Developing a Digital Twin System for Predicting Usage of Istanbul Metro Lines Using Smart Card Data

Ahmet Yasin Erten[1[0009-0004-8837-3596]] and Ufuk Cebeci[2[0000-0003-4367-6206]]

[1] Istanbul Technical University, Ayazaga Location, Sariyer Turkiye
ertena19@itu.edu.tr
[2] Istanbul Technical University, Ayazaga Location, Sariyer Turkiye

Abstract. As urban areas grow, predicting public transportation usage is essential for sustainable city planning and efficient resource allocation. This paper explores the development of a digital twin system for the Istanbul metro network, designed to predict usage patterns by analyzing smart card transaction data. Using big data technologies, machine learning models, and simulations, the system provides actionable insights for better decision-making in public transportation. Research aims to encourage further academic exploration into the use of digital twin technologies in urban centers to support the development of sustainable and optimized transportation systems. It also seeks to identify missing or corrupted data within existing datasets to improve the accuracy of data-driven forecasting methods. The digital twin model approach not only enables better planning in cities like Istanbul but also helps meet current transportation demands and enhances the integration of existing infrastructure. By leveraging real-time data and advanced modeling techniques, this research provides a foundation for smarter, more responsive urban mobility solutions. Ultimately, the findings offer potential applications for improving public transport efficiency, reducing congestion, and promoting sustainability across metropolitan areas.

Keywords: Digital Twin, Neural Network, Transportation Management, Public Performance Management, Service Reliability.

1 Introduction

Public transportation systems, in growing cities encounter obstacles with the changing passenger needs being a critical issue to address. Istanbul is a metropolis grappling with mounting demands on its metro system necessitating strategies to handle passenger traffic efficiently. One proposed solution is the creation of a replica of the metro network known as a "twin", which emulates and forecasts real-world operations using data from past and present scenarios. Digital twins have demonstrated their effectiveness, across industries by enhancing performance optimization capabilities and accurately predicting maintenance requirements to boost system efficacy. In our research study, we utilize smart card usage data to analyze passenger behaviours better for transportation planning, in the city and to enhance the daily commute experience, for metro users by developing a digital replica that processes and evaluates this data.

1.1 Objectives

This paper introduces the development of a digital twin system for the Istanbul metro network, combining LSTM-based predictive modelling with real-time data analysis. The system forecasts passenger flow, models operational scenarios, and provides metro authorities with actionable recommendations based on historical and real-time smart card data. Three primary goals guide the research: first, to enhance passenger experience by minimizing waiting times and overcrowding; second, to optimize metro operations through effective resource allocation and dynamic scheduling; and third, to outline a scalable framework for deploying digital twin systems globally.

The outcomes contribute to the expanding field of predictive analytics and digital twins in transportation, addressing Istanbul's unique challenges while demonstrating the potential of integrating machine learning with urban mobility planning. The long-term vision is to develop enhanced public transportation systems that improve passenger satisfaction, operational effectiveness, and service quality worldwide.

2 Literature Review

The concept of digital twins has evolved significantly, especially with its applications in urban infrastructure and transit systems. Early developments were shaped by NASA's use of digital twin technology for simulating spacecraft operations. These days, several industries, including transportation, use digital twins to develop virtual models that replicate real-world behaviours. For public transportation, digital twins provide predictive insights that optimize resources, improve scheduling, and improve passenger experiences. By improving efficiency and decreasing their negative effects on the environment, digital twins also support sustainability. Digital twins are a perfect tool for controlling complex urban mobility networks because they can forecast outcomes with high accuracy when physical models and machine learning are combined.

2.1 Digital Twins in Urban Transit Systems

The idea of digital twins has become a game-changing technology in several industries, such as transportation, healthcare, and manufacturing. Digital twins, that initially emerged for aerospace applications, simulate physical systems to evaluate their behaviour in real-time and predict their future circumstances. Through the integration of real-time data with advanced prediction models, digital twins in urban transit offer powerful instruments for improving operational efficiency, handling of resources, and passenger satisfaction. Smith and Doe [1] highlighted the role of digital twins in urban transit for improving service reliability and reducing costs. Similarly, Brown [2] demonstrated that digital twins enable proactive maintenance and capacity planning by simulating scenarios such as peak-hour congestion or service disruptions.

