Email: editor@ijerst.com or editor.ijerst@gmail.com

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp997-1005 Vol. 21, Issue 2, 2025

PREDICTING HOURLY BOARDING DEMAND OF BUS PASSENGERS USING IMBALANCED RECORDS FROM SMART-CARDS A DEEP LEARNING APPROACH

[1]Pinjari Muthu, MCA Student, Department of MCA
[2]P. Rohini Bai, M.Tech, Assistant, Professor, Department of MCA
[12]Dr KV Subba Reddy Institute of Technology, Dupadu, Kurnool

ABSTRACT

Data from tap-on smart cards is a useful tool for predicting future travel demand and learning about customers' boarding habits. In contrast to negative instances (not boarding at that bus stop at that time), positive instances (i.e., boarding at a particular bus stop at a certain time) are uncommon when looking at the smart-card records (or instances) by boarding stops and by time of day. It has been shown that machine learning algorithms used to forecast hourly boarding numbers from a certain site are far less accurate when the data is unbalanced. Prior to using the smart card data to forecast bus boarding demand, this research resolves the problem of data imbalance. In order to supplement a synthetic training dataset with more evenly distributed travelling and non-traveling cases, we suggest using deep generative adversarial networks (Deep-GAN) to create fake travelling instances. A deep neural network (DNN) is then trained using the synthetic dataset to forecast the travelling and non-traveling instances from a certain stop within a specified time range. The findings demonstrate that resolving the problem of data imbalance may greatly enhance the prediction model's functionality and better match the true profile of passengers. A comparison of the Deep-GAN's performance with that of other conventional resampling techniques demonstrates that the suggested approach may generate a synthetic training dataset with more variety and similarity, and therefore, a better prediction capability. The study emphasises the need of enhancing data quality and model performance for travel behaviour prediction and individual travel behaviour analysis, and it offers helpful advice in this regard.

I. INTRODUCTION

Rapid urbanisation increases the number of people living in urban areas, raises travel demand, and has negative consequences on air pollution and traffic congestion [1]–[3]. It is well acknowledged that public transit is an environmentally friendly and sustainable way to address these transportation issues. Buses have long dominated passenger transit as a traditional public transportation modality [4], [5]. However, poor levels of bus services have been caused by congestion, bus bunching, and unpredictable journey times [6]–[8]. Due in large part to the rise of ride-hailing services in recent years, bus ridership has declined in several cities [9]–[11]. Bus operators must figure out how to boost the bus's performance, appearance, and appeal in order to maintain and grow its customer base. Bus ridership may be increased by using advanced operating and management strategies for bus systems, which can greatly enhance level-of-service and service dependability [12]–[14]. This necessitates comprehending the temporal and geographical fluctuations in passenger demand and implementing the required supply-side adjustments [15]–[18]. Automatic fare collection is the original purpose of the smart card system. Smart-card data has become a ready-made and valuable data source

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

https://doi.org/10.62643/ijerst.2025.v21.i2.pp997-1005

Vol. 21, Issue 2, 2025

for spatiotemporal demand analysis [19], public transport planning [20]–[23], and further analysis of emission reduction for sustainable transport [24], [25] because the system also records boarding information, such as who boards buses, where they board, and when they board. We can simply examine the passenger movement on bus lines and at bus stops using the smart-card data, which allows us to determine the temporal and geographical aspects of bus journeys [26], [27]. However, there is still a lot of work to be done in automatically extracting valuable information from huge data. Large smart-card datasets may now be analysed effectively and efficiently thanks to machine learning methods. For example, Liu et al. [28] used a decision tree model to capture important aspects in the forecast of passenger flow in public transportation. Using a neural network model, Zuo et al. [29] developed a three-stage framework to predict individual accessibility in bus systems.

