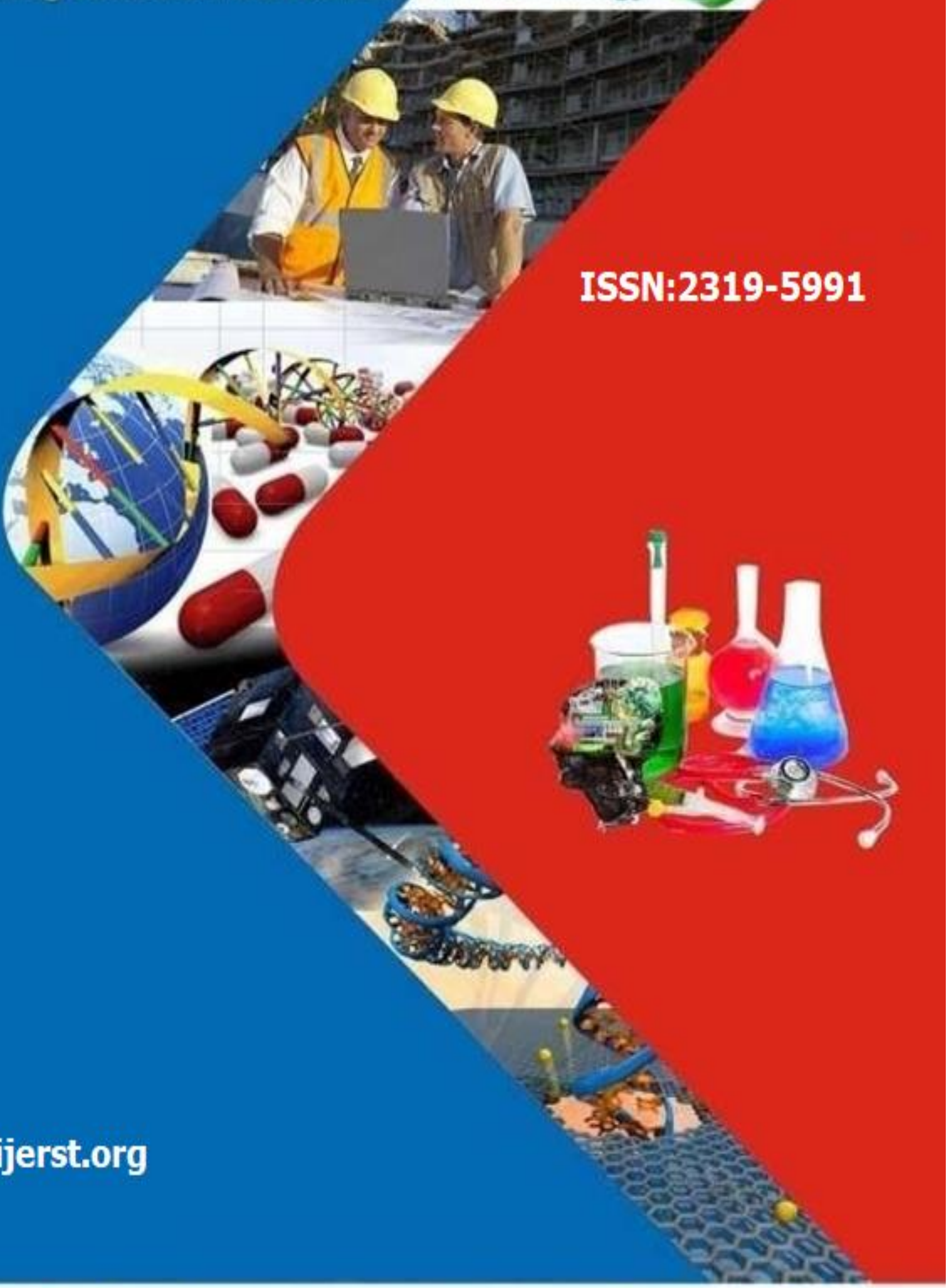
IJERST

International Journal of Engineering Research and Science & Technology

ISSN:2319-5991



www.ijerst.org

E-mail: editor@ijerst.org or ijerst.editor@gmail.com

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

DEEP LEARNING FOR HOURLY BUS BOARDING DEMAND PREDICTION FROM IMBALANCED SMART-CARD DATA

[1]Mrs. Nimmala Bhavani, M.Tech
Assistant professor , Department of MCA
Bapatla Engineering College
Bapatla, Andhra Pradesh

[2]Dasari Jyothirmai
Department of MCA
Bapatla Engineering College
Bapatla,Andhra Pradesh

[3]Murla Sai kalyani
Department of MCA
Bapatla Engineering College
Bapatla,Andhra Pradesh

[4]Yarlagadda Baby vasantha
Department of MCA
Bapatla Engineering College
Bapatla,Andhra Pradesh

[5]Konudula Thirsa
Department of MCA
Bapatla Engineering College
Bapatla,Andhra Pradesh

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

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

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].

