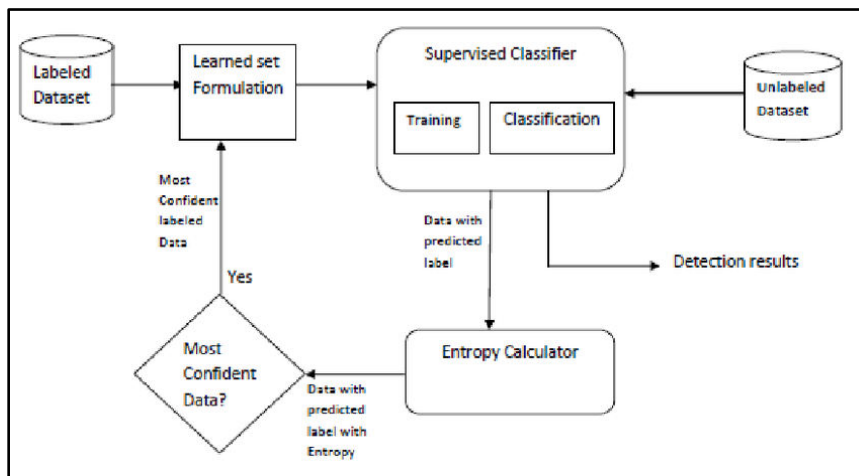


Fig 1. System Architecture

Data Preprocessing: Since the obtained data was raw and unprocessed, there were a lot of cleaning processes that had to be done to make the data usable. The actual timetable was obtained through web scraping. The obtained information was further cleaned and processed and added along with the delay data. The arrival and departure schedule for each bus was recorded for all the dates and was added to the data including the dates when the bus had arrived on time. Next, the scheduled time and the delayed time were added which helped to derive the actual time of

bus arrival [14]. Further comparisons were done and through a few formulas, the target column was marked (0 or 1) by looking at the delay data.

**Data Preparation:** Data Preparation will transform the data. By getting rid of missing data and removing some columns. First it will create a list of column names that want to keep or retain. Next drop or remove all columns except for the columns that want to retain. Finally drop or remove the rows that have missing values from the data set. Split into training and evaluation sets.

**Model Selection:** The principal component analysis is the technique that is used, especially for the reduction of the dimension of the given dataset. The principal component analysis is one of the most efficient and an accurate method for reducing the dimensions of data, and it provides the desired results. This method reduces the aspects of the given dataset into a desired number of attributes called principal components [15]. This method takes all the input as the dataset which is having a high number of attributes so as the dimension of the dataset is very high. This method reduces the size of the dataset by taking the data points on the same axis.

Accuracy on Test Set got an accuracy of 99.1% on test set.

**Saving the Trained Model -** Once you're confident enough to take your trained and tested model into the production-ready environment,

- The first step is to save it into a .h5 or .pkl file using a library like pickle.
- Make sure you have pickle installed in your environment.
- Next, let's import the module and dump the model into .pkl file.

## VI. DISCUSSION

A dynamic model to predict bus arrival time using only limited data sets based on machine learning. Based on the previous study results, the base travel time is predicted using the machine learning model. The performance analysis of KNN and random forest has been denoted as graphical representation in Fig 2.

