Asia-Pacific Journal of Information Technology and Multimedia 15(1) 2026: 193 – 210
https://doi.org/10.17576/apjitm-2026-1501-11

Modeling Rail Ridership using Long Short-Term Memory and Gated Recurrent Unit: A Case Study in Malaysia

Pemodelan Penumpang Rel Menggunakan Long Short-Term Memory dan Gated Recurrent Unit: Kajian Kes di Malaysia

Tok Chia Wen[1], Wong Zi Ming[1], Nor Azuana Ramli[1], Nur Haizum Abd Rahman[1], Wan Nur Syahidah Wan Yusoff[1], Mohd Azri Rosli[2] & Nadzri Hj. Yusof[2]*

[1]Centre for Mathematical Sciences, Universiti Malaysia Pahang Al-Sultan Abdullah, Lebuh Persiaran Tun Khalil Yaakob, 26300 Kuantan, Pahang, MALAYSIA.
[2]Prasarana Malaysia Berhad, B-20-1, Level 20, Menara UOA Bangsar, No. 5, Jalan Bangsar Utama 1, 59000 Kuala Lumpur, MALAYSIA.

**Corresponding author: azuana@umpsa.edu.my*

Received 17 September 2025
Accepted 6 May 2026, Available online 30 June 2026

ABSTRACT

Efficient management of rail transportation systems is important for achieving sustainable urban mobility. Identifying the factors influencing rail ridership and accurately forecasting future ridership patterns are critical for optimizing operational strategies, infrastructure planning, and resource allocation. This study investigates the primary determinants of rail ridership and develops predictive methods to forecast ridership trends. The methodology uses multivariate linear regression to identify significant predictors, along with deep learning approaches such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models for multivariate time series forecasting. The LSTM model obtained a mean absolute percentage error (MAPE) of 21.32%, an R-squared (R^2) of 0.74, and a root mean squared error (RMSE) of 1,009.03; whereas the GRU model obtained a MAPE of 17.36%, an R^2 of 0.83, and an RMSE of 803.09. Both models effectively captured the complex patterns in the ridership data. However, the GRU model achieved a slightly more accurate result than the LSTM model. The successful use of deep learning models in this research study indicates that they may provide a strong method for representing the complexities of ridership data and identifying valuable areas for public transportation systems' research and future operational improvements.

Keywords**:** deep learning; multivariate linear regression; rail ridership; predictive modeling.

ABSTRAK

Pengurusan sistem pengangkutan rel yang cekap adalah penting untuk mobiliti bandar yang mampan. Mengenal pasti faktor yang mempengaruhi penumpang rel dan meramalkan corak penumpang pada masa hadapan dengan tepat juga penting bagi mengoptimumkan strategi operasi, merancang infrastruktur dan memperuntukkan sumber dengan berkesan. Kajian ini bertujuan untuk meneroka faktor utama yang mendorong penumpang rel serta membangunkan model ramalan bagi meramal tren aliran penumpang rel. Kaedah yang digunakan ialah regresi

linear berbilang variasi untuk mengenal pasti faktor utama jumlah penumpang, manakala teknik pembelajaran mendalam, Memori Jangka Pendek Panjang (LSTM) dan unit berulang berpagar (GRU) digunakan untuk ramalan siri masa berbilang variasi. Model LSTM mencapai min ralat peratusan mutlak (MAPE) sebanyak 21.32%, R-kuasa dua (R2) sebanyak 0.74, dan punca ralat min kuasa dua (RMSE) sebanyak 1009.03; sementara itu, model GRU menunjukkan prestasi yang lebih baik dengan MAPE sebanyak 17.36%, R2 sebanyak 0.83, dan RMSE sebanyak 803.09. Kedua-dua model sangat efektif dalam mengenal pasti corak kompleks dalam data penumpang, dengan model GRU lebih tepat sedikit berbanding model LSTM. Kejayaan penggunaan model pembelajaran mendalam menggariskan potensi model ini untuk menggarap corak penumpang yang rumit, membuka jalan untuk penyelidikan masa lanjut pada masa depan serta meningkatkan operasi dalam sistem pengangkutan awam.

Kata kunci**:** pembelajaran mendalam; regresi linear berbilang variasi; penumpang rel; pemodelan ramalan.

INTRODUCTION

Malaysia’s National Artificial Intelligence Roadmap (AI-RMap) was launched in March 2021 by the Minister of Science, Technology, and Innovation (MOSTI) (MOSTI, 2021). With its determined goal of positioning Malaysia as a leading technology hub by 2030, this tactical blueprint is a significant leap in Malaysia’s technological development and innovation journey. An AI-RMap has been created to support Malaysia's goal to become an international leader in artificial intelligence (AI). It will also encourage the creation of new economic opportunities through the use of AI by developing a sustainable ecosystem of innovation regarding AI. The structure of the AI-RMap consists of 6 main foundations: the creation of governance for AI, the advancement of R&D, the development of new digital infrastructure to enable AI, the development of talent in AI, the integration of AI into social systems, and the building of an AI national innovation ecosystem. The collective goal of these strategies is to create a foundation or framework for developing AI in all sectors of Malaysia's economy and society.

Investing in research and development (R&D) for AI is very important in many industries, and one example of this is in the transportation industry. By putting more importance on R&D in AI, Malaysia can create new ideas for better efficiency, safety, and reliability of its transportation systems. Dikshit et al. (2023) explain that AI will be a major advance in transportation because businesses are using AI-based predictions to make their service deliveries better, while at the same time, decreasing the amount of people experiencing traffic in cities. Because of the rapid growth that Malaysia's transportation industry (mostly led by Prasarana Malaysia Berhad (Prasarana)) is experiencing, placing more focus on the development of AI is extremely relevant. This is because both urbanization and economic development are driving major growth in this sector, especially in areas like Selangor and Kuala Lumpur. Figure 1 shows how the population of Kuala Lumpur has grown since 1950 and is projected to continue growing until 2030. The growth from 1950 up to now has been quite consistent, with the greatest rates of growth occurring in the last several decades. Based on current projections, by the end of 2030, the population of Kuala Lumpur will be more than 8.8 million; this represents a 2.25% increase over the previous year (DOSM, 2024).

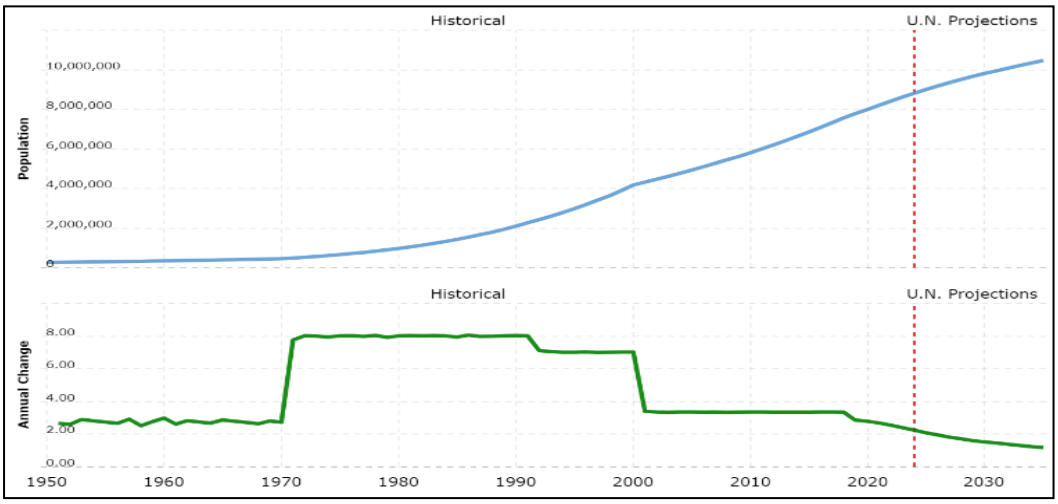

FIGURE 1. The population of Kuala Lumpur, year by year, from 1950 to 2030 (Macrotrends, 2024)

Urbanization is driving demand for public transit systems because they can reduce traffic congestion, lower environmental impact, and provide mobility and access in large metropolitan areas. This has resulted in a steady increase in the use of public transport, with Prasarana reporting an average of 1.1 million commuters each day, including 835,119 on rail and 258,710 on buses (MyRapid, 2024). The increase in public transport use illustrates the importance of public transport in terms of daily commuting and urbanization. There is also a need to continue to develop infrastructure to keep up with population growth.

