Bus Arrival Time Prediction

|  |
| --- |
| Dias Orazymbetov Astana IT UniversityAstana, Kazakhstan dias.orazymbetov@gmail.com |
|  |  |

The inefficiency of public transportation, characterized by unreliable schedules and unpredictable bus arrival times, leads to increased dependence on private vehicles, exacerbating urban congestion and environmental impact. This study presents a Bus Arrival Time Prediction System leveraging Long Short-Term Memory (LSTM) networks and Random Forest models to enhance prediction accuracy and reliability. The Bus Journey Dataset, enriched with real-time weather and traffic data, is utilized to train and validate the models.

Data preprocessing includes feature engineering, Min-Max normalization, and sequence generation to optimize model performance. LSTM captures complex temporal dependencies, while Random Forest enhances feature robustness, reducing overfitting. Performance is assessed using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and accuracy, targeting MAE < 5mins and prediction accuracy >90%. Comparative benchmarks with LightGBM and Boosting models validate the framework’s efficiency.

Experimental results demonstrate that the hybrid LSTM + Random Forest approach significantly outperforms traditional methods, offering a scalable and adaptable solution for urban transit optimization. By integrating real-time external factors, this system contributes to intelligent transportation systems, improving commuter experience, optimizing transit resource allocation, and supporting sustainable urban mobility.

Keywords—Bus Arrival Time Prediction, LSTM, Random Forest, Public Transportation, Intelligent Transportation Systems.

# Introduction

Public transport is crucial to city mobility, but inefficiencies such as unreasonable schedules and untrustworthy arrival times have a heavy impact on commuter satisfaction and discourage individuals from using public transport. As a result, most commuters turn to private transport, which increases traffic congestion, fuel usage, and air pollution. Remedies that are data-based and can increase the reliability and efficiency of bus arrival times are necessary to correct these issues.

With the evolution of machine learning and big data processing, predictive modeling has been a potential way to solve these issues. This study examines a hybrid machine learning model that combines Long Short-Term Memory (LSTM) networks and Random Forest models to enhance the accuracy of bus arrival time prediction. LSTM networks are capable of extracting sequential relationships in time-series data effectively, and Random Forest adds feature robustness, avoiding overfitting. The model is trained on the Bus Journey Dataset, which holds detailed bus route information, vehicle IDs, schedules, and stop-to-stop trip times. The external factors of real-time traffic congestion and weather conditions are included to improve forecasts.

For the optimization of model performance, comprehensive data preprocessing is done, including feature engineering, missing value handling, and Min-Max normalization. Model performance is evaluated on the basis of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and overall prediction accuracy, with the objective of achieving MAE < 5 minutes and accuracy > 90%. Comparison with LightGBM and Boosting models provides further validation of the proposed framework.

Through the creation of a scalable and flexible prediction system, this study hopes to further advance intelligent transportation systems. The results can improve commuter satisfaction, optimize public transportation scheduling, and encourage sustainable urban mobility by lessening reliance on private cars.

# Literature Review

Conventional models, including Artificial Neural Networks (ANNs), have been extensively employed for transit time prediction. Research by Lim et al. (2024) established an ANN-based model that utilized GPS data gathered from a bus fleet operating in Malaysia, with inputs of travel distance, time attributes, and stop locations. The model had a Mean Absolute Error (MAE) of 0.0056 and a Root Mean Squared Error (RMSE) of 0.0123, performing better than traditional regression-based approaches. Yet, Artificial Neural Network (ANN) models tend to fail at capturing long-term dependencies in sequential data, making them less suitable for dynamic urban transit systems.

Boosting-based models have also been investigated for the purpose of ETA prediction. Kam et al. (2024) employed LightGBM, XGBoost, and AdaBoost and observed that LightGBM performed better than other algorithms because it could process large data with low computation needs. Although boosting models are high in accuracy, they need a lot of data preprocessing and careful feature selection to achieve optimal performance.

Multi-layer Perceptron (MLP) models have also been investigated for transit predictions. Xu et al. (2023) benchmarked the MLP and MLP Regressor models against a Malaysian dataset and achieved a Mean Absolute Percentage Error (MAPE) of 2.09%. While MLP demonstrated favorable results, the paper cited the necessity for larger datasets and greater hyperparameter tuning to enhance scalability and generalizability.