In our own recent study [30], we show that combining machine learning methods with smartcard data may be a potent way to forecast the temporal and geographical patterns of bus boarding. When averaged over all travellers, the projections were shown to be very accurate. However, when attempting to forecast travel behaviour at the level of individual travellers and precise spatial-temporal features, our study has also shed light on the problems of data imbalance. An individual smart-card holder boarding at a specific stop during a specific time window (for example, an hour) is an uncommon occurrence; the majority of the records would indicate negative (non-traveling, or not boarding at this bus stop during this time window) instances, while only a small number would indicate positive (travelling, boarding at this stop at this time) instances. The effectiveness and precision of machine learning models used to forecast travel behaviour at the level of individual travellers and precise spatial-temporal features might be severely lowered by such data imbalance problems. This drives the current study, in which we address the problem of data imbalance in predicting disaggregate boarding demand (i.e., individual passengers' boarding behaviour during each hour of the day) by proposing an over-sampling technique, deep generative adversarial nets (Deep-GAN) model (originally developed in the context of image generation). We demonstrate that the prediction accuracy greatly increases with the synthesised and more balanced database. The suggested strategy, which is based on the Deep-GAN method, performs better when compared to alternative resampling techniques, such as Random Under-Sampling and Synthetic Minority Oversampling Technique.

This is how the remainder of the paper is structured. The main similar techniques and their uses in transport studies are reviewed in Section II. The particular data imbalance problem in forecasting the hourly boarding demand is explained in Section III. In Section IV, a deep neural network (DNN) is used to forecast the boarding behaviours (boarding or not boarding) of individual smart-card holders at any time of day using a Deep-GAN to provide a synthesised, more balanced training data sample. The suggested approach is applied to a real-world case study in Section V, and Section VI discusses the findings. Section VII concludes by summarising the primary conclusions and contributions of this work and outlining potential directions for further research.

II. LITERATURE SURVEY

"Timetable coordination of first trains in urban railway network: A case study of Beijing," by X. Guo, J. Wu, H. Sun, R. Liu, and Z. Gao Utilising Mathematical Models

This study proposes a model of first train schedule coordination in urban railway networks based on the significance of lines and transfer

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp997-1005
Vol. 21, Issue 2, 2025

stations. The proposed model is solved using a mathematical programming solver and a sub-network connection mechanism is devised. To confirm the efficacy of our proposed model, a genuine Beijing urban railway network and a basic test network are modelled. The results show that by drastically cutting down on connection time, the suggested methodology is successful in enhancing transfer performance.

W. Wu, P. Li, R. Liu, W. Jin, B. Yao, Y. Xie, and C. Ma, "A newsvendor model for forecasting bus route peak load using supply optimisation and scaled shepard interpolation," Part E of Transportation Research: Review of Logistics and Transportation

For on-demand public transit to function, precise short-term demand forecasting is necessary. Knowing where and when future travel needs are anticipated enables operators to swiftly modify schedules, enhancing service quality and dependability and drawing more people to public transportation. By using AI-based deep learning models to forecast bus passenger expectations using real-world patronage data from Melbourne's smart-card ticketing system, this work fills this requirement. Real-world data from 18 bus routes and 1,781 bus stops was used to construct the models, which take into account the temporal features of transit demand for some of Melbourne's busiest routes. Using the same data set, five traditional deep learning models were compared and assessed with LSTM and BiLSTM deep learning models. Additionally, a desktop comparison was conducted against many well-known demand forecasting models that have been documented in the literature throughout the last ten years. The findings of the comparative study showed that BiLSTM models performed better than the other models examined and had an accuracy of over 90% in predicting passenger requests.

"Artificial intelligence in railway transport: Taxonomy, regulations, and applications," by N. Beɿsinovi´c, L. De Donato, F. Flammini, R. M. Goverde, Z. Lin, R. Liu, S. Marrone, R. Nardone, T. Tang, and V. Vittorini

Railway transport is not an exception to the growing prevalence of artificial intelligence (AI) in most technical fields. However, because there are so many new terms and meanings attached to them, there is a chance that railway practitioners, like many other groups, will become lost in these ambiguities and blurred lines and miss the true opportunities and potential of big data analytics, machine learning, and artificial vision, to name a few of the most promising AI-related approaches. This paper's goal is to familiarise railway scholars and practitioners with the fundamental ideas and potential uses of artificial intelligence. In order to help researchers and practitioners better understand AI techniques, research fields, disciplines, and applications—both generally and in relation to railway applications like autonomous driving, maintenance, and traffic management—this paper offers a structured taxonomy. Additionally presented are the significant ethical and explicable issues of AI in railroads. Relevant research addressing current and upcoming applications has promoted the link between AI principles and railway subdomains in order to provide some indications of possible possibilities.

