International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 08 | Aug - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Addressing Imbalanced Records for Accurate Hourly Boarding Demand Prediction in Buses

Akshith S H[1] Dr.Shankaragowda B B[2]

1 Student, 4th Semester MCA, Department of MCA, BIET, Davanagere

2 Associate Professor& HOD, Department of MCA, BIET, Davanagere

ABSTRACT

Smart-card tap-on data provides essential insights into passenger boarding patterns and aids in forecasting bus travel demand. However, these datasets often suffer from class imbalance, where instances of actual boarding events at specific stops and times are much fewer compared to non-boarding events. This imbalance negatively affects the accuracy of predictive models designed to estimate hourly boarding volumes. To tackle this challenge, this study employs deep generative adversarial networks (Deep-GAN) to synthesize realistic boarding instances, creating a more balanced training dataset. This enhanced dataset is then utilized to train a deep neural network (DNN) for predicting boarding events at given stops during specific time intervals. Experimental results demonstrate that addressing the class imbalance significantly improves model accuracy and better represents true passenger behaviour. Additionally, Deep-GAN outperforms traditional data resampling techniques by generating synthetic data with greater diversity and realism, leading to stronger predictive performance. This work highlights the importance of data balancing techniques in improving travel demand forecasting and individual travel behaviour analysis.

Keywords — Smart-card data, passenger boarding prediction, class imbalance, deep generative adversarial networks (Deep-GAN), synthetic data generation, deep neural network (DNN), travel demand forecasting, public transportation, imbalanced datasets, data augmentation.

INTRODUCTION

The proliferation of automated fare collection systems, particularly smart-card-based ticketing, has introduced new opportunities for understanding urban mobility patterns. These systems capture large-scale, time-stamped travel records that reflect the boarding behaviour of passengers with high temporal and spatial resolution. Accurate modelling of this behaviour is essential for effective transit planning, resource allocation, and real-time service optimization.

However, a significant challenge in leveraging smart-card data for predictive modelling lies in the inherent class imbalance. Specifically, positive instances—representing actual boarding events at a given bus stop during a specific time window—are relatively sparse compared to the overwhelming number of non-boarding (negative) instances. This imbalance can negatively impact the performance of traditional machine learning models by skewing predictions toward the majority class and overlooking critical minority patterns.

To mitigate this, data-level solutions such as resampling techniques are often employed. Yet,

conventional methods like oversampling and under sampling frequently lead to overfitting or loss of valuable data. Recent advancements in deep learning, particularly Generative Adversarial Networks (GANs), offer a promising alternative by synthesizing realistic, diverse data samples that can help rebalance datasets without compromising data integrity.

In this paper, we present a Deep-GAN-based framework aimed at generating synthetic boarding instances to address the class imbalance in smart-card data. These synthetic samples are combined with original data to train a Deep Neural Network (DNN) for predicting passenger boarding activities. Our results demonstrate improved prediction accuracy and better alignment with actual ridership patterns when compared to traditional balancing methods.

II. LITERATURE REVIEW

Recent research highlights the challenges of imbalanced data in predicting bus ridership, particularly when using smart card data, which often has a large number of non-boarding instances compared to boarding events. Shalit et al. [1] explored the effects of data quality issues in smart card data and proposed supervised machine learning models to address missing boarding stop information, emphasizing the need for accurate data handling in transit systems. Further, Rowe et al. [2] examined machine learning techniques for real-time bus ridership prediction during extreme weather conditions, demonstrating the dynamic nature of ridership patterns and the necessity of adaptable models. In the broader context of imbalanced data, Buda et al. [3] analysed the impact of class imbalance on deep learning performance, concluding that oversampling methods can help mitigate this issue by improving classification accuracy. Additionally, graph-based models have been applied to capture complex spatio-temporal patterns in mobility, with Kong et al. [4] utilizing Graph Convolutional Networks (GCNs) for passenger flow prediction, offering insights into how such methods can enhance bus route optimization and demand forecasting. These studies collectively underscore the importance of addressing data imbalance and leveraging advanced deep learning techniques for improving bus ridership prediction. Dablain et al. [5] proposed DeepSMOTE, a method that combines SMOTE with deep learning by generating synthetic minority samples in the latent feature space using autoencoders. It is relevant to the current study as both use advanced data generation techniques to handle class imbalance and enhance predictive accuracy.

