Bus arrival time prediction

|  |  |
| --- | --- |
|  | Adel Yerkebulan Astana IT UniversityAstana, Kazakhstan [220782@astanait.edu.kz](mailto:220782@astanait.edu.kz) |
|  |  |

Abstract Bus arrival time prediction plays a vital role in enhancing the efficiency and reliability of urban public transportation systems. This study applies various machine learning techniques to predict bus arrival times, using the MBTA Bus Arrival Departure Times 2024 dataset. Its methodology includes data preprocessing, feature extraction, and model evaluation across five different predictive approaches: Linear Regression, Random Forest, XGBoost, Long Short-Term Memory (LSTM), and Convolutional Neural Networks (CNN). The experimental results show that deep learning models, particularly CNN, outperform traditional statistical and ensemble methods, achieving lower prediction errors. Despite their success, these models face challenges such as high computational costs and potential overfitting. Our findings provide insights into the trade-offs between different modeling approaches and suggest future directions for enhancing prediction accuracy, including the integration of additional data sources and the development of hybrid models.

Keywords—bus arrival time prediction, machine learning, deep learning

# Introduction

Accurate bus arrival time prediction is a fundamental component of intelligent transportation systems (ITS), contributing to improved commuter experience, route planning, and transit efficiency. Urban commuters often experience delays due to unpredictable traffic conditions, scheduling inefficiencies, and system failures, leading to frustration and reduced trust in public transportation. To address these issues, researchers have explored machine learning-based predictive models that analyze historical and real-time data to provide more reliable bus arrival estimates.

Previous studies, such as Baimbetova et al. (2021) and Seitbekova et al. (2020), have demonstrated that real-time GPS data combined with machine learning algorithms can significantly enhance bus arrival predictions. However, a major limitation of these models is their lack of adaptability to different urban environments—what works well in one city may not generalize to another due to variations in traffic conditions, bus schedules, and infrastructure quality.

To overcome these challenges, more advanced deep learning techniques have been introduced. Rashvand et al.

(2024) and Lee & Yoon (2022) showed that models like LSTMs can effectively capture temporal dependencies in bus schedules, while CNNs can learn spatial-temporal patterns in transit data. However, these methods require large, high-quality datasets and significant computational power, which may not always be available in cities with limited open-data infrastructure.

This study aims to evaluate the effectiveness of different machine learning models in predicting bus arrival times using the MBTA Bus Arrival Departure Times 2024 dataset. Unlike prior studies that incorporate traffic and weather conditions, our approach focuses on historical arrival data due to the unavailability of external contextual information.

By systematically evaluating different models, this study contributes to the ongoing development of more reliable, scalable, and efficient bus arrival time prediction systems.

# Methodology

*2.1 Research Scope and Dataset*

The study focuses solely on historical bus arrival and departure data. The dataset used in this study is the MBTA Bus Arrival Departure Times 2024 dataset, sourced from the MBTA Open Data Portal. It includes:

* Bus schedules (planned arrival and departure times)
* Actual recorded arrival and departure times
* Stop sequences and identifiers (to track bus movement along routes)
* Route identifiers and trip IDs (to distinguish different routes)

By leveraging this dataset, we aim to evaluate the effectiveness of machine learning models in predicting bus arrival times without real-time contextual variables, focusing on the feasibility of using historical trends alone.

To formalize the problem, we define bus arrival time prediction as:

where:

* is the predicted arrival time for bus,
* represents input features such as scheduled arrival time, stop sequence, and route ID,
* is the error term capturing noise and unaccounted factors,
* is the predictive function learned by machine learning models.

2.2 Data Preprocessing

Raw datasets often contain inconsistencies such as missing values, outliers, and unstructured categorical features, which can negatively impact model performance. The following preprocessing steps were applied:

Handling Missing Values:

* Rows with critical missing values (arrival or departure times) were removed.
* Less critical missing values were imputed using median values based on similar bus routes and time slots.

Feature Engineering:

* Time-Based Features: Extracted hour of the day, day of the week, and weekend indicators.
* Route-Based Features: Encoded bus route IDs and stop sequences using one-hot encoding and embeddings for deep learning models.

This structured preprocessing ensures that all models receive cleaned, standardized input data, improving their predictive performance.

*2.3 Models*

Multiple machine learning models were implemented and evaluated to predict bus arrival times:

* Linear Regression (LR): Establishes a simple baseline model by fitting a linear relationship between input features and arrival time:
* Random Forest Regressor (RF): Uses an ensemble of decision trees to model non-linear relationships and capture feature interactions.
* XGBoost: An ensemble learning technique that enhances accuracy through gradient boosting, optimizing an objective function iteratively.
* Long Short-Term Memory (LSTM): A recurrent neural network (RNN) variant designed to capture temporal dependencies in sequential data.
* Convolutional Neural Network (CNN): Utilizes convolutional operations to detect spatial-temporal relationships in time-series data. The input sequence is reshaped into a matrix representation for convolutional layers.

*2.4 Evaluation Metrics and Validation*

The models were evaluated using:

Mean Absolute Error (MAE): Measures the average magnitude of prediction errors:

Root Mean Squared Error (RMSE): Penalizes larger errors more heavily, providing a more sensitive evaluation metric:

To ensure robust model performance, we employed:

* Train-test split: An 80%-20% split to separate training and evaluation data.
* Cross-validation: 5-fold cross-validation to assess generalization capability.
* Time-based splitting: To maintain temporal integrity, ensuring that training data precedes test data to avoid data leakage.

# Results

3.1 Model Performance Comparison

The performance of each model was assessed using RMSE and MAE, as shown in Table 1.