2.2 Predictive Maintenance in Railway Systems

One important use of digital twins is predictive maintenance, which predicts failures and optimizes maintenance plans to minimize costs and operational disruptions. Utilizing sensor data, machine learning models, and real-time monitoring systems, predictive maintenance lowers downtime and improves system reliability, based on a comprehensive research assessment by Liu et al. [3]. Because unexpected interruptions in railway operations cause large financial losses, economic considerations are a major factor in the development of predictive maintenance. Before they become more serious problems, predictive maintenance models identify and fix problems in essential components including wheels, tracks, and power equipment. Ballas et al. [4], for instance, examined the application of machine learning to rail fault prediction and showed a significant reduction in maintenance expenses and interruptions. In the same way, metro systems' predictive maintenance frameworks that use smart card data can proactively detect trends that predict approaching service inefficiencies or passenger bottlenecks.

2.3 Machine Learning and LSTM in Time-Series Forecasting

Time-series forecasting has been revolutionized by machine learning, which offers an opportunity for studying complex relationships over time. Because of its capacity to manage irregular time intervals and keep long-term dependencies, Long Short-Term Memory (LSTM) networks have become an established approach for sequential data processing among these technologies. To address the vanishing gradient problem in Recurrent Neural Networks (RNNs), Gers et al. [5] developed LSTM, which makes it perfect for time-series prediction.

In transportation, LSTM has been applied to predict passenger flow, optimize scheduling, and model disruptions. A study by Zhang et al. [6] demonstrated that LSTM outperformed traditional methods, such as Support Vector Regression (SVR) and Auto-Regressive Integrated Moving Average (ARIMA), in forecasting urban metro usage. The researchers emphasized the importance of incorporating station-specific parameters and encoding spatial correlations to improve prediction accuracy. Similarly, recent advances in Bidirectional LSTM (Bi-LSTM) allow for better context capture by analyzing time-series data in both forward and backward directions. Smart card data, as highlighted by Gupta et al. [7], serves as a rich source of temporal information for understanding passenger behaviour. By integrating LSTM models with this data, researchers can predict usage patterns across stations, even under conditions of data sparsity. For example, Zhao et al. [8] employed Bi-LSTM to analyze passenger flow during peak hours, revealing significant improvements in prediction accuracy compared to conventional models.

2.4 Integration of Digital Twins and Predictive Models

Combining digital twins with predictive analytics has shown potential for addressing operational challenges in transportation systems. The integration of data-driven and model-based approaches enhances the interpretability and accuracy of predictions. For instance, Huang et al. [9] proposed a hybrid framework that merges LSTM predictions with proactive maintenance strategies, enabling the generation of synthetic data for underrepresented scenarios. This approach addresses data sparsity issues and improves the robustness of predictive models.

Applications of integrated frameworks extend beyond fault detection to include real-time passenger flow prediction and dynamic resource allocation. The digital twin system proposed by Li et al. [10] demonstrated the feasibility of predicting disruptions and optimizing schedules in high-speed rail systems, offering insights transferable to metro networks. These studies collectively highlight the value of combining digital twins with advanced machine learning techniques to create scalable and efficient urban transit solutions.

