Real-time Bus arrival time updating using Speed variations

Shiham Farook
Department of Computer Science and Engineering
University of Moratuwa
Sri Lanka
shiham.23@cse.mrt.ac.lk

Uthayasanker Thayasivam
Department of Computer Science and Engineering
University of Moratuwa
Sri Lanka
rtuthaya@cse.mrt.ac.lk

***Keywords*—Real-time, speed, GPS data, prediction accuracy, public transportation**

I. Introduction

Public transportation systems play a critical role in urban mobility, but their efficiency is often hampered by unpredictable delays due to traffic congestion, weather conditions, and other external factors. Accurate real-time bus arrival time predictions are essential to improve the passenger experience, optimize transit operations, and reduce waiting times. Traditional schedule-based estimation methods rely on static timetables, which making them ineffective in dynamic urban environments. This research introduces a data-driven approach of predicting bus arrival in real time using historical and real-time speed variations. By analyzing GPS-based trajectory data and incorporating temporal, spatial, and weather-related characteristics, we aim to improve the accuracy of arrival time estimations.

II. Literature Review

Previous studies have explored different methodologies for the prediction of bus arrival times in real time. Traditional schedule-based methods, which rely on fixed timetables and historical averages, fail to account for real-time disruptions such as traffic congestion, varying road conditions, and fluctuating passenger demand [1]. To overcome these limitations, researchers have explored statistical and regression-based models, including Kalman filters and autoregressive integrated moving averages (ARIMA). Although these models perform well under stable conditions, they struggle to handle dynamic, nonlinear variations in travel times [2]. Machine learning techniques such as Decision Trees, Random Forests, and Support Vector Machines (SVM) have demonstrated improved predictive accuracy by analyzing large datasets and identifying complex travel patterns [3]. However, these models still require extensive training data and feature engineering to generalize well across different transit networks [4].

Most AI-driven models face challenges related to data availability, computational efficiency, and generalizability. Many models rely on location-specific datasets, limiting their adaptability across different transit networks [5], [8]. Moreover, while deep learning particularly Long Short-Term Memory (LSTM) network models achieve high accuracy, they require significant computational power, posing challenges for real-time deployment in resource-constrained environments [9]. Despite these advancements, existing research has not fully explored the integration of historical segment-wise speed variations into predictive models.

Although prior studies focus on real-time traffic conditions, they often overlook the potential to use historical speed fluctuations to improve prediction accuracy. Additionally, there is limited discussion of how different machine learning architectures compare in their ability to model speed variations dynamically [6], [7]. Addressing these gaps, this study incorporates segment-wise speed variations into machine learning models such as XGBoost, Random Forest, and LSTM to improve arrival time predictions. Furthermore, emerging technologies such as the Internet of Things (IoT) and smart traffic management systems have shown potential to improve real-time data processing for public transportation systems [10]. However, the lack of standardized real-time data sharing frameworks and the need for multimodal transportation integration remain critical challenges [11].

III. Methodology

The GPS dataset used in this study was sourced from the SCDP project, covering July 2022 to October 2022 for bus routes 654 (Kandy-Digana, 16.6 km). GPS data preprocessing was conducted with extraction of bus trip data [12]. The dataset was structured with key attributes such as trip id (one terminal-to-terminal journey), device id (unique trip identifier), stop id, arrival time, duration (segment travel time), distance (between stops), and segment speed (bus speed between segments). Each route contained 15 segments. Only one direction was considered, and incomplete trips missing any essential attributes were removed to ensure data integrity.

https://doi.org/10.31705/ADScAI.2025.23

Feature engineering involved in deriving day type (weekday/Saturday/Sunday), public holidays, time of the day (early morning to night), and weather conditions from Sri Lanka's Meteorological Department. To achieve the research objective, the data were further transformed into two variations: (1) standard features and (2) standard features with lagged speed variables (segment speed from one week, two weeks, three weeks, one hour, two hours, and three hours before).

For Model Development, Random Forest, Gradient Boosting, XGBoost, and LSTM were implemented. Hyperparameter tuning was performed using grid search and Bayesian optimization. Model evaluation employed MAE, RMSE, and R^2 to compare model performance across datasets.

IV. Results and Discussion

The evaluation matrices of prediction is shown in the below Table 1.