"A review on the co-benefits of mass public transport in mitigating climate change," by S. C. Kwan and J. H. Hashim Sustainable Society and Cities

Due to the desired win-win outcomes of such policies towards both local and global aims, the scale of co-benefits from policies addressing climate change mitigations has been actively pushed. Studies on the quantitative health and environmental co-benefits of different modal transitions to public transport scenarios are

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp997-1005 Vol. 21, Issue 2, 2025

examined in this study. The evaluation of articles from 2004 to August 2015 was done by a systematic review. Nine of the 153 papers that were found met all of the review's requirements. Numerous studies have only examined the advantages for the environment, particularly the decreased air pollution caused by public transit in urban areas.

III. SYSTEM ANALYSIS AND DESIGN

EXISTING SYSTEM

Recent years have seen the emergence of smart card data, which offers a complete and affordable supply of information for public transport system planning and management. Using data from smart cards, this study proposes a multi-stage machine learning architecture to forecast passengers' boarding stops. The framework divides the prediction of hourly ridership into three stages: whether to travel or not during that one-hour time slot, which bus line to use, and at which stop to board. This helps to address the challenges caused by the imbalanced nature of the data (such as many non-traveling data) and the "many-class" issues (such as many possible boarding stops). Two deep learning architectures are implemented: recurrent neural networks (RNN) and long short-term memory networks (LSTM), in addition to a basic neural network architecture called fully connected networks (FCN). A real-world bus network is used to test the suggested methodology.

We demonstrate that, at the individual level, the data imbalance significantly affects prediction accuracy. When taken as a whole, FCN can estimate passengers at individual stops with high accuracy, but it struggles to capture the ridership's temporal dispersion. While LSTM and RNN can measure the temporal distribution, they are unable to use bus lines to capture the spatial distribution.

Disadvantages

• Outliers may affect the data produced by SMOTE and ADASYN. Due to minority outlier cases (often noisy data), they may produce some data in the majority data space, blurring classification borders and making the classification model more challenging to train.

• Because they must exclude a portion of the data, under-sampling techniques often come with the cost of losing some of the information in the rest of the data.

• There has been little research on the loss brought on by the data imbalance problem in the public transport system, despite the fact that the Easy Ensemble and Balance Cascade attempted to address the issue of lost information by expanding the number of models tens of times, which greatly increased the computing load. Additionally, no studies have been conducted to confirm the effectiveness of the current resampling techniques on unbalanced data in the boarding prediction job.

Proposed System

• This work is the first to concentrate on the problem of data imbalance in the public transport system and to provide a deep learning method, Deep-GAN, to address it.

• This research examined the variations in diversity and similarity between the synthetic and real-world travelling instances produced by Deep-GAN and other over-sampling techniques. In order to increase the quality of the data, it also examined various resampling techniques by assessing how well the next travel behaviour prediction model performed. Based on actual data from the public transport system, this is the first validation and assessment of the effectiveness of various data resampling techniques.

• Individual boarding behaviour was creatively simulated in this work, which is unusual in prior travel demand prediction projects. This individual-based model may provide more information about the passengers' conduct than

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp997-1005
Vol. 21, Issue 2, 2025

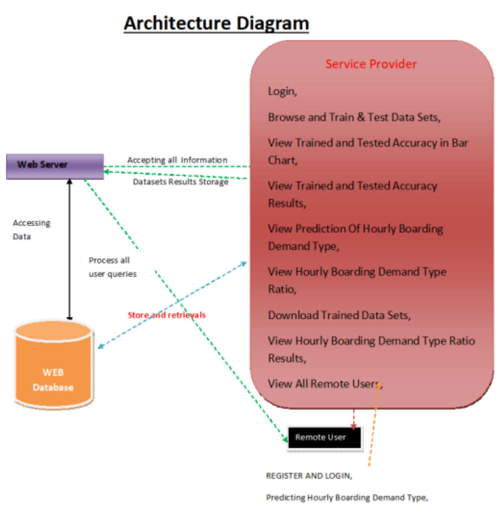
the widely used aggregated forecast, and the findings will help with the examination of the similarities and differences.