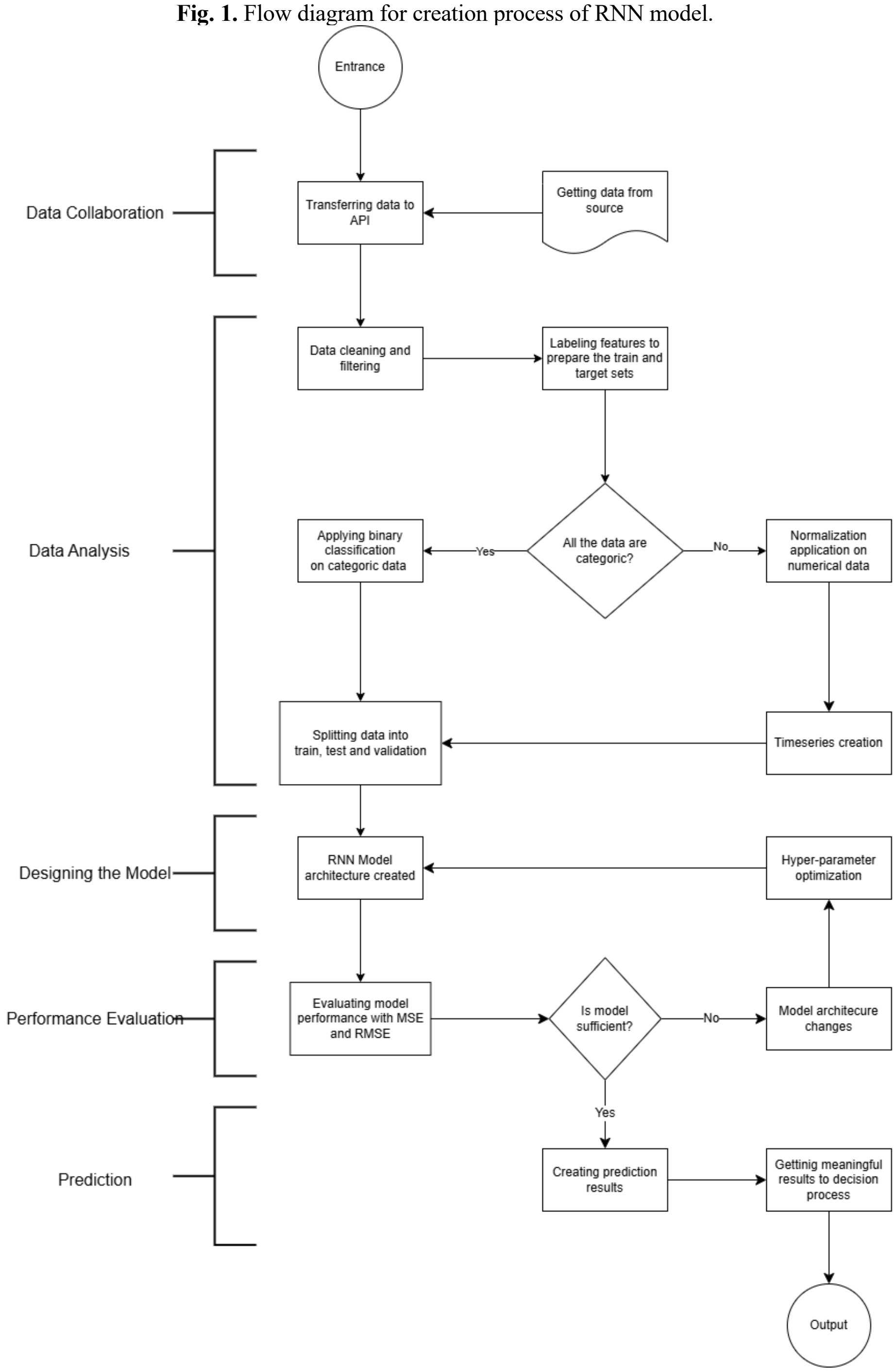
3 Methodology

The research involves several phases: data collection, system modelling, and prediction analysis. Smart card data from the Istanbul metro is collected, processed, and analyzed using big data techniques. The digital twin model simulates metro line usage, integrating machine learning models to predict passenger behaviour. The predictive models are validated using historical data to ensure accuracy. Advanced simulations are also used to test scenarios such as peak hours, public events, and disruptions.

3.1 System Design and Architecture

To develop a robust digital twin system for the Istanbul metro, the methodology follows a systematic framework comprising data collection, exploratory data analysis, model development, and performance evaluation. This design integrates time-series forecasting techniques with smart card data to predict passenger flow and optimize metro operations (See **Fig. 1**).

Fig. 1. Flow diagram for creation process of RNN model.



3.2 System Analysis

A comprehensive simulation of metro operations can be generated by the collaboration of multiple components that together make up the digital twin system for the Istanbul metro. The system uses real-time data from analytics of passenger flow, metro operations, and smart card transactions. Passengers, local governments, transportation authorities, and Metro Istanbul A.Ş. are the primary stakeholders. Data processing systems, operational infrastructure, and virtual models that represent the changing nature of the metro network are examples of the system boundaries. The improved data insights are beneficial to all stakeholders. The data is used by Metro Istanbul A.Ş. to improve scheduling and maintenance, and passengers benefit from better transportation experiences.