TABLE I
EVALUATION MATRICES OF PREDICTION

Model	**Without speed**			**With speed**		
	MAE	***RMSE***	***R2***	***MAE***	***RMSE***	***R2***
random forest	50.92	102.12	0.51	49.92	101.47	0.51
gradient boost	51.05	102.75	0.50	49.56	101.20	0.52
xgboost	44.88	97.32	0.56	43.54	96.74	0.57
LSTM	55.76	150.76	0.34	54.30	143.68	0.37

The evaluation results demonstrate that incorporating lagging speed features significantly enhances the predictive performance of all models in estimating bus arrival times. The study compared Random Forest, Gradient Boosting, XGBoost, and LSTM models using MAE, RMSE, and R^2 metrics, where XGBoost exhibited the highest predictive accuracy. This suggests that tree-based models effectively capture complex non-linear relationships between input features and travel duration, making them well-suited for structured transit data. The inclusion of weekly and hourly past speed variations improved real-time predictions by capturing traffic congestion patterns and temporal dependencies, confirming that historical speed trends are crucial in mitigating sudden fluctuations in travel speed.

While deep learning models like LSTM are generally expected to excel in sequential data problems, their performance in this study was relatively lower, possibly due to limited training data, hyperparameter sensitivity, and the structured nature of the dataset, which favors gradient boosting techniques. These findings align with real-world traffic behaviors, where congestion and travel speed exhibit strong periodic patterns. However, challenges such as GPS data inconsistencies, model generalizability across different cities, and computational efficiency in real-time deployment need to be addressed.

V. Conclusion and future works

This study highlights the effectiveness of incorporating historical speed variations into machine learning models to improve bus arrival time predictions, with XGBoost outperforming other models due to its ability to handle non-linearity and sequential dependencies effectively. The results emphasize that past speed trends provide valuable insights into traffic flow, enabling more accurate real-time estimations. Although the study successfully demonstrates the benefits of integrating lagging speed features, several challenges remain, including data quality issues, model scalability, and real-time implementation constraints.

Future research should explore hybrid modeling approaches that combine tree-based models with deep learning techniques to leverage their respective strengths. Additionally, further studies on feature selection, automated hyperparameter tuning, and real-time optimization could improve predictive performance. Expanding the dataset across multiple transit networks and incorporating external factors like weather conditions, road events, and multimodal transport data would enhance model generalizability, ensuring scalable and adaptive bus arrival time predictions for diverse urban environments.

References

[1] W. Fan and Z. Gurmu, "Dynamic Travel Time Prediction Models for Buses Using Only GPS Data," International Journal of Transportation Science and Technology, vol. 4, no. 4, pp. 353–366, 2015..

[2] Y. P. Huang et al., "Bus arrival time prediction and reliability analysis: An experimental comparison of functional data analysis and Bayesian support vector regression," Applied Soft Computing, vol. 111, p. 107663, 2021.

[3] P. He, G. Jiang, S.-K. Lam, and Y. Sun, "Learning heterogeneous traffic patterns for travel time prediction of bus journeys," Information Sciences, vol. 512, pp. 1394–1406, 2020.

[4] H. Liu, H. Xu, Y. Yan, Z. Cai, T. Sun, and W. Li, "Bus Arrival Time Prediction Based on LSTM and Spatial-Temporal Feature Vector," IEEE Access, vol. 8, pp. 11917–11929, 2020.

[5] D. Tsitsokas, A. Kouvelas, and N. Geroliminis, "Modeling and optimization of dedicated bus lanes space allocation in large networks with dynamic congestion," Transportation Research Part C: Emerging Technologies, vol. 127, p. 103082, 2021.

[6] O. Alam, A. Kush, A. Emami, and P. Pouladzadeh, "Predicting irregularities in arrival times for transit buses with recurrent neural networks using GPS coordinates and weather data," Journal of Ambient Intelligence and Humanized Computing, vol. 12, no. 7, pp. 7813–7826, 2021.

[7] N. Singh and K. Kumar, "A review of bus arrival time prediction using artificial intelligence," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 12, no. 4, 2022.

[8] C. Csiszár and Z. Sándor, "Method for analysis and prediction of dwell times at stops in local bus transportation," Transport, vol. 32, no. 3, pp. 302–313, 2016.

[9] B. Ghosh, "Predicting the duration and impact of the non-recurring road incidents on the transportation network," Nanyang Technological University, 2019.

[10] Y. Yuan et al., "Bus Dynamic Travel Time Prediction: Using a Deep Feature Extraction Framework Based on RNN and DNN," Electronics, vol. 9, no. 11, p. 1876, 2020.

[11] P. Wang, X. Chen, Y. Zheng, L. Cheng, Y. Wang, and D. Lei, "Providing real-time bus crowding information for passengers: A novel policy to promote high-frequency transit performance," Transportation Research Part A: Policy and Practice, vol. 148, pp. 316–329, 2021.

[12] S. Ratneswaran and U. Thayasivam, "Extracting potential Travel time information from raw GPS data and Evaluating the Performance of Public transit - a case study in Kandy, Sri Lanka," in 2023 3rd International Conference on Intelligent Communication and Computational Techniques, ICCT 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICCT56969.2023.10075789.