Malaysia has made considerable investments historically in the construction of the public transportation infrastructure to modernize and grow its public transportation network. One way that this has been done is through the introduction of new trains for operations on the Kelana Jaya Light Rail Transit (LRT) line and the Ampang LRT line, as well as by the commissioning of the Mass Rapid Transit (MRT) Putrajaya Line, which has significantly increased rail service capacity and frequency (MyRapid, 2024). In 2017, nearly 60% of Malaysians used public transportation, which reflects the importance of public transportation in supporting sustainable urban mobility. However, prior research has illustrated how there are ongoing challenges in meeting the increasing demand for public transportation in urbanized areas due to difficulties in maintaining schedules, inefficiencies in scheduling, and capacity limitations (Dimitriou & Gakenheimer, 2011; World Bank, 2017; Asian Development Bank, 2020).

Although recent developments have led to better public transport systems with greater standards of quality, many remain challenged to offer reliable and consistent service levels to meet fluctuating demands. Due to dramatically increasing demand for public transit, the associated infrastructure and services are experiencing significant strains on resources (maintenance and scheduling) and capacity constraints. In particular, rail systems (such as LRTs) are experiencing far greater ridership than other forms of public transportation. Although weekday ridership can generally be handled by the train system without disrupting normal operations, peak periods during holidays or major events can create a significant increase in ridership on trains and through stations. This will lead to overcrowding, which may create issues for passengers and disrupt passenger service. Therefore, this study identified what factors affect deviations from typical ridership patterns.

Our hypothesis is that abnormal spikes in ridership can occur during public holidays, special events, and adverse weather conditions. Although previous studies have examined the impacts

of holidays and weather on train ridership (Zhou et al., 2017) and changes in ridership based on events such as COVID-19 (Halvorsen et al., 2023), very few studies have been conducted on these three factors affecting train ridership in Malaysia. The findings of this study not only contribute to the literature on transportation but also have practical implications for citizens and transportation services. It can also significantly improve the daily commute experience for citizens and enhance the efficiency of transportation services by mitigating the factors that lead to overcrowding and service delays, especially at busy LRT stations.

By advancing AI research and development initiatives, this study also aims to accurately forecast ridership trends, particularly by identifying abnormal patterns. Typical forecasting models used to estimate ridership (e.g., autoregressive integrated moving average (ARIMA), regression, and tree-based models) do not adequately represent non-linear behavior or long-run temporal dependencies, nor do they allow for an adequate representation of factors external to ridership (e.g, weather, holidays, and socio-economic indices). Recurrent Neural Networks (RNNs) help with issues related to the sequential dependencies of data, but are still affected by the vanishing gradient problem. Through the use of Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, which are designed to address the challenges of complexity, non-linearity, and long-range temporal dependencies seen in sequential data, the goal of this research is to provide improved accuracy and stability in the forecast of rail ridership. It is hoped that this will address issues associated with crowding, delays, and overall passenger safety, thereby increasing the reliability and sustainability of Malaysia's urban transport systems and giving the Malaysian public more confidence in the delivery of reliable and safe public transportation services in the future.

This paper's goal is to develop an extensive methodological framework to analyze the factors influencing deviations in public transport ridership patterns in Malaysia. For this study, real-time ridership data are collected from the Malaysian public transport operator, Prasarana, along with external datasets including publicly recognized holidays, major events, and weather data. The collected data are then preprocessed and analyzed to identify relevant correlations across the dataset. Following this, an analysis of the multivariate linear relationship between rail ridership and load factor is conducted using multivariate linear regression models with advanced machine learning algorithms. The combination of traditional statistical regression analysis and deep learning algorithms helps develop accurate forecasts of public transport ridership variations and strategies to provide a higher level of service to customers using public transport systems. The following methodology section describes the steps for data collection, analysis, and model development in detail.

LIMITATIONS OF PRIOR STUDIES

Studies on predictive models for rail ridership are not new, including those that incorporate weather and special events. However, traditional statistical techniques, such as autoregressive integrated moving average (ARIMA), regression, and Random Forest, have limitations because they fail to capture nonstationary and nonlinear demand dynamics. Models that incorporate many different types of variables have been found to perform poorly when there is an imbalance in the data. This is especially true when the data contains very few occurrences of a particular outcome (Zhao et al., 2021). The integration of weather data into the model has been investigated by both Guang et al. (2017) and Alver and Ercan (2024), and they both found a lack of quantitative analysis, indicating that the addition of weather data would improve the model's likelihood of success. The authors Zou et al. (2024) and Toque et al. (2018) also noted that adding an additional variable (individually measured), such as weather, to a traditional model will present difficulties when capturing the spatiotemporal dynamics of the data. This

problem has been documented in many prior studies and is summarized in Table 1 below, along with a list of authors who have researched various aspects regarding this issue.

TABLE 1. Research gap

Aspect	Limitations in Previous Studies	Supporting Studies	Research Gap Addressed by Current Study
Data Availability and Quality	Traditional models, such as ARIMA and regression, struggle with imbalanced and sparse data, particularly during rare or irregular events. Reliance on Automatic Fare Collection (AFC) data lacks complete event and weather annotation.	Zhao et al. (2021); Ni et al. (2017)	Develop a multivariate, event-aware framework that integrates both structured ridership data and contextual variables (such as holiday, event type, venue, and weather) to overcome data sparsity.
Weather Integration	Weather effects are often modeled using simplistic or lagged variables, providing limited insight into temporal–spatial dynamics. Many models lack a quantitative evaluation of weather impact.	Xu et al. (2019); Guang et al. (2017); Zou et al. (2024)	Apply deep learning models (LSTM, GRU) capable of learning complex non-linear relationships between weather patterns and ridership fluctuations over time.
Model Architecture and Temporal Dynamics	Traditional and shallow ML models assume linearity and stationarity, limiting their ability to represent long-term temporal dependencies or multi-factor interactions.	Toque et al. (2018); Guang et al. (2017); Zhao et al. (2021)	Utilize LSTM and GRU architectures to capture long-term dependencies, sequential memory, and interaction effects between multiple time-dependent variables.
Performance Evaluation	Prior studies often report qualitative performance claims without consistent quantitative benchmarking (RMSE, MAPE, R^2). Limited transparency restricts replicability and comparison.	Chen et al. (2020); Zou et al. (2024)	Conduct systematic performance evaluation using multiple error metrics (MAPE, RMSE, R^2) for robust benchmarking between LSTM and GRU models.

METHODOLOGY

DATA COLLECTION AND ACQUISITION

This research used secondary data downloaded from various sources, including Microsoft Azure Storage Explorer and the company website. The collected data included ridership, rail load factors, events, public holidays, and weather. The primary focus was on the number of riders departing from and arriving at Bukit Jalil station, while the supported data included the rail load factor for the Ampang Line. The rail load factor is the dependent variable, and the ridership variables are treated as independent variables in multivariate analysis. Dependent variables included events and activities in Bukit Jalil, public holidays, and weather in Kuala Lumpur. The ridership transaction data was obtained from Prasarana's database, which holds daily records of train riding transactions for all riders who have stopped (tapped out) at their final destination after tapping on at the beginning of their travels. The records consist of Kajang Line transaction records as well as transaction records from other lines and are categorized in the same manner as shown in Table 2. All records included consist of complete ridership data from January 2023 through April 2024 and will be presented during the Results and Discussion section through the preparation stage.