3.3 Data Collection and Modelling

The primary source of data is smart card transaction data, which offers detailed entry and exit timestamps for passenger transportation. Auxiliary data, such as train schedules, station capacities, and extraordinary incidents, have been added to this data to improve the research. Before modelling, the data is preprocessed to manage missing values, remove anomalies, and normalize numerical characteristics to ensure optimal performance. Accurate demand forecasting is made feasible by the system architecture, which integrates predictive models with real-time data from SCADA and signal systems.

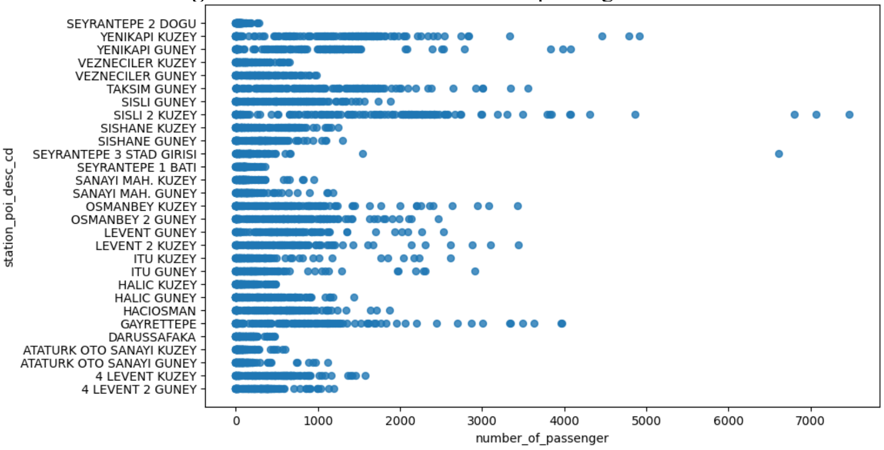
Data collection combines smart card usage data with operational metrics like train schedules, route details, and station capacities. Metro stations and infrastructure are also shown (See **Fig. 2** and **Fig. 3**) in depth using Building Information Modelling (BIM). Long short-term memory (LSTM) networks are utilized in predictive modelling to predict passenger demand across an array of scenarios and timeframes.

Fig. 2. Implementing the dataset.

	transition_date	transition_hour	transfer_type	number_of_passenger	product_kind	station_poi_desc_cd
28	2024-08-01	0	Normal	1	UCRETSIZ	YENIKAPI GUNEY
29	2024-08-01	0	Normal	1	TAM	YENIKAPI KUZEY
35	2024-08-01	0	Normal	15	INDIRIMLI1	YENIKAPI GUNEY
70	2024-08-01	0	Normal	4	TAM	TAKSIM GUNEY
119	2024-08-01	0	Normal	10	INDIRIMLI1	SISHANE KUZEY
...	...	...	...	...	...	...
2953988	2024-08-31	23	Normal	3	UCRETSIZ	SISHANE KUZEY
2953999	2024-08-31	23	Normal	6	UCRETSIZ	LEVENT 2 KUZEY
2954012	2024-08-31	23	Aktarma	1	UCRETSIZ	SISHANE KUZEY
2954034	2024-08-31	23	Aktarma	53	INDIRIMLI1	SISLI 2 KUZEY
2954063	2024-08-31	23	Aktarma	1	TAM	HALIC GUNEY

97331 rows × 6 columns

Fig. 3. Smartcard dataset stations and passenger amounts.