Advantages

In order to solve the problem of data imbalance in predicting disaggregate boarding demand—that is, individual passengers' boarding behaviour throughout the day—the system suggests an over-sampling technique called deep generative adversarial nets (Deep-GAN) model, which was first created in the context of image generation.

The system demonstrates that the prediction accuracy greatly increases with the synthesised and more balanced database. The suggested strategy, which is based on the Deep-GAN method, performs better when compared to alternative resampling techniques, such as Random Under-Sampling and Synthetic Minority Oversampling Technique.

SYSTEM ARCHITECTURE



IV. IMPLEMENTATION

Modules

Service Provider

The Service Provider must use a working user name and password to log in to this module. He can do a number of tasks after successfully logging in, including browsing and training and testing data sets. View Results of Trained and Tested Accuracy, View Trained and Tested Accuracy in Bar Chart, See Hourly Boarding Demand Type Prediction, Hourly Boarding Demand Type Ratio, and Get Trained Data Sets here. View the Results of the Hourly Boarding Demand Type Ratio See Every Remote User,

View and Authorize Users

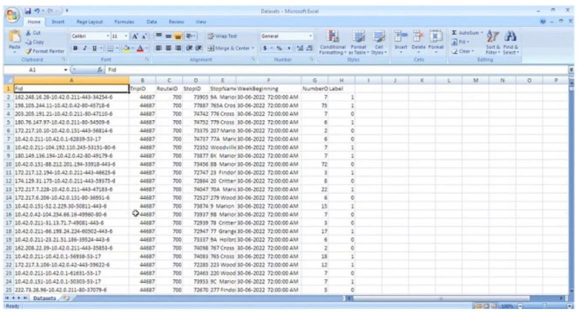
The administrator may see a list of all registered users in this module. Here, the administrator may see the user's information, like name, email, and address, and they can also grant the user permissions.

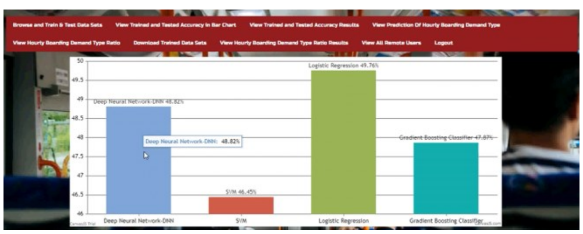
Remote User

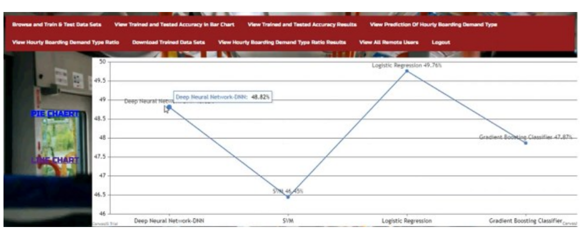
A total of n users are present in this module. Before beginning any actions, the user needs register. Following registration, the user's information will be entered into the database. Following a successful registration, he must use his password and authorised user name to log in. Following a successful login, the user will do tasks such as registering and logging in, predicting the kind of hourly boarding demand, Examine your profile.

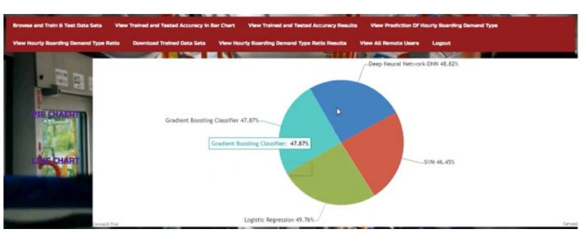
V. RESULTS

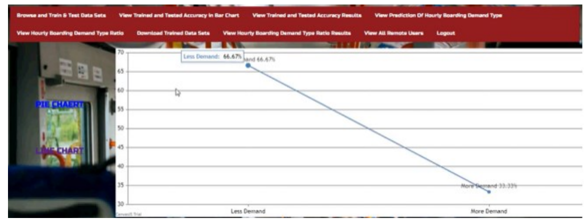
International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp997-1005
Vol. 21, Issue 2, 2025