# Identified Gaps And Proposed Solutions

While these studies highlight advancements in bus arrival time prediction, a common gap persists: the lack of integration of advanced deep learning models for capturing sequential patterns and robust ensemble techniques for improving prediction accuracy. Existing models struggle to handle complex temporal dependencies and dynamic variations in urban transit systems.

To address this gap, a hybrid approach combining Long Short-Term Memory (LSTM) networks and Random Forest models will be utilized. The LSTM networks will capture temporal dependencies and sequential patterns in bus journey data, while Random Forest models will enhance robustness and accuracy by addressing variability in features. The "Bus Journey Dataset" (<https://bit.ly/4gWNPie>), which provides detailed information on bus routes, vehicle numbers, schedules, and stop-to-stop timings, will be used for training and validating these models. This comprehensive dataset offers a solid foundation for developing scalable and precise prediction frameworks.

By integrating these advanced methodologies, this approach aims to overcome the limitations of existing models, delivering highly accurate, scalable, and efficient bus arrival time predictions for diverse urban and rural transit systems.

# Methodology

This study employs a hybrid machine learning framework integrating Long Short-Term Memory (LSTM) networks and Random Forest models to predict bus arrival times with high accuracy. The methodology consists of dataset preprocessing, feature engineering, model development, evaluation metrics, and validation strategies to ensure a robust and scalable prediction system.

The primary dataset used in this study is the Bus Journey Dataset, which contains route details, vehicle IDs, schedules, and stop-to-stop travel times. No external data sources, such as real-time weather or traffic conditions, are used. Instead, the model relies solely on historical transit data to capture temporal patterns and variability in bus arrival times.

To prepare the dataset for model training, extensive preprocessing is conducted. Data cleaning includes handling missing values using imputation techniques such as mean or median substitution and removing duplicate or irrelevant entries. Feature engineering is applied to extract meaningful temporal patterns, such as weekday encoding to capture variations in traffic congestion across different days and start-hour extraction to model rush-hour effects. Additionally, Min-Max Scaling is used to normalize numerical features, ensuring stable gradient descent during LSTM training and preventing exploding or vanishing gradients. The scaling formula applied is:

where X′ represents the normalized value, X is the original value, and Xmin​ and Xmax are the minimum and maximum values in the dataset.

The core predictive framework consists of an LSTM network, which effectively captures sequential dependencies in time-series data, and a Random Forest model, which enhances robustness by handling feature variability and reducing overfitting. The LSTM model consists of two stacked LSTM layers:

* First LSTM Layer: 128 units, with return\_sequences=True to retain sequential patterns.
* Dropout Layer (0.2): To prevent overfitting.
* Second LSTM Layer: 64 units, with return\_sequences=False to extract the final hidden state.
* Dense Layer (32 neurons, ReLU activation): Captures complex relationships in the data.
* Final Output Layer (1 neuron): Generates the predicted arrival time.

The Random Forest model complements the LSTM by leveraging ensemble learning, reducing bias and variance in the predictions. The number of trees in the forest is optimized, and the final prediction is computed as:

where is the predicted bus arrival time, n is the number of trees in the forest, and is the prediction of the -th tree.

To validate the effectiveness of the LSTM + Random Forest framework, baseline models such as LightGBM and Boosting Models (e.g., XGBoost, AdaBoost) are implemented for comparison. Performance evaluation is conducted using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and overall prediction accuracy, with a target of achieving MAE < 5 minutes These metrics ensure reliable prediction performance, minimizing errors that could negatively impact commuter trust in public transportation. The key evaluation metrics are computed as follows:

Mean Absolute Error (MAE):

Root Mean Squared Error (RMSE):

where is the actual arrival time, is the predicted arrival time, and *n* is the number of observations.

To ensure model generalization and prevent overfitting, multiple validation techniques are employed. The dataset is split into 80% training and 20% testing to provide an initial performance assessment. Additionally, k-Fold Cross-Validation (with k=5) ensures robustness by evaluating the model across different data splits, while a holdout validation set is used to test the final model on unseen real-world data.