Table 1

|  |  |  |
| --- | --- | --- |
| **Model** | **RMSE** | **MAE** |
| Linear Regression | 4.8835 | 3.6424 |
| Random Forest | 5.3264 | 3.5640 |
| XGBoost | 5.2691 | 3.6030 |
| LSTM | 4.8508 | 3.6277 |
| CNN | 4.6956 | 3.4668 |

Deep learning models (LSTM and CNN) outperformed traditional and ensemble-based methods, achieving the lowest RMSE and MAE values. The CNN model exhibited the best predictive accuracy, likely due to its ability to extract hierarchical features from temporal data.

3.2 Feature Importance Analysis

A feature importance analysis using XGBoost highlighted the most influential factors in predicting arrival times. Figure 1 illustrates the top-ranked features.

A bar graph with text

AI-generated content may be incorrect.

Figure 1

Scheduled arrival time and route identifiers emerged as the strongest predictors. Meanwhile, stop sequences, directions identifiers and days of the week played a moderate role in influencing the arrival time predictions.

The study suggests that incorporating additional contextual factors, such as real-time traffic data, could further improve model accuracy.

3.3 Limitations and Future Directions

While the models performed well, there were several limitations:

* Exclusion of external variables: Due to data constraints, real-time traffic and weather conditions were not included.
* Computational cost: Deep learning models, while accurate, require significant computational resources, making real-time deployment challenging.
* Potential overfitting: LSTM and CNN models may have been overfitted due to limited training data.

3.4 Future Improvements

* Integrating additional external factors like real-time GPS and traffic data.
* Exploring hybrid models that combine deep learning and traditional approaches for better interpretability and efficiency.
* Optimizing computational resources to ensure scalability for real-world deployment.

# Discussion

The results of this study demonstrate the potential of machine learning models for bus arrival time prediction, with deep learning techniques, particularly CNNs, exhibiting the best performance. This section interprets these findings in relation to existing literature, discusses their implications for real-world deployment, and outlines possible improvements for future research.

4.1 Interpretation of Results

The CNN model achieved the lowest RMSE and MAE, surpassing both traditional regression-based models and tree-based ensemble approaches. This aligns with prior studies that emphasize the advantage of deep learning in capturing complex temporal dependencies. LSTM, another deep learning model, also performed well, although its slightly higher error rates suggest that CNNs may be more suitable for structured transit data where spatial relationships (such as stop sequences) are critical.

Despite their predictive power, tree-based methods such as XGBoost and Random Forest underperformed compared to deep learning models. This could be due to their inability to fully capture the sequential nature of bus arrival patterns, as they treat time-series data as independent instances rather than interdependent sequences.

One interesting finding is that Linear Regression performed comparably to tree-based models, indicating that bus arrival times may exhibit strong linear relationships with scheduled times and route information. However, its higher errors in peak-hour scenarios suggest a lack of robustness when dealing with unpredictable delays.

4.2 Implications for Public Transport Systems

The improved accuracy of CNN and LSTM models presents an opportunity for transit authorities to implement real-time bus arrival prediction systems with higher reliability. More accurate predictions could enhance passenger satisfaction by reducing wait times and enabling better trip planning.

However, deploying these models in real-world transit networks poses challenges:

* Computational constraints: Running deep learning models in real-time requires efficient hardware, which may not be feasible for all transit systems.
* Data dependency: The models rely on high-quality, comprehensive datasets. Incomplete or inconsistent data could lead to unreliable predictions.
* Scalability**:** While CNN and LSTM performed well on a single dataset, their effectiveness across different cities and transit networks remains uncertain.

4.3 Limitations and Future Research Directions

While this study provides valuable insights, certain limitations should be acknowledged:

1. The study primarily used schedule-based features, stop sequences, and route identifiers. The absence of real-time traffic, weather conditions, and unexpected disruptions likely reduced model accuracy.
2. The dataset may contain biases due to incomplete data collection or inconsistencies in reporting arrival times. Future work should explore methods to clean and augment datasets for greater reliability.
3. Deep learning models require significantly more computational power than traditional models. Future research should explore optimization techniques, such as model pruning or quantization, to reduce inference time.

To address these limitations, future studies should explore:

* The integration of real-time traffic data, GPS updates, and weather conditions to improve robustness.
* The use of hybrid models that combine CNN/LSTM architectures with rule-based or statistical approaches for better generalizability.
* Deployment of models in multiple urban environments to assess scalability and adaptability to different transit systems.

4.4 Conclusion

This study demonstrates that deep learning models, particularly CNNs, outperform traditional and ensemble-based methods in bus arrival time prediction. While challenges remain, the findings highlight the potential for AI-driven transit systems to enhance urban mobility. Future improvements, including real-time data integration and computational optimizations, could pave the way for more accurate and efficient public transportation networks.

##### References

1. Baimbetova, Aidana & Konyrova, Kulyash & Zhumabayeva, Aigerim & Seitbekova, Yerkezhan. (2021). Bus Arrival Time Prediction: a Case Study for Almaty. 1-6. 10.1109/SIST50301.2021.9465963.
2. Lee, Chanjae & Yoon, Young. (2022). A Novel Bus Arrival Time Prediction Method Based on Spatio-Temporal Flow Centrality Analysis and Deep Learning. Electronics. 11. 1875. 10.3390/electronics11121875.
3. Seitbekova, Y. et al. (2020) The bus arrival time prediction using LSTM neural network and location analysis, Journal of Advances in Technology and Engineering Research.
4. Rashvand, N., Hosseini, S. S., Azarbayjani, M., & Tabkhi, H. (2024). Real-Time Bus Arrival Prediction: A Deep Learning Approach for Enhanced Urban Mobility.