TABLE 2. Characteristics of daily rail ridership transactions

Variable	Description	Data Type
RECORD_DATE	Date and time of each transaction.	Integer
IDENTIFIER	The identified transaction for the person.	String
ORIGIN	The station that person departs from.	String
DESTINATION	The station that person arrived at.	String

The public holiday data for Kuala Lumpur in 2023 and 2024 was extracted from the website https://publicholidays.com.my/kuala-lumpur/ and compiled into an Excel file, including the URLs of the holidays. The Excel file of event data was manually generated from the ridership line chart at Bukit Jalil station. Peak spikes that are indicative of a significant event were identified using the line chart via the ridership line chart at Bukit Jalil. Furthermore, the types of events, for example, the Black Pink concert on March 4, 2023, the Coldplay concert on November 22, 2023, and the Malaysia Cup Football Match Final in December 2023, were derived from the website. The data is readily available throughout the line chart, as seen in Figure 2 of the appendix. It is important to consider the weather, as it has a significant impact on ridership (Zhou et al., 2017). In this case, rainfall data was identified by utilizing the Malaysian Meteorological Department (MetMalaysia) Web Service API, which provides an accessible and programmable means of accessing the weather data. The use of this API to extract raw data from its endpoints involves using web scraping techniques. The MetMalaysia Web Service API has a limit of 50 records per request, so to compile the entire dataset, multiple requests were required. The MetMalaysia Web Service API provides detailed weather data and guidelines on specifying the location to search for and defining the different data types, helping you gain the best understanding of the content within the API. Subsequently, the raw data were processed and integrated to form a consolidated dataset primed for analysis and application. Finally, the Rapid Rail team has provided the rail load factor dataset. This file contains data on the load factor among the following rail lines: the Kelana Jaya line, Ampang line, Monorail, and MRT Kajang line. The rail load factor percentage can be calculated using the following equation:

$$\text{Rail Load Factor (\%)} = ((\text{Total Ridership})/(\text{Total Capacity})) \times 100 \quad (1)$$

The rail load factor percentage is divided into three periods for each line. For instance, the AM% column is calculated for the peak time from 7:00 am to 9:00 am on weekdays, the PM% column for the peak time from 5:00 pm to 7:00 pm, and the NPH% column for the non-peak time hours. In this study, only the Ampang line data columns were extracted and used as response variables in a multivariate linear regression to identify the significant predictors influencing rail ridership and rail load factor. However, this dataset only consists of 2024 data, so other data were filtered from January 1, 2024, to April 30, 2024, to satisfy this objective.

MODELING

Various models were employed to achieve the objectives of this study, including multivariate linear regression, Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU). This study aims to utilize these models to acquire insight, anticipate ridership, and support decision-making based on the data's outcomes and objectives.

A. Multivariate Linear Regression

Multivariate Linear Regression is a statistical analysis technique used to model the relationship between multiple dependent and numerous independent variables, an extension of multiple regression. Multivariate regression analysis is used to forecast the values of two or more dependent variables from a set of independent variables (Janssen et al., 2019). The objective of this model is to evaluate if the independent variables have statistical significance and influence on the dependent variables. The method estimates a coefficient (or slope/weight) for each independent variable to show the expected change of the dependent variable with a one-unit change in that independent variable when all other independent variables are held constant. The general equation for a multivariate regression model is:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{2}$$

where

$$\boldsymbol{Y} = \begin{bmatrix} Y_{11} & \cdots & Y_{1m} \\ \vdots & \ddots & \vdots \\ Y_{n1} & \cdots & Y_{nm} \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} X_{10} & \cdots & X_{1r} \\ \vdots & \ddots & \vdots \\ X_{n0} & \cdots & X_{nr} \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} \beta_{01} & \cdots & \beta_{0m} \\ \vdots & \ddots & \vdots \\ \beta_{r1} & \cdots & \beta_{rm} \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_{11} & \cdots & \varepsilon_{1m} \\ \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \cdots & \varepsilon_{nm} \end{bmatrix}$$

$\boldsymbol{Y}$ are the dependent variables, $\boldsymbol{X}$ are the dependent variables, $\boldsymbol{\beta}$ are the unknown parameters, and $\boldsymbol{\varepsilon}$ are the model's errors.

B. Deep Learning Models

The Long Short-Term Memory (LSTM) and gated recurrent unit (GRU) models were chosen as the study's method. Both methods are improved versions of recurrent networks (RNNs), as they have been particularly successful across various tasks due to their ability to distinguish between recent and earlier examples by assigning different weights to each, while forgetting unimportant information for predicting future output. In this sense, both methods are better at handling long input sequences than RNNs, which can only memorize short sequences. For LSTM, the initial step is to decide whether to keep or discard information from the previous timestamp. The forget gate equation is as follows:

$$f_t = \sigma(X_t * U_f + H_{t-1} * W_f) \tag{3}$$

where f_t is forget gate, X_t is the input to the current timestamp, U_f is the weight associated with the input, H_{t-1} is the hidden state of the previous timestamp, and W_f is the weight matrix associated with the hidden state. Then, a sigmoid function is applied to it. If f_t is 0, the network will forget everything, and if f_t is 1, the network will remember nothing.

$$\begin{aligned} C_{t-1} * f_t &= 0 \quad \text{... if } f_t = 0 \\ C_{t-1} * f_t &= C_{t-1} \quad \text{... if } f_t = 1 \end{aligned} \tag{4}$$

where C_{t-1} is the cell state at the current timestamp. The input gate is then utilized to measure the significance of the additional data carried by the input, as shown in the equation below:

$$i_t = \sigma(X_t * U_i + H_{t-1} * W_i) \tag{5}$$

where i_t is the input gate. The new data that must be supplied to the cell state is now a function of a hidden state at timestamp t-1 and input x at timestamp t.

$$N_t = \tanh(X_t * U_c + H_{t-1} * W_c) \quad (6)$$
$$C_t = (f_t * C_{t-1} + i_t * N_t) \quad (7)$$

where N_t is referred to as new information and C_t represents long-term memory. The following equations are the output equations, where o_t is the output gate, H_t is the hidden state, and Output is the prediction.

$$o_t = \sigma(X_t * U_o + H_{t-1} * W_o) \quad (8)$$
$$H_t = o_t * \tanh(C_t) \quad (9)$$
$$Output = Softmax(H_t) \quad (10)$$

GRU is similar to an LSTM with a gating mechanism that allows for input or forgetting certain features (Gers et al., 1999). However, it lacks a context vector or output gate, resulting in fewer parameters than an LSTM (WildML, 2015). Both methods were compared to determine which provided the best results for forecasting ridership, using mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE). The equations for MAE, RMSE (Wilmott & Matsuura, 2005), and MAPE (Montaño Moreno et al., 2013) are presented as follows:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t| \quad (11)$$
$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n}} \quad (12)$$
$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \text{ x } 100\% \quad (13)$$

where y_t is the actual value and $\hat{y}_t$ is the predicted value at data point, t.

RESULTS AND DISCUSSION

Following data collection, the next stage is data preparation, which is essential for cleaning and refining the datasets prior to analysis and model development. This section outlines the data preprocessing procedures applied to all datasets and presents the results obtained from modeling with the described methods. The analysis of the most effective modeling approach is discussed at the conclusion of this section.

DATA PREPROCESSING

During data collection, the raw ridership data for each month were split into two: one for the Kajang line and another for other rail lines excluding the Kajang line. After the data was cleaned and combined from both sources into one full monthly dataset per line, the combined datasets needed to be further checked for any erroneous values. There were several reasons some of the ridership data were invalid; for example, the times listed at the LRT stations were originally not open outside normal business hours. Normally, an LRT station should be listed each day to be open from 6 AM to 12 AM (midnight). Therefore, any invalid times should be removed as part of the verification process. For example, if Merdeka Day in Malaysia is on August 31 and Prasarana has its LRT stations open from 6 AM to 2 AM for those celebrating Merdeka Day, Prasarana will not have ridership records for those arriving at 3 AM.

The data was grouped according to the station, and the data was aggregated by day to generate ridership counts for every date at each station after cleaning the dataset of invalid entries. The date column was formatted to yyyy-mm-dd for all dates in the standardized format. The ridership datasets were consolidated into a single dataset for comprehensive analysis after each month's data between 2023 and April 2024 was preprocessed separately. The datasets were cleaned of duplicates to ensure data quality. The study's analysis included the date range from January 1, 2023, to April 30, 2024. After all datasets had completed cleaning, the final cleaned version of each dataset was saved as a new CSV file. The date column in the public holiday dataset and the date column in the weather dataset were also standardized in order to merge both datasets for ease of integration. All datasets were merged to form a complete dataset for modeling and analysis following the thorough cleaning process. Due to the lack of event data for the analysis portion of the study, the analysis was conducted using the ridership dataset for the Bukit Jalil station. The ridership dataset was filtered for trips where the origin or destination of the trip was Bukit Jalil station to ensure that the data would meet the objectives of the study.