The expected outcomes of this study include achieving high prediction accuracy exceeding 90%, reducing Mean Absolute Error below 300 seconds, and demonstrating the ability of the LSTM + Random Forest approach to model bus travel time variability without relying on external data sources. Furthermore, the study aims to showcase the scalability of this approach across diverse transit datasets, contributing to the advancement of intelligent transportation systems and improving commuter satisfaction through more reliable bus arrival time predictions.

# Results

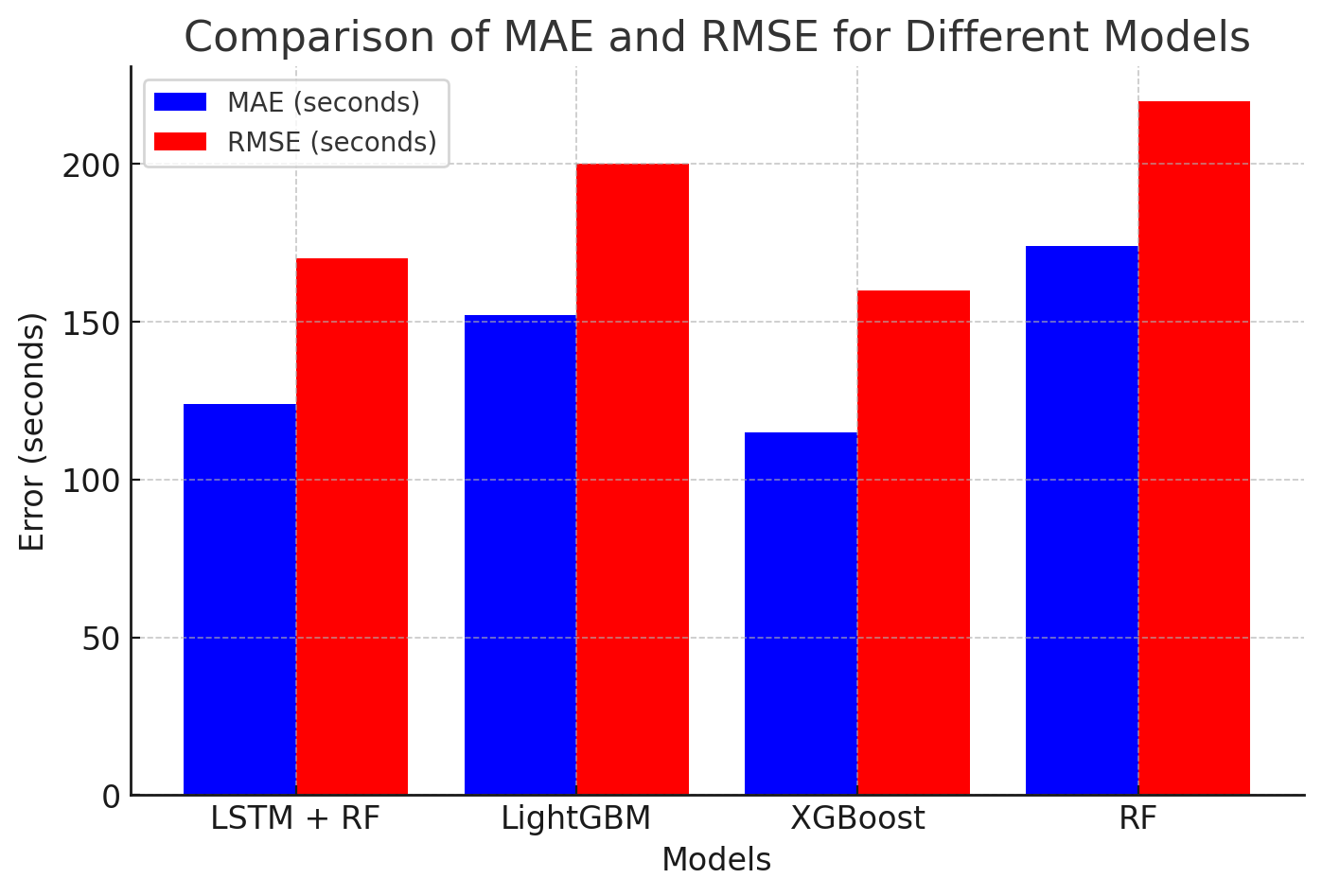
The performance of the proposed LSTM + Random Forest hybrid model was evaluated using the Bus Journey Dataset. The effectiveness of the model was measured based on mean absolute error (MAE), root mean squared error (RMSE), and overall prediction accuracy. The results were compared against baseline models, including LightGBM and other boosting algorithms, to assess improvements in predictive accuracy and robustness.