The smart-card system is initially designed for automatic fare collection. As the system also records the boarding information, for example, who gets on buses, where and when, smart-card data has become a ready-made and valuable data source for spatio-temporal demand analysis [19], public transport planning [20]–[23], and further analysis of emission reduction for the sustainable transport [24], [25]. From the smart-card data, we can easily observe the passenger flow at bus stops and on bus lines, and from which to derive the spatial and temporal characteristics of bus trips [26], [27]. However, extracting useful information from big data automatically still poses a significant challenge. In recent years, machine learning techniques have emerged as an efficient and effective approach to analyzing large smart-card datasets. For instance, Liu et al. [28] captured key features in public transport passenger flow prediction via a decision tree model. Zuo et al. [29] built a three-stage framework with a neural network model to forecast the individual accessibility in bus systems.

In our own recent research [30], we demonstrate that smartcard data combined with machine learning techniques can be a powerful approach for predicting the spatial and temporal patterns of bus boarding. The predictions were found to be highly accurate at an aggregated level, averaged over all travelers. However, our research has also thrown light on the data imbalance issues, when trying to predict travel behavior at the level of individual travelers and fine spatial-temporal details. For instance, the boarding of an individual smart-card holder at a specific stop during a particular time window (e.g. an hour) is a rare event: most of the records would denote negative (non-travelling, or not boarding at this bus stop during this time window) instances, and only a few are positive (travelling, boarding at this stop at this time) instances. Such data imbalance issues can significantly reduce the efficiency and accuracy of machine learning models deployed for predicting travel behavior at the level of individual travelers and fine spatial-temporal details. This motivates this current study where we propose an over-sampling method, deep generative adversarial nets (Deep-GAN) model (initially developed in the context of

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

image generation) to address the data imbalance issue in predicting disaggregate boarding demand (i.e. individual passengers boarding behavior during each hour of the day). We show that, with the synthesized and more balanced database, the prediction accuracy improves significantly. The performance of the proposed approach, based on the Deep- GAN method, is further benchmarked against other resampling methods (including Synthetic Minority Oversampling Technique and Random Under-Sampling) and is shown to have superior performance.

The rest of the paper is organized as follows. Section II reviews the key resembling methods and their applications in transport studies. Section III describes the specific data imbalance issue in predicting the hourly boarding demand. Section IV uses a Deep-GAN to provide a synthesized, more balanced training data sample and a deep neural network (DNN) to predict the individual smart-card holders' boarding actions (boarding or not boarding) in any hour of a day. Section V applies the proposed method to a real-world case study, and the results are discussed in Section VI. Finally, Section VII summarizes the main findings and contributions of this paper and suggests future investigations.

II. LITERATURE SURVEY

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

A model of timetable coordination of first trains in urban railway networks, based on the importance of lines and transfer stations, is proposed in this paper. A sub-network connection method is developed, and a mathematical programming solver is utilized to solve the suggested model. A simple test network and a real network of Beijing urban railway network are modeled to verify the effectiveness of our suggested model. Results demonstrate that the proposed model is effective in improving the transfer performance in that they reduce the connection time significantly.

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

Accurate short-term forecasting of public transport demand is essential for the operation of on-demand public transport. Knowing where and when future demands for travel are expected allows operators to adjust timetables quickly, which helps improve service quality and reliability and attract more passengers to public transport. This study addresses this need by developing AI-based deep learning models for prediction of bus passenger demands based on actual patronage data obtained from the smart-card ticketing system in Melbourne. The models, which consider the temporal characteristics of travel demand for some of the heaviest bus routes in Melbourne, were developed using real-world data from 18 bus routes and 1,781 bus stops. LSTM and BiLSTM deep learning models were evaluated and compared with five conventional deep learning models using the same data set. A desktop comparison was

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

also undertaken against a number of established demand forecasting models that have been reported in the literature over the past decade. The comparative evaluation results showed that BiLSTM models outperformed other models tested and was able to predict passenger demands with over 90% accuracy.

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, regulations and applications,

Artificial Intelligence (AI) is becoming pervasive in most engineering domains, and railway transport is no exception. However, due to the plethora of different new terms and meanings associated with them, there is a risk that railway practitioners, as several other categories, will get lost in those ambiguities and fuzzy boundaries, and hence fail to catch the real opportunities and potential of machine learning, artificial vision, and big data analytics, just to name a few of the most promising approaches connected to AI. The scope of this paper is to introduce the basic concepts and possible applications of AI to railway academics and practitioners. To that aim, this paper presents a structured taxonomy to guide researchers and practitioners to understand AI techniques, research fields, disciplines, and applications, both in general terms and in close connection with railway applications such as autonomous driving, maintenance, and traffic management. The important aspects of ethics and explainability of AI in railways are also introduced. The connection between AI concepts and railway subdomains has been supported by relevant research addressing existing and planned applications in order to provide some pointers to promising directions.