Several additional variables were also incorporated into the data through feature engineering methods (e.g., day of the week, day type, quarter of the year). The final dataset contained a significant amount of additional contextual data, which will improve future analysis or build better models. The main benefit of using one-hot encoding on these categorical variables was that they were converted from categorically described values to binary-coded variable data. This process increased the information value of each of the data points as well as made possible correlation analyses. During the data preparation processes, it was noted that data relating to the rail load factor only contained valid dates in 2024.

CORRELATION ANALYSIS

The correlation analysis is displayed in the Appendix. The variable “existEvent” exhibits high correlation, contributing 0.5, while the event location, specifically the national stadium, contributed 0.7 to the correlation with ridership. The correlations exceed 0.01, with the ridership target variable (“count”) further analyzed and modelled. In order to eliminate any weak but possibly significant correlations being excluded from the analysis, a cutoff level of 0.01 was chosen as the minimum value of correlation required to qualify as a correlation for the analysis. This cutoff is relatively low but was selected because rail ridership is a multi-factorial issue (i.e., unique factors have small effects at the individual level, which can wax large when looked at in the context of many complicated nonlinear multivariable relationships).

MULTIVARIATE DATA ANALYSIS

Multivariate linear regression is used with four dependent variables: rail load factor (AM%, PM%, and NPH%) and ridership. For predictors, public holidays, events, weather, quarters, days of the week, types of events, places of events, event days, sports days, and countries of participating teams in multi-day events are used. Of the four multivariate regression models analyzed, only three have high multiple R-squared values of 0.9037, 0.9009, and 0.8734 for ridership, AM%, and PM%, respectively. The models are significant based on the p-value (<0.05) from the F-statistic test (ANOVA). Multivariate test statistics are necessary to determine which predictors should be included in a multivariate regression model. Since it involves multiple responses, modified hypothesis tests, such as the ANOVA() function, are used to assess whether predictors contribute significantly to the models. Based on the results shown in Figure 3, the following factors have p-values less than 0.05 (the significance level):

holiday, event, multi-day event, Friday, Sunday, Saturday, concert, the specific location of the event (the National Stadium), and the second sports day. Therefore, these variables are significant predictors in this multivariate model.

```
Type II MANOVA Tests: Pillai test statistic
                                  Df test stat approx F num Df den Df    Pr(>F)
isHoliday                          1   0.61063   37.247      4     95 < 2.2e-16 ***
existEvent                         1   0.32468   11.418      4     95 1.310e-07 ***
`more than 1 day`                  1   0.50336   24.071      4     95 9.086e-14 ***
Korea                              1   0.00404    0.096      4     95  0.983392
Malaysia                           1   0.02865    0.700      4     95  0.593577
UK                                 1   0.03264    0.801      4     95  0.527239
FGM_rain                           1   0.02149    0.522      4     95  0.720076
FSIGW_rain                         1   0.06451    1.638      4     95  0.171141
DayOfWeek_Fri                      1   0.17120    4.906      4     95  0.001222 **
DayOfWeek_Mon                      1   0.01865    0.451      4     95  0.771095
DayOfWeek_Sat                      1   0.69928   55.228      4     95 < 2.2e-16 ***
DayOfWeek_Sun                      1   0.71787   60.432      4     95 < 2.2e-16 ***
DayOfWeek_Thu                      1   0.03111    0.763      4     95  0.552124
DayOfWeek_Tue                      1   0.01387    0.334      4     95  0.854468
`quarter_Quarter 1`                1   0.04088    1.012      4     95  0.405140
TYPE_Concert                       1   0.49088   22.899      4     95 2.882e-13 ***
`sport day_1.0`                    1   0.00037    0.009      4     95  0.999846
`PLACE_National Stadium, Bukit Jalil` 1   0.61413   37.800      4     95 < 2.2e-16 ***
`sport day_2.0`                    1   0.46874   20.955      4     95 2.083e-12 ***
`sport day_3.0`                    1   0.01888    0.457      4     95  0.767037
`sport day_4.0`                    1   0.04093    1.014      4     95  0.404435
`sport day_5.0`                    1   0.00022    0.005      4     95  0.999943
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 3. Table summary of the ANOVA test

The significant predictor variables identified in Figure 3 were further analyzed to determine whether they fit as well as a model with all predictors. One way to do this is to fit a smaller model and then compare it to the larger model using the ANOVA() test. The ANOVA test's results show that the p-value is greater than 0.05 (the significance level), indicating that the result is insignificant. Hence, there is not enough evidence to support that the model with only a few predictors fits as well as the model with all predictors at the 5% significance level. Another method is used to conduct the same test to further validate the results, which involves using the linearHypothesis() function. This function can perform the test without fitting a separate model and returns all four multivariate test statistics.