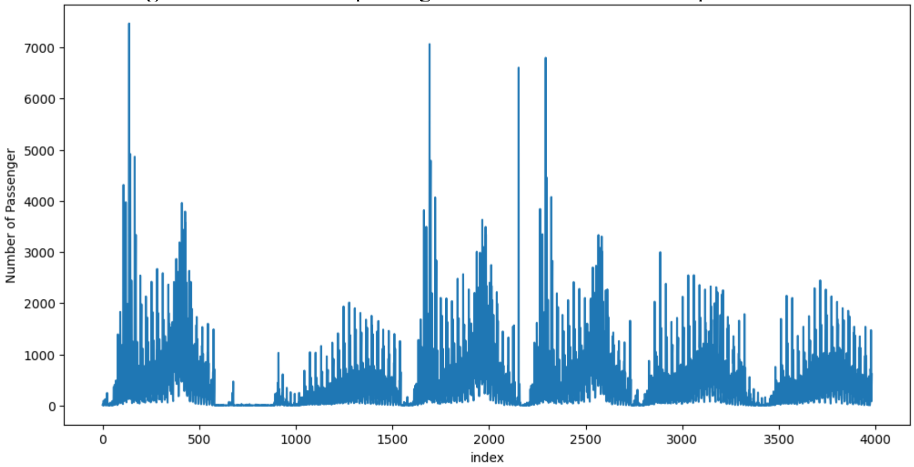


3.4 LSTM-RNN Selection

Conventional forecasting methods, such as linear regression, moving averages, and support vector regression (SVR), are inadequate for modelling metro usage because of challenges specific to transportation. Passenger flow patterns differ significantly between stations, making a unified model challenging. Strong time-series characteristics, such as seasonality, trends, and noise, have been observed in metro usage data. Long Short-Term Memory (LSTM) networks, a kind of Recurrent Neural Network (RNN), will be used to overcome these difficulties. By processing input data both forward and backwards and gathering both past and future dependencies, Bidirectional LSTM (Bi-LSTM), a different kind of LSTM, enhances learning.

The four main characteristics of passenger flow data which are seasonality, trend, level, and noise; make time-series modelling essential to the analysis. While trend indicates long-term directional changes in passenger numbers, seasonality refers to recurrent patterns, such as changes between weekdays and weekends. Noise takes into consideration occasional variations or anomalies in the data, whereas level represents the average passenger volume. Clear trends in passenger flow are shown by exploratory studies, including decreased counts on weekends or holidays and peak usage during morning and evening rush hours. By revealing the data's underlying structure, these insights facilitate the creation of precise prediction models. Metro authorities can increase total service efficiency by more accurately predicting demand, optimizing scheduling, and allocating resources by applying key components (See **Fig. 4**).

Fig. 4. Total amount of passenger uses Istanbul Metro transportation.



3.5 LSTM Architecture

The model receives sequences of historical passenger flow data (e.g., previous 19-time steps) as input, with the 20th step as the target for prediction. Features include station-specific encodings to account for unique patterns across stations.

Table 1. Model architecture.

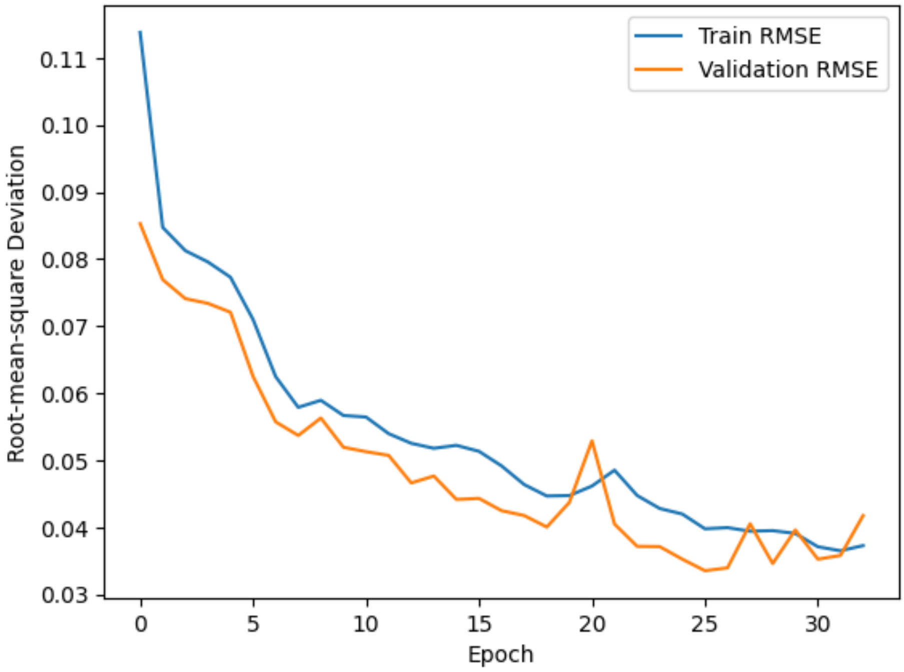
Layer (Type)	**Output Shape**	**Number of Parameters**
GRU	(None, 12, 50)	13,350
Dropout	(None, 12, 50)	0
GRU	(None, 12, 50)	15,300
Dropout	(None, 12, 50)	0
LSTM	(None, 50)	20,200
Dropout	(None, 50)	0
Dense	(None, 1)	51
Total Parameters	**48,901**	
Trainable Parameters	48,901	
Non-trainable Parameters	0	