III. EXISTING SYSTEM

Smart card data has emerged in recent years and provide a comprehensive, and cheap source of information for planning and managing public transport systems. This paper presents a multi-stage machine learning framework to predict passengers' boarding stops using smart card data.

The framework addresses the challenges arising from the imbalanced nature of the data (e.g. many non-travelling data) and the 'many-class' issues (e.g. many possible boarding stops) by decomposing the prediction of hourly ridership into three stages: whether to travel or not in that one-hour time slot, which bus line to use, and at which stop to board. A simple neural network architecture, fully connected networks (FCN), and two deep learning architectures, recurrent neural networks (RNN) and long short-term memory networks (LSTM) are implemented. The proposed approach is applied to a real-life bus network.

We show that the data imbalance has a profound impact on the accuracy of prediction at individual level. At aggregated level, FCN is able to accurately predict the rideship at individual stops, it is poor at capturing the temporal distribution of ridership. RNN and LSTM are able to measure the temporal distribution but lack

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

the ability to capture the spatial distribution through bus lines.

Disadvantages

• The data generated by SMOTE and ADASYN are susceptible to outliers. They may generate some data in the majority data space due to minority outlier instances (usually noisy data), causing blurred classification borderlines and making the learning difficulties of the classification model.

• The under-sampling methods usually have to pay the price of losing parts of the information of the majority of data because they have to remove a part of the data. Although the Easy Ensemble and Balance Cascade tried to solve the problem of lost information, they increased the number of models tens of times, significantly increasing the computational burden.

• Little study has noticed the loss caused by the data imbalance issue in the public transport system. There is also no research to validate the efficiency of the existing resampling methods on imbalanced data in the boarding prediction task.

PROPOSED SYSTEM

• The data imbalance issue in the public transport system has received little attention, and this study is the first to focus on this issue and propose a deep learning approach, Deep-GAN, to solve it.

• This study compared the differences in similarity and diversity between the real and synthetic travelling instanced generated from Deep-GAN and other over-sampling methods. It also compared different resampling methods for the improvement of data quality by evaluating the performance of the next travel behaviour prediction model. This is the first validation and evaluation of the performance of different data resampling methods based on real data in the public transport system.

• This paper innovatively modelled individual boarding behaviour, which is uncommon in other travel demand prediction tasks. Compared to the popular aggregated prediction, this individual-based model is able to provide more details on the passengers' behaviour, and the results will benefit the analysis of the similarities and heterogeneities.

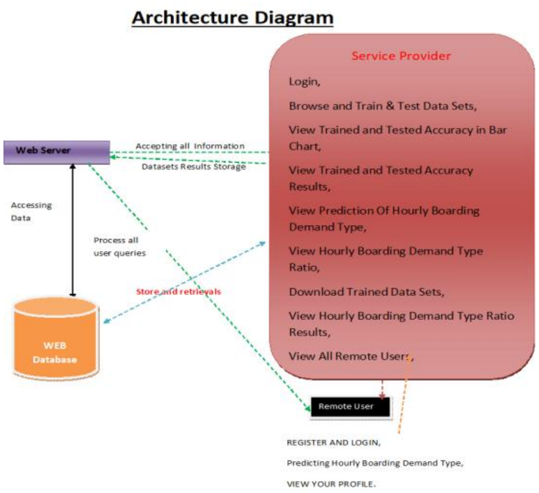
Advantages

The system proposes an over-sampling method, deep generative adversarial nets (Deep-GAN) model (initially developed in the context of image generation) to address the data imbalance issue in predicting disaggregate boarding demand (i.e. individual passengers boarding behavior during each hour of the day).

The system shows that, with the synthesized and more balanced database, the prediction accuracy improves significantly. The performance of the proposed approach, based on the Deep- GAN method, is further benchmarked against other resampling methods (including Synthetic Minority Oversampling Technique and Random Under-Sampling) and is shown to have superior performance.

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

SYSTEM ARCHITECTURE



IV. IMPLEMENTATION

Modules

Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Browse and Train & Test Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Prediction Of Hourly Boarding Demand Type, View Hourly Boarding Demand Type Ratio, Download Trained Data Sets, View Hourly Boarding Demand Type Ratio Results, View All Remote Users.