```
Multivariate Tests:
                 Df test stat approx F num Df den Df  Pr(>F)
Pillai            1 0.0645090 1.637736      4     95 0.17114
Wilks             1 0.9354910 1.637736      4     95 0.17114
Hotelling-Lawley  1 0.0689573 1.637736      4     95 0.17114
Roy               1 0.0689573 1.637736      4     95 0.17114
```

FIGURE 4. Table summary of the linearHypothesis test

The results in Figure 4 showed that the p-values of all the tests are insignificant at the 5% significance level. Therefore, the model with only a few predictors does not fit as well as the model with all predictors at the 5% significance level. Thus, the model with fewer predictors cannot be used for prediction. Hence, the list of the predictor variables models that can be used for prediction is shown in Figure 5.

Coefficients:

	count	AM_percent	PM_percent	NPH_percent
(Intercept)	6.575e+03	4.796e+01	4.865e+01	3.217e+01
isHoliday	-2.735e+03	-3.599e+01	-3.221e+01	-2.938e+00
existEvent	-1.575e+04	-5.082e+00	-5.172e+00	-1.126e+01
`more than 1 day`	1.834e+04	2.957e+00	1.874e+00	3.970e+00
Korea	-2.958e+02	1.736e+00	-8.671e-01	-2.380e+00
Malaysia	-1.173e+03	4.419e-02	-2.240e-01	6.405e+00
UK	1.620e+03	-1.174e+01	-1.672e+01	3.555e+00
FGM_rain	1.156e+02	6.170e-01	1.800e+00	-1.426e+00
FSIGW_rain	-7.088e+02	1.736e+00	-8.671e-01	-1.380e+00
DayOfWeek_Fri	-8.559e+02	-6.163e+00	-1.438e+00	1.531e+00
DayOfWeek_Mon	-3.656e+02	-1.429e+00	-2.468e+00	-1.712e+00
DayOfWeek_Sat	-2.715e+03	-4.583e+01	-4.376e+01	2.949e+00
DayOfWeek_Sun	-2.871e+03	-4.653e+01	-4.468e+01	-5.036e+00
DayOfWeek_Thu	-4.263e+02	-1.292e+00	-3.152e+00	-1.437e+00
DayOfWeek_Tue	-3.482e+02	-5.135e-02	-6.671e-01	-1.131e+00
`quarter_Quarter 1`	8.386e+01	-5.663e-01	-1.594e+00	-1.770e+00
TYPE_Concert	1.947e+04	1.785e+00	2.745e+00	7.296e+00
`sport day_1.0`	2.656e+02	1.219e+00	1.150e+00	-4.748e-01
`PLACE_National Stadium, Bukit Jalil`	2.333e+04	1.174e+01	1.672e+01	8.445e+00
`sport day_2.0`	1.364e+04	5.198e-01	2.278e-01	-4.603e-01
`sport day_3.0`	2.036e+03	3.112e+00	7.711e-01	1.339e+00
`sport day_4.0`	5.441e+02	-2.759e+00	7.485e+00	1.307e+00
`sport day_5.0`	1.680e+01	9.487e-01	3.329e-01	-1.305e-01

FIGURE 5. Linear Regression Model for each dependent variable

As shown in Figure 5, external factors had a significant impact on rail ridership behavior. In terms of the holiday effect, all holidays showed the same negative effect on daily ridership due to fewer commuters travelling for work on holidays compared to other days of the week. Conversely, multi-day events had a strong positive effect on ridership since large/recurring events stimulate demand, especially when these events are located at a large venue (like the National Stadium at Bukit Jalil), where ridership increased most significantly. It is also found that heterogeneous impacts of rainfall: where rain at some locations increased train travel, possibly because commuters avoided their private vehicles, while at other locations, it decreased train usage. These differences can be attributed to local variations in commuting trends. Overall, day-of-week variables had a negative relationship to the baseline, with Friday and Saturday showing the most significant decrease in ridership, confirming that most people commute Monday-Thursday rather than on weekends. In addition to the seasonal effect, event variables changed the phase of demand distribution by moving passenger loads from AM peak demand periods to PM and non-peak times, particularly for concerts and sports events, where the majority of the ridership increase occurred during the PM peak.

The overall results indicate that traditional linear models have limitations because they do not account for complex, non-linear, and context-dependent factors affecting ridership. Thus, this suggests that the use of advanced sequential models (i.e., LSTM and GRU) would be beneficial in capturing long-term dependencies and interactions between temporal, spatial and external factors.

MODELING USING DEEP LEARNING MODELS

Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models were utilized to identify the most effective prediction model, as introduced earlier in this paper. These techniques were applied to develop a model capable of accurately forecasting future ridership and detecting irregular patterns. This can be accomplished by integrating key variables influencing ridership, as determined through the multivariate linear regression analysis conducted in the preceding section. For the predictive ridership variable, the ridership data needed to be normalized before being used in the predictive model due to the presence of binary predictors and the extreme variability of ridership data. The total number of ridership ranged

from 2000 to 20000, which cannot be directly compared; hence, a Min-Max normalization was applied to the ridership variable, allowing the data to be placed in the normalized range [0,1]. The dataset had both seasonal influences and weekly patterns, and the sequence length was set to 7 (one week of data) due to the number of weeks in the dataset. After multivariate analysis, only significant variables were selected as predictors, with ridership as the target variable. Models received predictors and target variables separately when passing through the sequence creation function, allowing modeling to learn effectively from the dataset and also predict accurately based on patterns in the data. The dataset was split into training (80%) and testing (20%). The modeling process began with a GRU layer of 54 units, with ReLU activation, so that temporal dependencies of data could be captured. The last layer was a dense layer, condensing all the information learned into one output to allow for making an appropriate prediction for regression.

The model was compiled with ADAM optimizer and Mean Squared Error (MSE) as the loss function (enabling weights to adjust to optimize the model to minimize the prediction error). Training was conducted on the training dataset for 50 epochs with a batch size of 32, with 10% of the data being used for validation during training to measure performance and ensure generalizability of the model results. The model's predictive accuracy improved with each iteration of training as the internal parameters of the model were improved upon based on the errors calculated previously. Once trained, the model's performance was assessed against the validation dataset. The results of the model's performance can be seen in Figure 6..

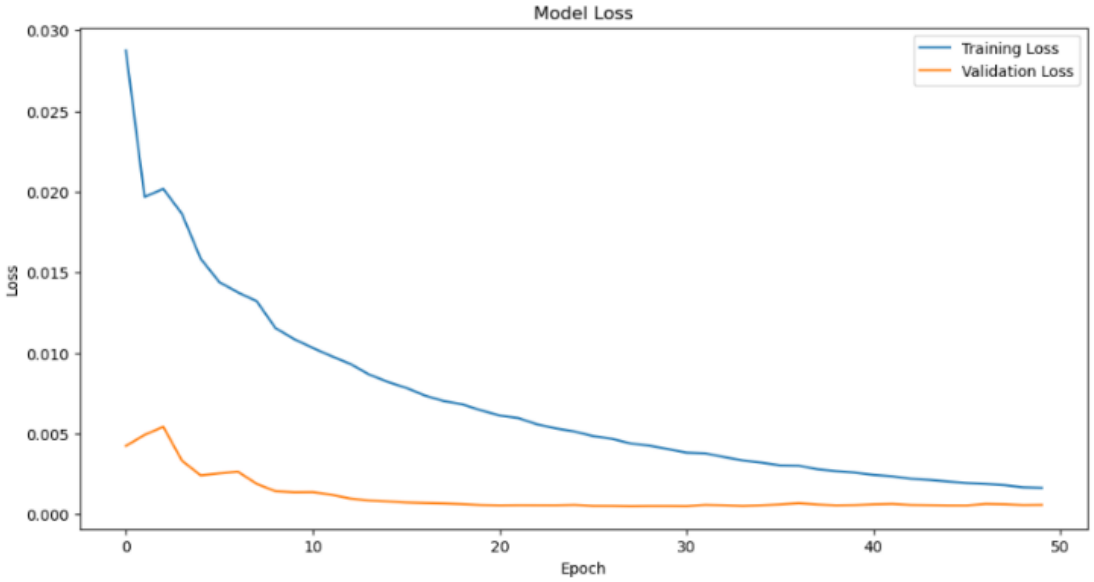

FIGURE 6. GRU graph of model loss between training and validation