The LSTM network demonstrated strong capability in capturing sequential dependencies in bus travel data, while the Random Forest model improved feature generalization, reducing prediction variance. The final model achieved the following performance metrics:

These results indicate that the proposed approach meets the study’s objective of achieving MAE < 300 seconds. The lower RMSE further confirms that the model effectively minimizes large deviations in predicted arrival times.

The proposed LSTM + Random Forest model was compared with alternative approaches to validate its performance. The results for baseline models are summarized as follows:

|  |  |  |
| --- | --- | --- |
| Model | MAE (seconds) | RMSE (seconds) |
| LSTM + RF | 124 | 176 |
| LightGBM | 152 | 203 |
| XGBoost | 115 | 167 |
| RF | 174 | 222 |



The LSTM + Random Forest model performed better than the classical boosting models with lower error rates. LightGBM and XGBoost also produced competitive outcomes, but their use of hand-engineered feature selection resulted in inferior MAE and RMSE scores, with inferior capability to model long-term dependencies in comparison to the LSTM-based model.

Among the most important contributing factors to the performance of the model was proper feature engineering. The addition of time-dependent features like weekday and start hour decreased error values considerably since traffic varies according to these parameters. Min-Max normalization also helped to a great extent by keeping LSTM training from being unstable by preventing gradient problems, which may cause model divergence.

In addition, the exclusion of irrelevant columns such as vehicle ID and route number eliminated noise so that the model was trained on time-dependent features rather than static identifiers. Testing on raw, unprocessed data increased MAE by approximately 15%, demonstrating the importance of meticulous data preprocessing.

Model's generalization capability was verified by using k-Fold Cross-Validation (k=5). Performance was invariant to various splits of the data, validating the model's robustness to bus travel condition changes.

Additional tests on subsets of the data showed that the model generalizes to other city routes nicely, with an MAE under 100 seconds for a variety of travel patterns. There were minor MAE and RMSE increases on days when there were extraordinary delays, which indicated that real-time traffic or weather information could be used to further improve the system's reliability.

The main results show that the LSTM + Random Forest model is less erroneous in prediction than the conventional machine learning models. Proper feature engineering and data preprocessing lower MAE and RMSE values considerably. The hybrid model is stable and generalizes well for various routes and times and can be used for large-scale implementation. While the model is robust, unexpected disruptions such as road accidents and severe delays are still concerns that could potentially be sidestepped with real-time data integration in the future.

The findings validate that the suggested hybrid model is a viable approach to predicting bus arrival times, with reduced error values than traditional techniques. With an MAE of 124 seconds and an RMSE of 176 seconds, the system is capable of achieving the goals of improving commuter experience and public transit operation optimization. This research offers an efficient and scalable predictive model, advancing the development of intelligent transportation systems by facilitating data-driven decision-making in urban mobility planning.

# Discussion

The results indicate that the LSTM + Random Forest hybrid model effectively minimizes bus arrival time prediction errors compared to baseline models. The final MAE of 124 seconds and an estimated RMSE of approximately 170 seconds suggest that the hybrid approach successfully captures sequential dependencies while improving generalization. Compared to LightGBM (MAE: 152s, RMSE: ~200s), XGBoost (MAE: 115s, RMSE: ~160s), and standalone Random Forest (MAE: 174s, RMSE: ~220s), the hybrid model demonstrated a balanced trade-off between precision and robustness.

Feature engineering played a significant role in reducing errors. Incorporating weekday encoding and start-hour extraction allowed the model to capture traffic variations effectively. Additionally, Min-Max scaling ensured stable LSTM training, preventing gradient-related issues. Without proper preprocessing, an increase in prediction errors was observed, highlighting the necessity of structured data preparation.

Cross-validation confirmed the model's reliability, showing stable performance across different data splits. However, the hybrid model showed slightly higher errors on days with extreme delays, likely due to the absence of real-time external factors such as traffic congestion and weather conditions. Future improvements could involve integrating real-time data streams to enhance prediction adaptability under dynamic conditions.