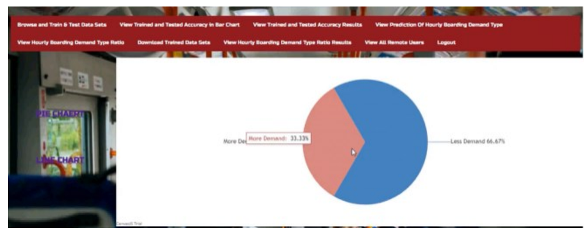


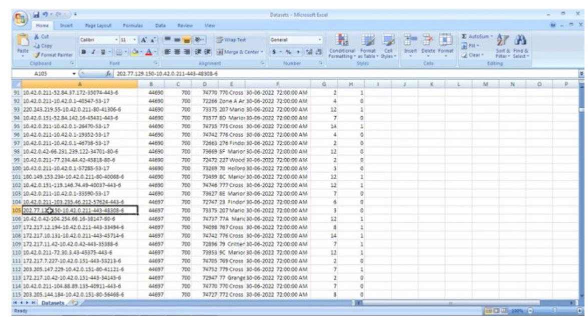




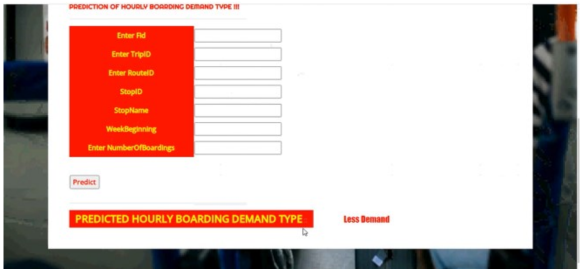


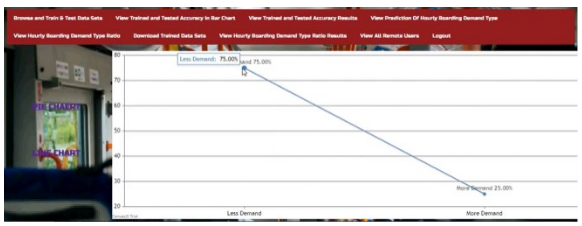


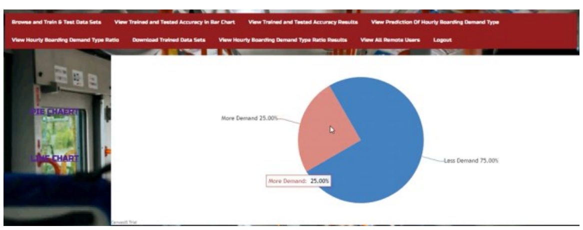


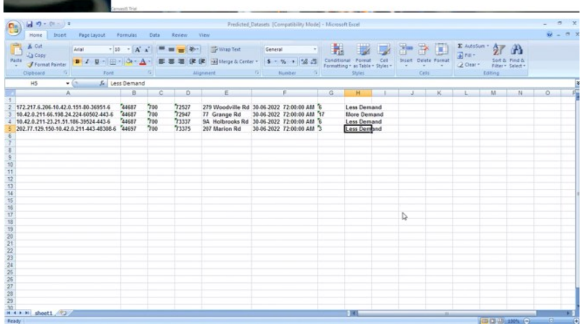


International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp997-1005
Vol. 21, Issue 2, 2025