As shown in Figure 6, the plot displays the model’s loss and compares the training and validation losses. Both the training and validation losses decrease steadily as the epochs progress. The training loss consistently decreases with each epoch, and the validation loss shows a decreasing trend without any significant increase. The training and validation losses have a moderate separation metric distance, indicating sufficient training and that the model has either not been overfit or underfit. The model has been shown not to have had its training process disrupted, as there is no intersection of the training and validation losses throughout the entire model training period. To validate the model's training, the model was used to estimate ridership. The estimated ridership had a significant correlation with the actual ridership (Figure 7). The estimated values (represented by the orange line) were nearly

equivalent to the actual ridership (represented by the blue line), indicating that the model performed well during training and is able to predict on previously unobserved data. Although there are some discrepancies, all differences fall within the acceptable range of difference.

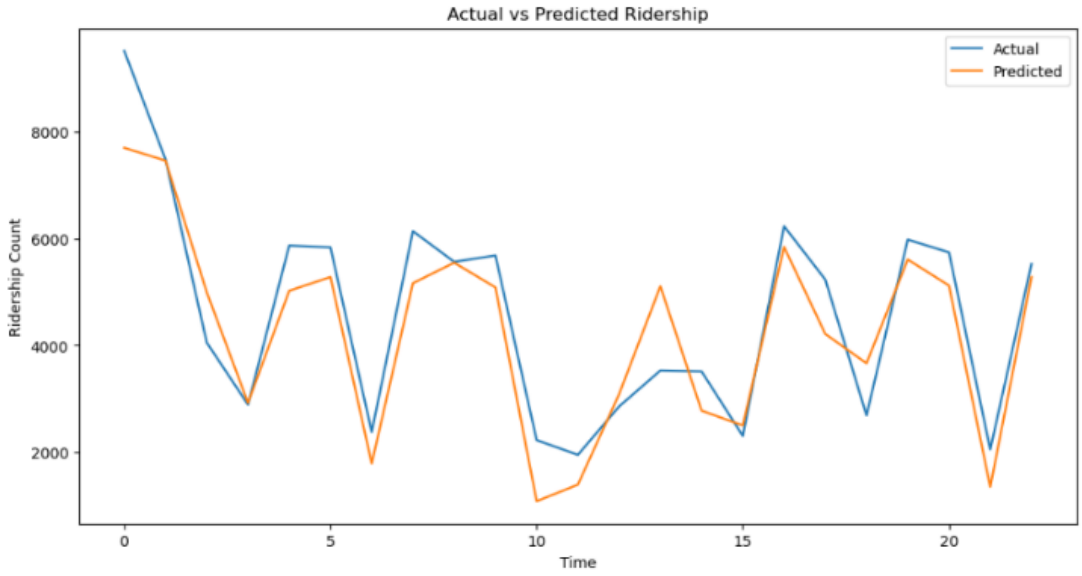

FIGURE 7. Graph of comparison between actual and predicted ridership using GRU

The LSTM model was configured with an input shape suitable for sequential data, specifying sequence length and feature dimensions. The first layer is an LSTM layer, with 64 LSTM cells that use ReLU activation functions and help to capture temporal correlation patterns in the data. The second layer is the dense output layer, which condenses all of the information learned from all of the LSTM cells into a single predicted output for regression. To optimize the model, both the ADAM optimizer and mean squared error (MSE) loss were used for compilation of the model, similar to the GRU-based models. The model was trained for 50 epochs with a batch size of 32. The model training results are shown in Figure 8, where it can be seen that the training and validation losses decline at nearly the same rate, i.e., both continue to decrease with each successive epoch. During training, the majority of the training loss is less than that of the validation loss (providing some evidence that the model was trained appropriately). The gap between the two training (training) and validation (validation) values is within an acceptable range, indicating that the model has been trained effectively and will likely not have issues related to overfitting or underfitting. In addition, the training and validation losses do not cross over each other (further evidence that model training has been done correctly).

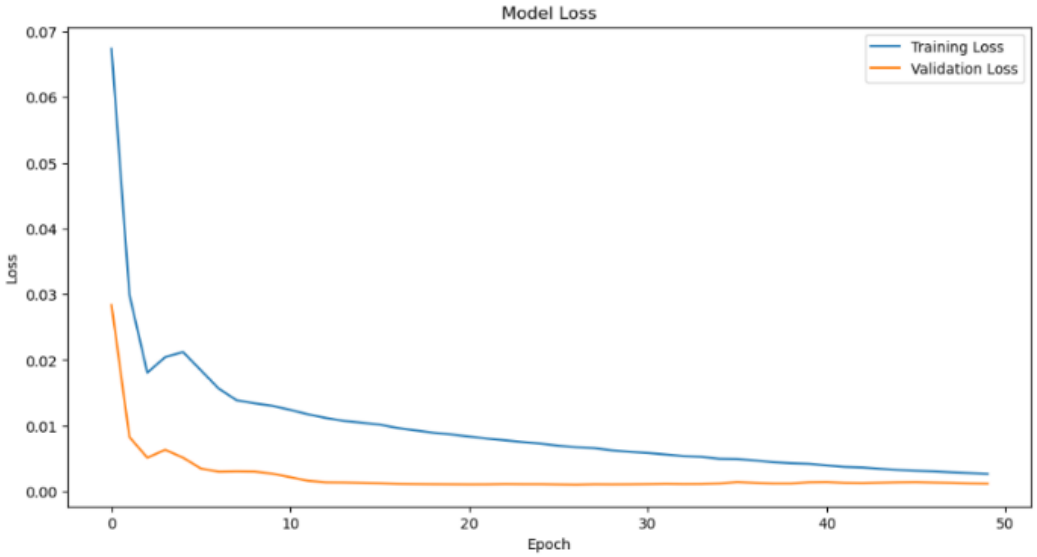

FIGURE 8. LSTM graph of model loss between training and validation

Figure 9 plots the predicted and actual values. Although slight deviations exist between predicted and actual values, these differences remain within an acceptable range. Overall, the LSTM model accurately predicts ridership across various scenarios. As it is not easy to compare performance through plots, the models were evaluated using R-squared (R2), mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE). Table 3 compares these two models.

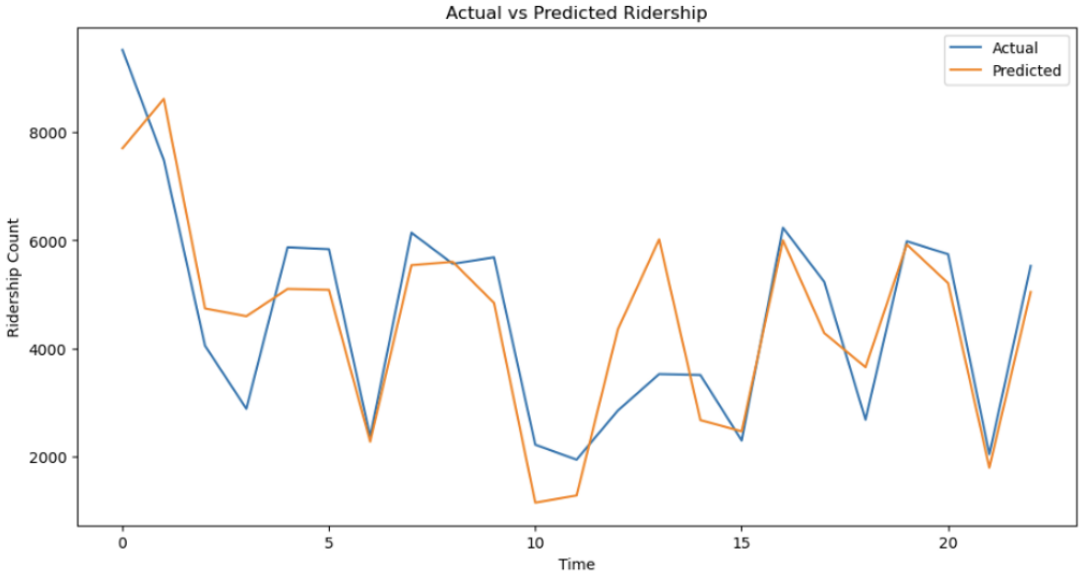

FIGURE 9. Graph of comparison between actual and predicted ridership using LSTM

In comparing GRU and LSTM, GRU shows higher predictive accuracy. GRU has a slightly higher R2 (0.8331) than LSTM (0.7366), indicating that GRU captures approximately 83% of the variability in the ridership data and can more accurately model/identify patterns and trends than LSTM. In addition, GRU has a slightly lower mean absolute percentage error (MAPE) of 17.36% than LSTMs' MAPE of 21.32%, indicating better relative accuracy. Finally, GRU has a lower RMSE (803.09) than the naïve baseline, further verifying that GRU is a more accurate model for ridership than LSTM.

TABLE 3. Comparison between LSTM and GRU

Performance Metrics	LSTM	GRU
R-squared (R^2)	0.7366	0.8331
MAE	810.0074	659.3715
RMSE	1009.0250	803.0852
MAPE (%)	21.3242	17.3642

CONCLUSION

In summary, this study aimed to identify the factors with the strongest influence on rail passenger numbers and to predict future rail passenger numbers using a range of analytical techniques, including multivariate linear regression and deep learning models such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). As a result of the multivariate linear regression analysis, a number of external or context-specific factors were identified as major contributors to shaping rail passenger trip demand. Statistically significant evidence suggested that holidays will generally lead to a reduction in the number of rail passengers, as there will be a lower level of commuting activity associated with these holidays. By contrast, larger-scale type events, particularly those held at major venues, for example, the National Stadium, Bukit Jalil, will lead to higher levels of rail passenger numbers, particularly during evening hours and off-peak times. There were also mixed results associated with weather, as there was evidence that rainfall in certain regions will increase the demand for rail travel, as people shift from the use of private vehicle transport. However, in other areas, it is marginally reducing demand for rail travel by discouraging it. Similarly, Friday and Saturday saw the largest reductions in rail passenger numbers compared to mid-week days, further demonstrating that weekday travel is an essential component of rail traffic. The results of this study confirm the complex interplay between several variables that impact rail ridership and how different types of travel drivers, such as calendar effects, weather, location, and events, interact and combine to influence demand for rail travel.

The development of the LSTM and GRU models and the prediction of ridership trends are based on these findings. The LSTM and GRU both underwent extensive model training and evaluation, with rigorous comparisons using stringent key metrics. LSTM scored an MAPE of 21.32%, R2 value of 0.74, and RMSE of 1009.02, indicating a good fit to the data and predictive value accurate to a reasonable extent. The GRU model produced superior results to the LSTM model, scoring an MAPE of 17.36%, an R2 value of 0.83, and an RMSE of 803.09. Both models captured the complex patterns in the ridership data very well; however, the GRU model produced superior predictive performance than the LSTM model. A visual inspection of model predictions vs. actual data provides further validation of the performance of both models, as both accurately forecast the ridership statistics for the two-month evaluation period. These results emphasize that deep learning methods, specifically the GRU model, are a good method for multivariable time series forecasting in rail ridership analysis.