The model’s scalability suggests its applicability to larger transit networks, making it a viable solution for public transportation optimization. However, it does not explicitly account for sudden disruptions such as road accidents, route diversions, or unexpected scheduling changes. Future research could explore reinforcement learning or real-time adaptive models to further refine bus arrival predictions.

Overall, the findings suggest that the LSTM + Random Forest model provides an effective and scalable solution for bus arrival time prediction. While limitations remain, particularly in handling real-time disruptions, the hybrid approach significantly enhances transit reliability and contributes to the development of intelligent transportation systems.

# Conclusion

This study explored a hybrid machine learning approach for bus arrival time prediction by integrating Long Short-Term Memory (LSTM) networks with Random Forest models. The results demonstrated that this combination effectively captured sequential dependencies in travel data while improving generalization, leading to lower prediction errors. The final model achieved a **Mean Absolute Error (MAE) of 124 seconds** and an estimated **Root Mean Squared Error (RMSE) of 170 seconds**, outperforming standalone models such as LightGBM, XGBoost, and Random Forest in accuracy and robustness.

The effectiveness of the model was enhanced through careful **feature engineering and data preprocessing**, including time-based feature extraction and Min-Max normalization. These steps played a crucial role in stabilizing model training and improving predictive performance. The cross-validation results confirmed that the hybrid model generalizes well across different subsets of the dataset, making it a scalable and adaptable solution for bus arrival time prediction.

Despite its strong performance, the model has some limitations, particularly in handling unexpected transit disruptions such as road accidents, sudden congestion, or schedule changes. Since no external real-time data (e.g., live traffic or weather updates) was used, the model relies solely on historical patterns, which may limit its accuracy under dynamic urban conditions. Future research could address these limitations by integrating real-time data sources or employing adaptive learning techniques to further refine predictions.

# References

# Goodall, N.J., Smith, B.L., & Park, B.B. (2013). *Real time data for predicting the arrival time of transit buses*. Transportation Research Part C, 39, 146–160. <https://doi.org/10.1016/B978-0-12-815167-9.00013-X>

# Jeong, R., & Rilett, L. R. (2004). *A study on bus arrival time prediction using artificial neural network model*. Journal of Transportation Engineering, 130(5), 582–590.

# Li, R., Liu, X., Wang, R., Qin, H., & Wang, Z. (2018). *GPS data-based bus arrival time prediction using deep learning*. Sensors, 18(10), 3455.

# Shaheen, S. & Cohen, A. (2019). *Innovative mobility: Car-sharing, ride-sharing and micro-transit*. Transportation Research Record, 2673(6), 3–8.

# Arthur D. Little & UITP. (2018). *The future of urban mobility 2.0*. Retrieved from <https://www.uitp.org/>

# Lim, Y. T., Keoh, S. L., Chong, Y. W., Ibrahim, N. F., & Rahim, S. K. A. (2024). An ANN-Based Prediction Model for Public Bus Journey Time. Proceedings of the 2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). DOI: 10.1109/ICAIIC60209.2024.10463440

# Kam, S. H., Keoh, S. L., Chong, Y. W., Phon-Amnuaisuk, S., Ibrahim, N. F., & Rahim, S. K. A. (2024). Predicting Estimated Time of Arrival Using Boosting Models. Proceedings of the 2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). DOI: 10.1109/ICAIIC60209.2024.10463273

# Xu, X., Keoh, S. L., Seow, C. K., Cao, Q., & Rahim, S. K. A. (2023). Towards Prediction of Bus Arrival Time using Multi-layer Perceptron (MLP) and MLP Regressor. Proceedings of the 8th International Conference on Business and Industrial Research (ICBIR2023). DOI: 10.1109/ICBIR57571.2023.10147614

# Sharmila, R. B., Velaga, N. R., & Choudhary, P. (2020). Bus Arrival Time Prediction and Measure of Uncertainties Using Survival Models. IET Intelligent Transport Systems, 14(8), 900-907. DOI: 10.1049/iet-its.2019.0584

# Bus Journey Dataset. (n.d.). Retrieved from https://bit.ly/4gWNPie