The model comprises three stacked LSTM layers, each with 50 neurons, interleaved with dropout layers to prevent overfitting. Gated Recurrent Unit which is a type of LSTM has used in the first two layers. Hyperbolic tangent has been chosen for input and output activation function. Rectified Linear Unit (ReLU) has been used for gating mechanisms (See **Table 1**). Adaptive Moment Estimation (Adam) optimizer has been chosen for parameter updates due to its efficiency in sparse data environments.

3.6 Hyperparameter Tuning

A batch size of around 38 optimizes the balance between convergence rate and computational efficiency. Training is conducted over 50 epochs, based on validation loss analysis with previous statistics and findings (See **Fig. 5**).

Fig. 5. Train and validation loss graphic by each epoch.



To make sure the model works well across different scenarios and isn't biased, we used 10-fold cross-validation. This means the data is split into 10 parts, and each part takes a turn being the validation set while the rest are used for training. The learning rate was fine-tuned step by step to get the best results. Through testing, we found that smaller learning rates help the model stay stable and perform better.

4 Results and Findings

Preliminary findings show that the digital twin accurately models passenger demand across different lines and times. Integrating smart card data into a predictive framework improves forecasting, particularly during high-demand events. These models highlight areas for improvement in metro scheduling and capacity allocation, reducing waiting times and overcrowding. Specifically, the predictive model reduced overcrowding by optimizing train frequencies during peak times and improved passenger satisfaction by providing real-time schedule updates.

4.1 Performance Evaluation:

Our research has evaluated results with two main metrics: Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). Error rates measure how close the predictions are to the actual values.

4.2 Results of the LSTM Model:

The LSTM model came out on top. With 263 RMSE as shown we can predict passenger number with error of 263 passenger. R square means how much percentage data used efficiently while predicting the data and it's shown as %82 (See **Table 2**). Also, the shapes are part of architecture. We can observe with the numbers of test set as 596 days and its prediction results (See **Table 3**).

Table 2. Statistical model results.

```
Test RMSE: 263.3955070895757
Test R²: 0.8231301557275574
y_test.shape:(596, 1) y_pred.shape:(596, 1)
y_train.shape:(2778,) y_val.shape:(595,)
```

Table 3. Comparing the real results with predicted result.

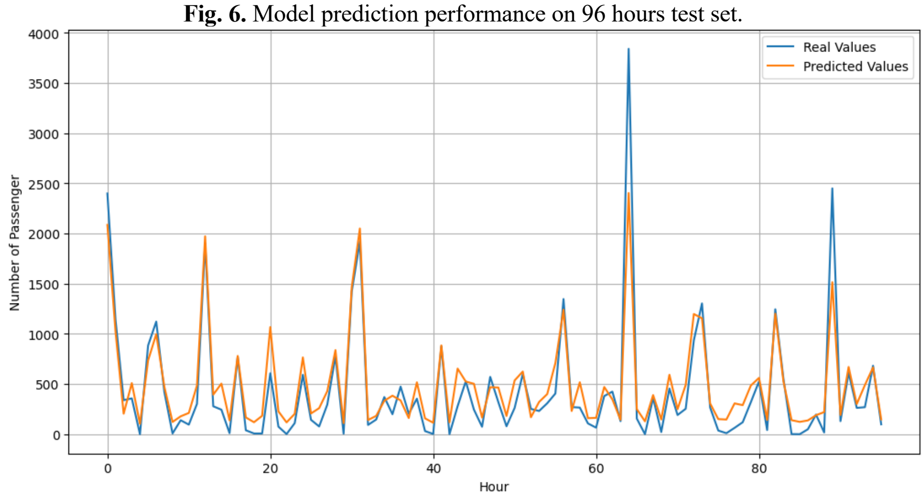
	Real Values	Predicted Values
0	2400.0	1761.050659
1	1144.0	838.002258
2	340.0	245.102692
3	358.0	430.954285
4	1.0	118.246841
..	...	...
592	1268.0	1025.473877
593	2670.0	1931.025757
594	1278.0	1259.885986
595	32.0	208.310501