The results and approach of this study are similar to the results and methods of other studies completed before it used traditional regression techniques or time-series modeling (for example, the autoregressive integrated moving average (ARIMA)). While these previous studies indicated very important variables to predict ridership (e.g., population density, income, and calendar event), this study includes multivariate analyses to provide a larger set of predictors than those included in the other studies. Additionally, by utilizing Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models, as compared with traditional methods, a clear shift toward deep learning has occurred, with better accuracy and the ability to recognize

non-linear relationships in ridership data unidentifiable by traditional models. These advancements contribute to the growing body of literature supporting the use of deep learning in urban transportation forecasting.

This study offers actionable insights for transportation stakeholders by identifying key factors that influence rail ridership. The successful deployment of deep learning models underscores their potential in capturing intricate ridership patterns, highlighting avenues for future research and operational enhancements in public transportation systems. As a whole, this body of work advances the intellectual foundations of transportation planning and management; however, the project also has limitations due to computational and data availability challenges. The need for scalable infrastructure to accommodate large real-time data sets is essential to overcoming the computational challenges posed by the dataset size and complexity. Further, future research should look at utilizing ensemble and hybrid modeling methods, which incorporate various forecasting techniques to improve upon the advances made through the utilization of deep learning and artificial intelligence. The use of advanced predictive analytics, including attention mechanisms and reinforcement learning approaches, will enable improvements in the quality of multivariate time series forecasting used in transport-related studies.

ACKNOWLEDGEMENT

This research was funded by a grant from the Ministry of Higher Education of Malaysia (FRGS Grant) RDU230127 (FRGS/1/2023/TK02/UMP/02/2).

REFERENCES

Alver, Y., & Ercan, H. 2024. Daily travel demand prediction in rail systems by using deep learning techniques. International Conference on Road and Rail Infrastructure.

Asian Development Bank. 2020. Sustainable Urban Transport in Asia. Manila: Asian Development Bank.

Bernama. 2023a. Merdeka Tournament: Services for Bukit Jalil LRT Station extended until 12.30am Saturday. New Straits Times, 12 October 2023.

Bernama. 2023b. Rapid KL to extend operating hours on New Year's Eve. New Straits Times, 29 December 2023.

Chen, E., Ye, Z., Wang, C., & Xu, M. 2020. Subway Passenger Flow Prediction for Special Events Using Smart Card Data. IEEE Transactions on Intelligent Transportation Systems, 21: 1109-1120.

Dikshit S., Atiq A., Shahid M., Dwivedi V. & Thusu A. 2023. The use of artificial intelligence to optimize the routing of vehicles and reduce traffic congestion in urban areas. EAI Endorsed Transactions on Energy Web 10: 1–13.

Dimitriou H.T. & Gakenheimer R. 2011. Urban Transport in the Developing World: A Handbook of Policy and Practice. Cheltenham: Edward Elgar Publishing.

Gers F.A., Schmidhuber J. & Cummins F. 1999. Learning to forget: Continual prediction with LSTM. In Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN 99) 2: 850–855. London: Institution of Engineering and Technology.

Guang, Z., Yang, J., & Li, J. 2017. Forecast of Short-Term Passenger Flow of Urban Railway Stations Based on Seasonal ARIMA Model. In: Jia, L., Qin, Y., Suo, J., Feng, J., Diao, L., An, M. (eds) Proceedings of the 3rd International Conference on Electrical and Information Technologies for Rail Transportation (EITRT) 2017. EITRT 2017. Lecture Notes in Electrical Engineering, vol 483. Springer, Singapore.

Halvorsen A., Wood D., Jefferson D., Stasko T., Hui J. & Reddy A. 2023. Examination of New York City Transit's bus and subway ridership trends during the COVID-19 pandemic. Transportation Research Record 2677(4): 51–64.

Janssen J., Brederode L., Wismans L. & Van Berkum E. 2019. Developing an artificial neural network for estimating road capacity values for weaving sections. CORE.

Malaysia, Department of Statistics (DOSM). 2024. W.P. Kuala Lumpur - The Population of Malaysia. OpenDOSM.

Malaysia National Artificial Intelligence Roadmap 2021–2025 (AI-RMAP). 2021. Ministry of Science, Technology and Innovation (MOSTI), Putrajaya. ISBN 9789671902554.

Macrotrends. 2024. Kuala Lumpur Population 1950–2024. Macrotrends.

Malaymail. 2023a. Madani Government One-Year Anniversary Programme at Bukit Jalil Stadium from December 8 to 10, says secretariat. Malay Mail, 15 November 2023.

Malaymail. 2023b. Malaysia Cup Final: Bukit Jalil LRT Station to extend operations until 1am on December 8, says Rapid Rail. Malay Mail, 6 December 2023.

Montaño Moreno J.J., Palmer Pol A., Sesé Abad A., Cajal Blasco B. 2013. Using the R-MAPE index as a resistant measure of forecast accuracy. Psicothema 25(4):500-6.

MyRapid. 2024. Prasarana's Ridership. MyRapid.

Ni, M., He, Q., & Gao, J. 2017. Forecasting the Subway Passenger Flow Under Event Occurrences with Social Media. IEEE Transactions on Intelligent Transportation Systems 18: 1623-1632.

Rozlan I. 2023. Here are the buses and trains with extended operating hours on Merdeka Day. Lowyat.NET, 30 August 2023.

Tan D. 2023a. Rapid KL extends LRT, MRT, monorail hours on August 31. Paul Tan's Automotive News, 23 August 2023.

Tan D. 2023b. Public transport ridership increase in Malaysia; daily average at 1.1m, highest since pandemic. Paul Tan's Automotive News, 4 December 2023.

Toque, F., Côme, É., Oukhellou, L., & Trépanier, M. 2018. Short-Term Multi-Step Ahead Forecasting of Railway Passenger Flows During Special Events With Machine Learning Methods. CASPT 2018, Conference on Advanced Systems in Public Transport and TransitData 2018.

UITP (International Association of Public Transport). n.d. Global Public Transport Report: Asia-Pacific Transport Trends. UITP.

WildML. 2015. Recurrent neural network tutorial, part 4 – Implementing a GRU/LSTM RNN with Python and Theano. WildML Blog. October.