VI. CONCLUSION

This work was motivated by the difficulty we had in predicting the boarding behaviour of passengers over a time frame using real-world bus smart-card data due to unbalanced data. In order to enhance a DNN-based prediction model of individual boarding behaviour, we suggested a Deep-GAN to over-sample the travelling instances and to rebalance the rate of travelling and non-traveling instances in the smart-card dataset. By using the models on actual smart-card data gathered from seven bus routes in Changsha, China, the effectiveness of Deep-GAN was assessed. By comparing the various imbalance ratios in the training dataset, we discovered that, generally speaking, the model performs better with more unbalanced data, with a 1:5 ratio between positive and negative examples showing the most increase. The high incidence of imbalance will result in deceptive load profiles, while the perfectly balanced data may overpredict the ridership during peak hours, according to the forecast accuracy of the hourly distribution of bus ridership. Both over-sampling and under-sampling improve the model's performance, according to a comparison of many similar approaches. Among the over-sampling techniques, Deep-GAN has the highest recall and accuracy ratings. The Deep-GAN also demonstrated a strong ability to enhance the quality of training datasets and the performance of predictive models, particularly when the under-sampling is inappropriate for the data, even though the performance of the predictive model trained using the Deep-GAN-data is not appreciably better than that of other similar approaches.

This study's contributions include:

- This work is the first to concentrate on the problem of data imbalance in the public transport system and to provide a deep learning method, Deep-GAN, to address it.
- This research examined the variations in diversity and similarity between the synthetic and real-world travelling instances produced by Deep-GAN and other over-sampling techniques. By assessing the efficacy of the upcoming travel behaviour prediction model, it also evaluated several similar approaches for enhancing the quality of the data. Based on actual data from the public transport system, this is the first

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp997-1005
Vol. 21, Issue 2, 2025

validation and assessment of the effectiveness of various data-resembling techniques.

- Individual boarding behaviour was creatively modelled in this work, which is unusual in prior travel demand prediction projects. This individual-based model may provide more information about the passengers' behaviour than the widely used aggregated forecast, and the findings will help with the examination of the similarities and differences.

Forecasting models will advance in sophistication as computer power and technology advance. Individual travel behaviour will eventually replace the bus network and bus lines as the objective in the area of demand prediction for public transportation systems. The digital twin of the public transport system is one example of how this development might significantly improve planning and administration of public transport. Future forecasting efforts in public transport systems are likely to face the problem of unbalanced data as well. In order to solve the problem of data imbalance in travel behaviour prediction, our study suggests a Deep-GAN model. When compared to other similar approaches, the validation using real-world data demonstrated that the Deep-GAN performed better in handling the problem of data imbalance and improved the prediction models. More researchers and managers will benefit greatly from this study's expertise with comparable data imbalance problems, particularly in public transit.

It should be mentioned that there are still some restrictions even with Deep-GAN and DNN models' excellent performance. First, Deep-GAN is only used for oversampling in this study. Nevertheless, a hybrid Deep-GAN variation exists in which negative cases are under-sampled and positive ones are over-sampled. The encouraging outcomes of the Deep-GAN oversampling encourage further study to evaluate the hybrid Deep-GAN's performance. Second, this research makes predictions at the individual level, which leads to a deluge of data and complicates computing. To reduce the size of the dataset, it could be helpful to classify the passengers (for example, by applying clustering techniques). Third, the spatiotemporal aspects of boarding behaviour are not taken into account by the existing Deep-GAN. The quality of the produced dummy travelling instances and the performance of the subsequent prediction models will be further enhanced by tailoring the networks of the discriminator and generator in GAN according to the features of the boarding behaviour. Lastly, the features and data augmentation variations were individually chosen by the suggested Deep-GAN. Therefore, it is probable that the improvements will fall short of ideal. Further gains are probably possible if the characteristics and the ideal imbalance ratio are chosen together, but the computing complexity will increase. In the future, this can be tested. Likewise, it has been believed that the optimal rate of imbalance for Deep-GAN is also the optimal rate for other similar techniques. Future research must test this assumption.

REFERENCES

[1] X. Guo, J. Wu, H. Sun, R. Liu, and Z. Gao, “Timetable coordination of first trains in urban railway network: A case study of beijing,” Applied Mathematical Modelling, vol. 40, no. 17, pp. 8048–8066, 2016.

[2] W. Wu, P. Li, R. Liu, W. Jin, B. Yao, Y. Xie, and C. Ma, “Predicting peak load of bus routes with supply optimization and scaled shepard interpolation: A newsvendor model,” Transportation Research Part E: Logistics and Transportation Review, vol. 142, p. 102041, 2020.