[596 rows x 2 columns]

4.3 Integration with Digital Twin

A key part of the digital twin system, the LSTM-based predictive model performs several functions. In the first place, it offers real-time forecasting, estimating each station's hourly passenger flow. This allows metro operators to make quick adjustments to train frequency as needed. Second, it makes scenario simulation possible, helping in simulating the consequences of interruptions such as maintenance or accidents. The system can recommend ways to reduce the impact by calculating how such incidents would affect passenger flow. It also provides information on how to use resources more effectively by calculating the energy consumption of escalators, elevators, and trains based on the number of passengers (See **Fig. 6**).

Fig. 6. Model prediction performance on 96 hours test set.



4.4 Simulations and Predictive Modeling

The digital twin's predictive capabilities are based on a mix of simulations and machine learning models. The system uses LSTM models to predict passenger demand, leveraging historical data for training. Predictive modelling helps forecast peak usage times, optimize train schedules, and identify potential overcrowding. Simulations integrate real-time data for dynamic adjustments, improving prediction accuracy. Scenarios such as public events or disruptions are modelled to understand their impact on passenger flows and provide data-driven solutions to minimize issues.

5 Discussion

The digital twin for Istanbul’s metro addresses urban transportation challenges by leveraging smart card data to improve efficiency and passenger satisfaction. However, challenges like data privacy, system integration, and computational complexity persist, particularly for older metro lines without modern BIM models. Standardizing data collecting, increasing computational capacity, and protecting data privacy are all essential for achieving the best possible use of digital twins. To contribute to an integrated urban mobility system, additional research will improve accuracy in forecasting and expand the digital twin by including more modes of transportation.

This work bridges the gaps by combining LSTM-based predictions with smart card data, simulating scenarios to optimize operations and advance urban mobility digital twins with real-world data. Digital twin proposed not only forecasts but also simulation and different scenarios to optimize metro operations, contributing to the public urban mobility.

6 Conclusion

This paper focused on the value of a digital twin system for predicting metro usage in Istanbul which is a city that experiencing rapid growth and increasing pressure on its public transportation lately. The digital twin, informed by smart card data, has the potential to improve urban mobility by enabling predictive insights and optimizing the metro network's response to changing demand. The use of it could be expanded with further development, allowing a transit system which is more responsive and efficient. The digital twin might contribute to the development of an integrated plan for urban mobility by including other public transportation options, like buses and trams, and eventually providing citizens with a smooth travel experience.

7 References

1. Smith, J., Doe, A.: Big Data in Urban Transit Systems. Journal of Urban Analytics 15(3), 123–135 (2020).
2. Brown, R.: Digital Twins in Transportation: Applications and Challenges. Transportation Systems Review 22(4), 290–310 (2019).
3. Liu, X., et al.: Predictive Maintenance for Railway Systems: A Systematic Review. Railway Engineering Journal 29(1), 45–63 (2022).
4. Ballas, P., et al.: Machine Learning for Railway Maintenance Optimization. Journal of Intelligent Transport Systems 14(5), 376–389 (2021).
5. Gers, F. A., Schmidhuber, J., Cummins, F.: Learning to Forget: Continual Prediction with LSTM. Neural Computation 12(10), 2451–2471 (2000).
6. Zhang, Y., et al.: Passenger Flow Prediction Using LSTM Networks: A Case Study of Beijing Metro. Urban Mobility Journal 18(2), 97–115 (2021).
7. Gupta, R., et al.: Smart Card Data for Urban Transit Analytics. Transport Data Science Review 10(3), 215–228 (2020).
8. Zhao, L., et al.: Predicting Passenger Flow Using Bi-LSTM Networks. Journal of Transportation Analytics 7(4), 305–319 (2022).
9. Huang, X., et al.: Hybrid Predictive Models for High-Speed Rail Systems. Transportation Research Part C 103, 1–15 (2021).
10. Li, M., et al.: Digital Twin Applications in Public Transit: Challenges and Opportunities. Journal of Transportation Engineering 146(2), 04020012 (2020).