Willmott C.J., Matsuura K. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim Res 30:79-82.

World Bank. 2017. Public Transport: Lessons from Developing Cities. Washington, DC: World Bank.

Xu, Y., Qiao, Q., & Wu, R. 2019. Urban Rail Transit Hourly Ridership Evolution Model under Rainfall Weather. CICTP 2019.

Zhao, Y., Ma, Z., Jiang, X., & Koutsopoulos, H.N. 2021. Short-Term Metro Ridership Prediction During Unplanned Events. Transportation Research Record 2676: 132 - 147.

Zhou M., Wang D., Li Q., Yue Y., Tu W. & Cao R. 2017. Impacts of weather on public transport ridership: Results from mining data from different sources. Transportation Research Part C: Emerging Technologies 75: 17–29.

Zou, L., Wang, Z., & Zhao, L. 2024. Data pre-integration in graph-transformer networks for enhanced spatiotemporal ridership prediction. Proceedings Volume 13421, Eighth International Conference on Traffic Engineering and Transportation System (ICTETS 2024); 1342147.

APPENDICES

Appendix A. Line chart

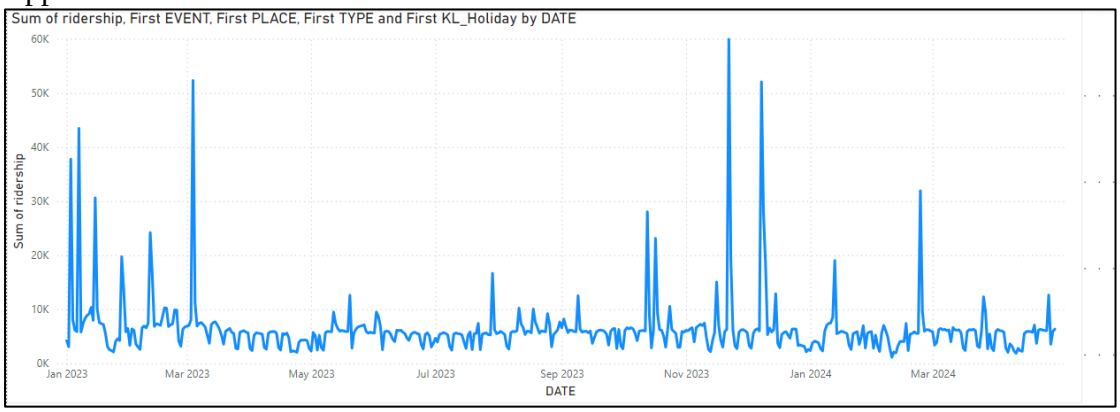

FIGURE 2. Line chart of ridership at Bukit Jalil station

Appendix B. Correlation matrix

	count	isHolliday	existEvent	more than 1 day	China	Taiwan	Hong Kong	Korea	Japan	Malaysia	UK	FGA_rain	FGM_
count	1.000000	-0.171328	0.503363	0.196431	0.000514	0.172629	0.057410	0.239432	0.051053	0.337128	0.303672	-0.003851	-0.02
isHolliday	-0.171328	1.000000	-0.082641	-0.054925	0.066029	-0.034768	-0.044553	-0.027401	-0.012204	-0.066443	-0.032488	0.049407	-0.02
existEvent	0.503363	-0.082641	1.000000	0.786029	0.229675	0.326444	0.418320	0.257274	0.114583	0.623856	0.305043	-0.009633	-0.00
more than 1 day	0.196431	-0.054925	0.786029	1.000000	0.212273	0.358378	0.442407	-0.031628	-0.014086	0.545003	0.084095	-0.004267	0.02
China	0.000514	0.066029	0.229675	0.212273	1.000000	-0.011736	-0.015039	-0.009250	-0.004119	-0.022429	-0.010967	0.039349	0.02
Taiwan	0.172629	-0.034768	0.326444	0.358378	-0.011736	1.000000	-0.021357	-0.013135	-0.005850	-0.031851	-0.015574	0.011611	0.03
Hong Kong	0.057410	-0.044553	0.418320	0.442407	-0.015039	-0.021357	1.000000	-0.016832	-0.007497	-0.040815	-0.019957	0.071606	-0.03
Korea	0.239432	-0.027401	0.257274	-0.031628	-0.009250	-0.013135	-0.016832	1.000000	-0.004610	-0.025102	-0.012274	-0.123423	0.06
Japan	0.051053	-0.012204	0.114583	-0.014086	-0.004119	-0.005850	-0.007497	-0.004610	1.000000	-0.011180	-0.005467	0.019614	-0.01
Malaysia	0.337128	-0.066443	0.623856	0.545003	-0.022429	-0.031851	-0.040815	-0.025102	-0.011180	1.000000	-0.029763	-0.062409	-0.03
UK	0.303672	-0.032488	0.305043	0.084095	-0.010967	-0.015574	-0.019957	-0.012274	-0.005467	-0.029763	1.000000	0.052216	-0.00
FGA_rain	-0.003851	0.049407	-0.009633	-0.004267	0.039349	0.011611	0.071606	-0.123423	0.019614	-0.062409	0.052216	1.000000	-0.05
FGM_rain	-0.029627	-0.027020	-0.006804	0.024098	0.022996	0.032657	-0.036897	0.067602	-0.019614	-0.034275	-0.004941	-0.059374	1.00
FGN_rain	0.083770	0.057091	-0.024844	-0.045796	0.034539	-0.015994	-0.033319	-0.074119	0.040047	-0.012609	0.037151	0.285858	-0.14
FSIGW_rain	-0.037496	0.056286	-0.042870	-0.053583	0.028618	-0.015702	0.052079	-0.110064	0.014265	-0.076154	0.037976	0.727295	0.13
DayOfWeek_Fri	0.048516	-0.065865	0.045883	0.081333	-0.036891	0.040241	0.042436	-0.041286	0.111665	0.026325	0.000507	0.094961	-0.04
DayOfWeek_Mon	-0.073811	0.122612	-0.144742	-0.106584	0.027636	-0.052831	-0.067700	-0.041636	-0.018544	-0.100963	-0.049367	0.000723	-0.03
DayOfWeek_Sat	0.109083	-0.017080	0.265524	0.099653	0.092474	0.065284	0.150129	0.248628	-0.018544	0.049963	0.098130	0.016760	-0.01
DayOfWeek_Sun	-0.095734	0.027746	0.057738	0.076759	0.026945	0.084032	0.076103	-0.041963	-0.018698	-0.001844	-0.000901	-0.044593	0.04
DayOfWeek_Thu	-0.033490	0.078090	-0.074498	-0.043154	-0.036891	-0.052389	-0.067134	-0.041286	-0.018389	0.001036	0.000507	-0.017952	-0.01
DayOfWeek_Tue	0.021906	-0.086926	-0.059270	-0.044713	-0.037202	-0.052831	-0.067700	-0.041636	-0.018544	0.049963	-0.049367	-0.063424	-0.00
DayOfWeek_Wed	0.024352	-0.039021	-0.091695	-0.063902	-0.036891	-0.052389	-0.067134	-0.041286	-0.018389	-0.024252	0.000507	0.014315	0.06
TYPE_Concert	0.330666	-0.050721	0.755571	0.513243	0.304240	0.432049	0.553648	0.340503	0.151651	0.022643	0.340900	0.026875	0.01
TYPE_Sport	0.362604	-0.063888	0.599960	0.570023	-0.021566	-0.030626	-0.039246	-0.024137	-0.010750	0.922303	0.048116	-0.047494	-0.02
PLACE_Axiata Arena, Bukit Jalil	0.134356	-0.064080	0.854343	0.624684	0.269066	0.222455	0.447676	0.234034	0.134118	0.526823	0.129730	-0.039125	0.00
PLACE_National Stadium, Bukit Jalil	0.737875	-0.049584	0.465557	0.105148	-0.016738	0.248090	0.041002	0.095537	-0.008343	0.300937	0.364893	0.048126	-0.01
quarter_Quarter 1	0.082948	0.045440	0.029555	-0.000551	-0.023113	0.000547	-0.127585	0.047770	-0.034947	0.119480	0.013875	-0.061407	-0.07
quarter_Quarter 2	-0.136397	0.053249	-0.018937	0.022401	-0.052200	0.037983	-0.006582	-0.011296	-0.026019	0.001171	-0.029359	-0.037852	-0.00
quarter_Quarter 3	-0.049630	-0.047561	0.021697	0.033229	-0.044111	0.019534	0.211338	0.002442	-0.021988	-0.119713	0.029223	-0.076225	0.10
quarter_Quarter 4	0.097867	-0.067212	-0.037420	-0.057417	0.130502	-0.062226	-0.047203	-0.049041	0.094013	-0.026806	-0.014085	0.194255	-0.00
event_day_1.0	0.365478	-0.065175	0.611950	0.240979	0.176821	0.251103	0.349807	0.420417	0.187242	0.055917	0.347714	0.006399	-0.03
event_day_2.0	0.054524	-0.034768	0.326444	0.415306	0.167242	0.364583	0.279290	-0.013135	-0.005850	-0.031851	0.120142	0.055879	0.07
event_day_3.0	0.002261	0.053639	0.257274	0.327306	0.216439	0.147114	0.362277	-0.010352	-0.004610	-0.025102	-0.012274	-0.011782	0.01
sport day_1.0	0.473036	-0.040897	0.383991	0.196306	-0.013805	-0.019605	-0.025122	-0.015451	-0.006881	0.556151	0.097783	0.027859	-0.02
sport day_2.0	0.067656	-0.021181	0.198872	0.253009	-0.007150	-0.010153	-0.013011	-0.008002	-0.003564	0.318779	-0.009488	-0.109789	-0.03
sport day_3.0	0.029660	-0.021181	0.198872	0.253009	-0.007150	-0.010153	-0.013011	-0.008002	-0.003564	0.318779	-0.009488	-0.037873	0.03
sport day_4.0	0.018876	-0.021181	0.198872	0.253009	-0.007150	-0.010153	-0.013011	-0.008002	-0.003564	0.318779	-0.009488	-0.037873	-0.03
sport day_5.0	0.016636	-0.021181	0.198872	0.253009	-0.007150	-0.010153	-0.013011	-0.008002	-0.003564	0.318779	-0.009488	0.034042	0.03
sport day_6.0	0.010537	-0.021181	0.198872	0.253009	-0.007150	-0.010153	-0.013011	-0.008002	-0.003564	0.318779	-0.009488	-0.037873	-0.03
Day_Category_Weekday	-0.009869	0.008375	-0.250251	-0.136775	-0.092473	-0.131321	-0.175299	-0.159612	0.028886	-0.037126	-0.075185	0.021728	-0.02
Day_Category_Weekend	0.009869	-0.008375	0.250251	0.136775	0.092473	0.131321	0.175299	0.159612	-0.028886	0.037126	0.075185	-0.021728	0.02