[3] N. Beˇsinovi´c, L. De Donato, F. Flammini, R. M. Goverde, Z. Lin, R. Liu, S. Marrone, R. Nardone, T. Tang, and V. Vittorini, “Artificial intelligence in railway transport: Taxonomy,

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp997-1005 Vol. 21, Issue 2, 2025

regulations and applications," IEEE Transactions on Intelligent Transportation Systems, 2021.

[4] S. C. Kwan and J. H. Hashim, "A review on co-benefits of mass public transportation in climate change mitigation," Sustainable Cities and Society, vol. 22, pp. 11–18, 2016.

[5] Y. Wang, W. Zhang, T. Tang, D. Wang, and Z. Liu, "Bus od matrix reconstruction based on clustering wi-fi probe data," Transportmetrica B: Transport Dynamics, pp. 1–16, 2021, doi: 10.1080/21680566.2021.1956388.

[6] S. J. Berrebi, K. E. Watkins, and J. A. Laval, "A real-time bus dispatching policy to minimize passenger wait on a high frequency route," Transportation Research Part B: Methodological, vol. 81, pp. 377–389, 2015.

[7] A. Fonzone, J.-D. Schm¨ocker, and R. Liu, "A model of bus bunching under reliability-based passenger arrival patterns," Transportation Research Part C: Emerging Technologies, vol. 59, pp. 164–182, 2015.

[8] J. D. Schm¨ocker, W. Sun, A. Fonzone, and R. Liu, "Bus bunching along a corridor served by two lines," Transportation Research Part B: Methodological, vol. 93, pp. 300–317, 2016.

[9] D. Chen, Q. Shao, Z. Liu, W. Yu, and C. L. P. Chen, "Ridesourcing behavior analysis and prediction: A network perspective," IEEE Transactions on Intelligent Transportation Systems, pp. 1–10, 2020.

[10] E. Nelson and N. Sadowsky, "Estimating the impact of ride-hailing app company entry on public transportation use in major us urban areas," The B.E. Journal of Economic Analysis & Policy, vol. 19, no. 1, p.
20180151, 2019.

[11] Z. Chen, K. Liu, J. Wang, and T. Yamamoto, "H-convlstm-based bagging learning approach for ride-hailing demand prediction considering imbalance problems and sparse uncertainty," Transportation Research Part C: Emerging Technologies, vol. 140, p. 103709, 2022.

[12] R. Liu and S. Sinha, "Modelling urban bus service and passenger reliability," 2007. [13] J. A. Sorratini, R. Liu, and S. Sinha, "Assessing bus transport teliability using micro-simulation," Transportation Planning and Technology, vol. 31, no. 3, pp. 303–324, 2008.

[14] Y. Wang, W. Zhang, T. Tang, D. Wang, and Z. Liu, "Bus od matrix reconstruction based on clustering wi-fi probe data," Transportmetrica B: Transport Dynamics, pp. 1–16, 2021.

[15] Y. Hollander and R. Liu, "Estimation of the distribution of travel times by repeated simulation," Transportation Research Part C: Emerging Technologies, vol. 16, no. 2, pp. 212–231, 2008.

[16] W. Wu, R. Liu, and W. Jin, "Modelling bus bunching and holding control with vehicle overtaking and distributed passenger boarding behaviour," Transportation Research Part B: Methodological, vol. 104, pp. 175–197, 2017.

[17] W. Wu, R. Liu, W. Jin, and C. Ma, "Stochastic bus schedule coordination considering demand assignment and rerouting of passengers," Transportation Research Part B: Methodological, vol. 121, pp. 275–303, 2019.

[18] W. Wu, R. Liu, and W. Jin, "Designing robust schedule coordination scheme for transit networks with safety control margins," Transportation Research Part B: Methodological, vol. 93, pp. 495–519, 2016.

[19] S. Zhong and D. J. Sun, A Spatio-temporal Distribution Model for Determining Origin–Destination Demand from Multisource Data. Springer, Singapore, 2022, pp. 33–52.

[20] M. Bordagaray, L. dell'Olio, A. Fonzone, and Ibeas, "Capturing the conditions that introduce systematic variation in bike-sharing travel behavior using data mining techniques," Transportation Research Part C: Emerging Technologies, vol. 71, pp. 231–248, 2016.