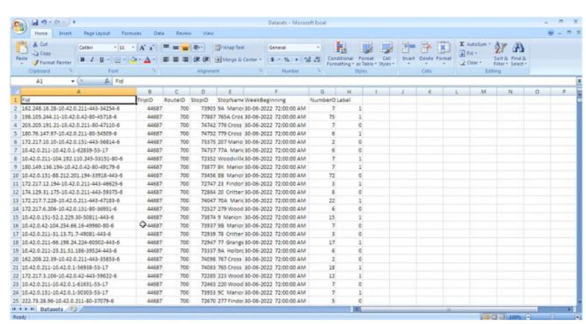
View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

Remote User

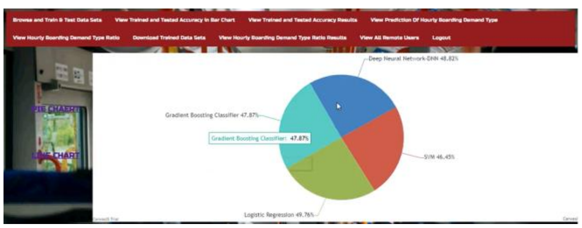
In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, Predicting Hourly Boarding Demand Type, VIEW YOUR PROFILE.

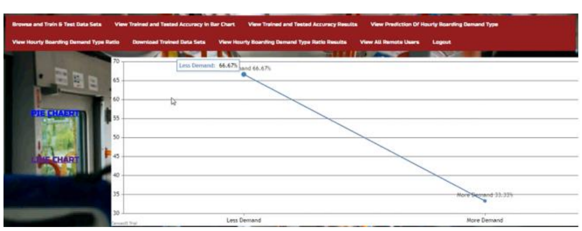
V. RESULTS

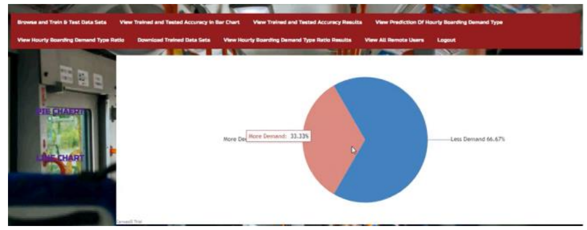


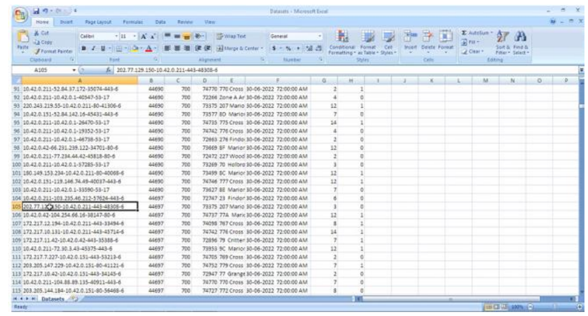
International Journal of Engineering Research and Science & Technology

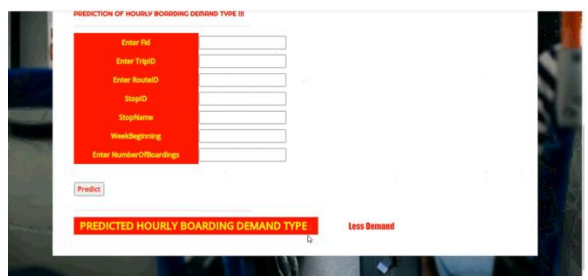
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025



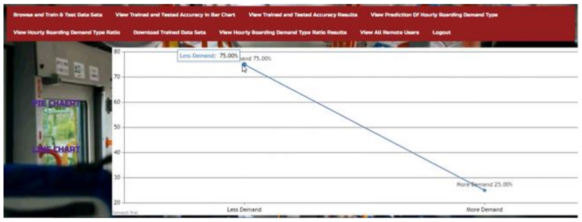


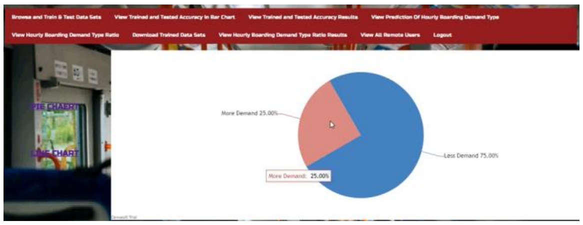


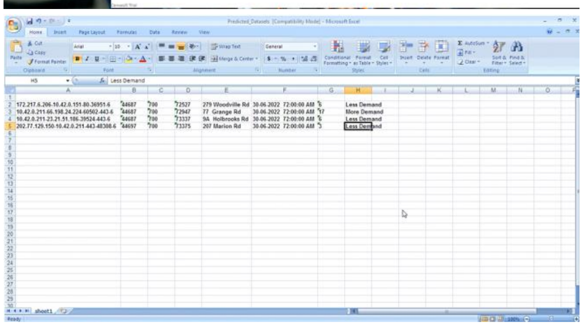




ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025







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

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

• 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. Further study is required to test this notion.

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

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

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, 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.

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

[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.