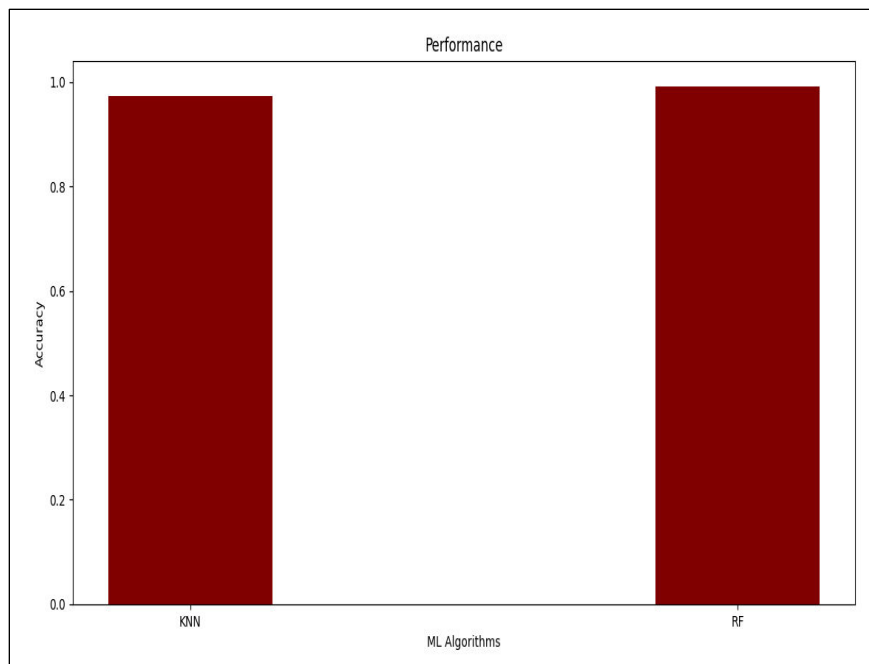


Fig 2. Comparison Graphical representation of KNN and RF

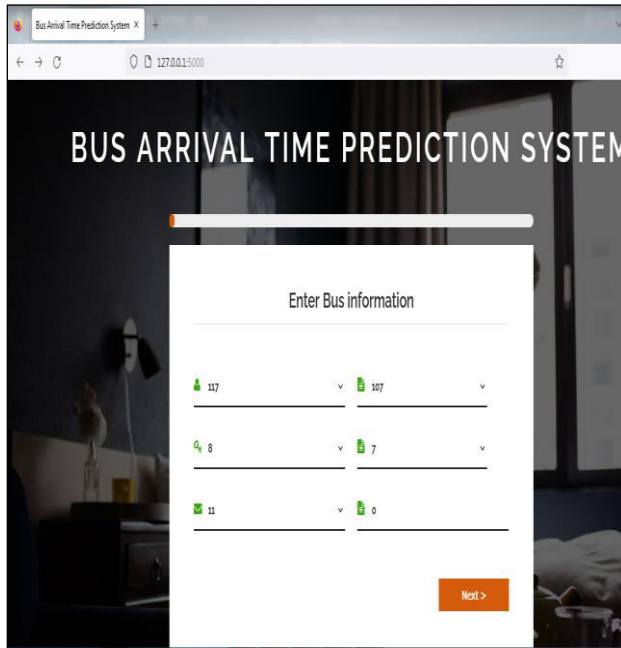


Fig 3. Home Page

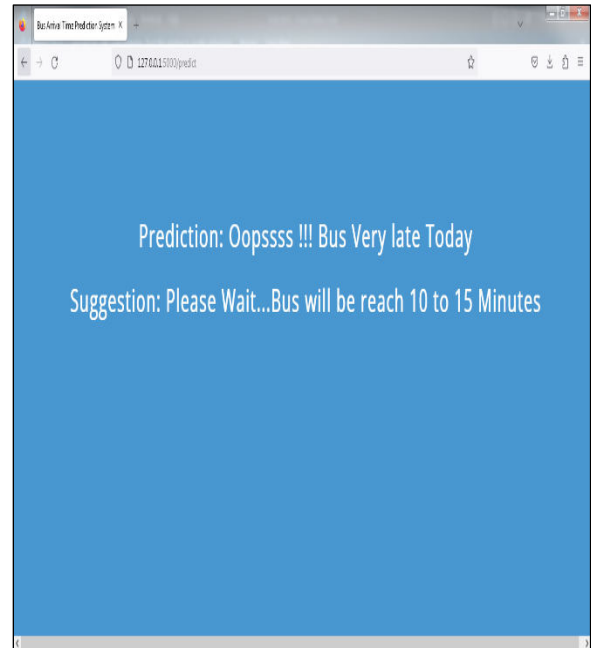


Fig 4. Bus time has been predicted to very late

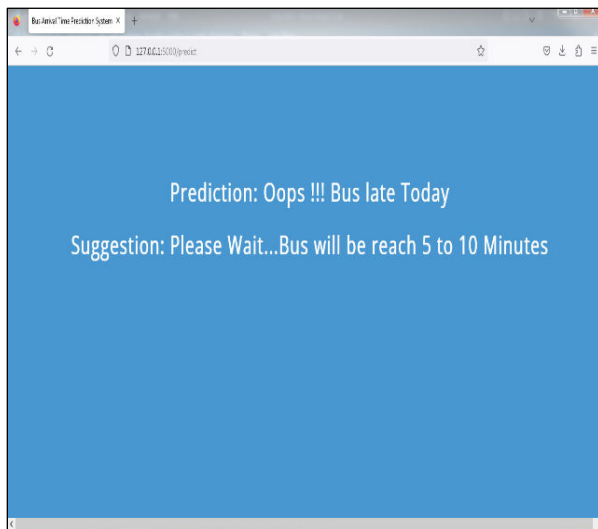


Fig 5. Bus time has been predicted late

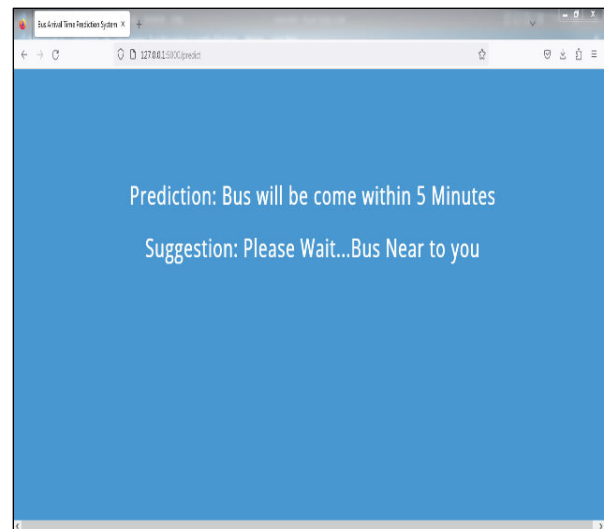


Fig 6. Bus time has been predicted to on time

## VII. CONCLUSION AND FUTURE ENHANCEMENT

The proposed Bus Arrival Time Estimation Model estimates the bus arrival time for individual bus stops for a selected route in India. The sections between the bus stops were divided into two discrete spatial patterns: a section with intersections and sections without intersections. Each spatial pattern was modeled separately using the Random forest and KNN and the proposed model. The proposed model estimated the bus arrival time based on both





machine learning model results and the preceding trip travel time and demonstrated better accuracy than the Random forest and KNN model for both spatial patterns. The preceding trip within 30 minutes to that of the current demonstrated a positive correlation, illustrating its usefulness in forecasting travel times. The proposed model exhibited better accuracy based on the R-squared values. The proposed model can be suggested for bus arrival time forecasting in the study route and extended to other routes to predict the citywide bus arrivals. Similar cities can employ the proposed model as well. If available, the up-to-the-minute details of the traffic flow, incidents, and other delays can further improve the forecasting. The model can be updated in the future once the up-to-the-minute data stream is available. Overall, with the available limited datasets, the proposed model is a promising option for forecasting the bus arrival time. In this project the travel time considered only a single bus in a journey, which is the basis of passenger-oriented travel time predictions. The prediction model should be expanded to be more consistent with the actual application scenario of a conventional large-scale composite transit network. Due to the data's particularity and the demand's diversity and complexity, future research should focus on these points.

### VIII. ACKNOWLEDGEMENT

I would like to express my sincere thanks to Valli Rathi I for her valuable guidance and support in completing this article. I would also like to express my gratitude towards our College.

### REFERENCES

- [1] S. Joshi, S. Saxena, T. Godbole, En Shreya, "Developing Smart Cities: An Integrated Framework", *Procedia Comput. Sci.*, vol 93, no. September, pp. 902–909, 2016, Doi: 10.1016/J.Procs.2016.07.258.
- [2] M. S.Kumbhar En P. S.Yalagi, "Urban Resources for Smart City Application", *Int. J. Eng. Trends Technol.*, vol.40, no.6, pp. 366–370, 2016, Doi: 10.14445/22315381/Ijett-V40p259.
- [3] Polytechnique Montreal, "Smart Cities and Integrated Mobility: A White Paper", *Next Gener. Integr. Mobility Drive. Smart City*, pp. 1– 78, 2018.
- [4] R. M. Savithramma, B. P. Ashwini, En R. Sumathi, "Smart Mobility Implementation in Smart Cities: A Comprehensive Review on State-of-Art Technologies", 2022 4th Int. Conf. Smart Syst. Inven. Technol., pp. 10–17, 2022, Doi: 10.1109/Icssit53264.2022.9716288.
- [5] M. Kumar, N. Kumari, S. tomar, En T. Kumar, "Smart Public Transportation for Smart Cities", *SSRN Electron. J.*, pp. 1–6, 2019, Doi: 10.2139/SSrn.3404487.
- [6] S. Porru, F. E. Misso, F. E. Pani, En C. Repetto, "Smart Mobility and Public Transport: Opportunities and Challenges in Rural and Urban Areas", *J. Traffic Transp. Eng. English Ed*, vol 7, no 1, pp. 88–97, 2020, Doi: 10.1016/J.Jtte.2019.10.002.
- [7] R. S. Chhillar, "A Review of Intelligent Transportation Systems in Existing Framework Using Iot", vol.70, no. 6, pp. 137–143, 2022.
- [8] S. Muthuramalingam, A. Bharathi, S. Rakesh Kumar, N. Gayathri, R. Sathiyaraj, En B. Balamurugan, "Iot Based Intelligent Transportation System (Iot-Its) for Global Perspective: A Case Study", *Intell. Syst. Ref. Libr.*, vol.154, pp. 279–300, 2019, Doi: 10.1007/978-3-030-04203-5\_13.
- [9] Gopikrishnan Nair, B. Anil Kumar and Lelitha Vanajaskshi "Mapping bus and stream travel time using machine learning approaches" Vol.2022, Doi:10.1155/2022/9743070.
- [10] BP Ashwini ,R Sumathi and H S Sudhira " Bus travel time prediction : A comparative Study of Linear and Non- Linear Machine Learning Models" con. Series, vol.2161, Oct 2021, DOI 10.1088/1742-6596/2161/1/012053.
- [11] Leone Pereira Masiero, Marco Casanova, Marcelo Tilio "Travel Time Prediction using Machine Learning" Conf. 4<sup>th</sup> ACM SIGSPATIAL International Workshop on Computational Transportation Science at Chicago, USA.
- [12] Lukasz Palys , Maria Ganzha and Marcin Paprzycki "Machine learning for bus travel prediction", 2022, DOI: 10.1007/978-3-031-08754-7\_72
- [13] W. Fan, Z. Gurmu, Dynamic Travel Time Prediction Models for Buses Using Only GPS Data, *Int. J. of Transportation Science and Technology*, 2015, 353-366.



[14] M. Jiwon, D. Kim, S. Kho, C. Park, Travel Time Prediction Using k Nearest Neighbor Method with Combined Data from Vehicle Detector System and Automatic Toll Collection System, Transportation Research Record Journal of the Transportation Research Board 2256:51-59, 2012.

[15]P. Vidnerov´a, RBF-Keras: an RBF Layer for Keras Library.,2019, Available at [https://github.com/PetraVidnerova/rbf\\_keras](https://github.com/PetraVidnerova/rbf_keras)



Impact Factor: 8.379



# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  [ijircce@gmail.com](mailto:ijircce@gmail.com)



[www.ijircce.com](http://www.ijircce.com)

Scan to save the contact details