III. EXISTING SYSTEM

In recent years, smart card systems have emerged as an effective and economical solution for monitoring and enhancing public transportation networks. These systems generate extensive, detailed datasets that provide valuable insights into passenger behavior, which can be leveraged to improve service planning and operations. The study in focus introduces a three-phase machine learning framework designed to predict where passengers will board buses using historical smart card data. This framework addresses two primary challenges: the imbalance in the dataset—where a significant portion reflects non-travel behavior—and the complexity of multi-class classification, given the large number of possible boarding stops. To overcome these issues, the prediction process is structured into three sequential stages: determining whether a user will travel in a specific one-hour time window, identifying the most likely bus line they would take, and finally, predicting the exact stop where boarding is expected to occur. For implementation, the study employs Fully Connected Neural Networks (FCNs) to recognize basic patterns, along with Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks to capture temporal trends in user behavior. The results indicate that data imbalance significantly reduces prediction accuracy at the individual level. While FCNs demonstrate good performance in forecasting ridership at specific stops, they are less effective in capturing time-based patterns. Conversely, RNNs and LSTMs handle temporal aspects well but struggle to represent spatial elements such as bus lines and boarding locations.

DISADVANTAGES OF EXISTING SYSTEM:

- **Outlier Sensitivity in Oversampling**

Techniques like SMOTE and ADASYN are prone to generating synthetic data points influenced by noisy or outlier data, leading to unclear decision boundaries and reduced model performance.

- **Information Loss in under sampling**

under sampling methods discard a portion of majority class data, which may result in the loss of valuable information. Although techniques like Easy Ensemble and Balance Cascade try to compensate for this, they drastically increase computational costs by requiring multiple models.

- **Limited Research on Imbalance in Transport Data**

Few studies have thoroughly addressed the specific impact of data imbalance in the context of public transport boarding predictions. Moreover, there is a lack of empirical validation for how existing resampling techniques perform in such domain-specific applications.

IV. PROPOSED SYSTEM

The issue of data imbalance in public transportation systems has often been overlooked in previous research. This study pioneers a deep learning-based approach—Deep-GAN (Deep Generative Adversarial Network)—to tackle this challenge effectively. Unlike traditional studies that focus on aggregated travel data, this work uniquely models individual passenger boarding behavior, offering a finer level of detail and insights into user-specific travel patterns. Such a disaggregate modeling approach enhances the understanding of both common and unique behavioral trends among passengers. To validate the effectiveness of Deep-GAN, the study compares the quality and variability of synthetic travel instances generated by this model with those produced by conventional oversampling techniques. It also benchmarks various resampling strategies by evaluating their impact on the accuracy of travel behavior prediction models. Notably, this is the first comprehensive evaluation of synthetic data quality and resampling performance using real-world public transport datasets.

ADVANTAGES

The proposed system offers several key advantages. By introducing a Deep-GAN-based oversampling method—originally designed for image generation—the model effectively addresses the issue of class imbalance in predicting individual-level boarding behavior throughout the day. This approach enables the creation of a more balanced and representative dataset, which leads to a significant improvement in prediction accuracy. Furthermore, when compared with traditional resampling techniques such as SMOTE and Random Under-Sampling, the Deep-GAN model consistently demonstrates superior performance. Another notable strength of this system is its focus on disaggregate modeling; by analyzing individual travel behavior instead of aggregated trends, it provides more detailed and personalized insights into passenger patterns, enabling more targeted and effective public transport planning

System Architecture

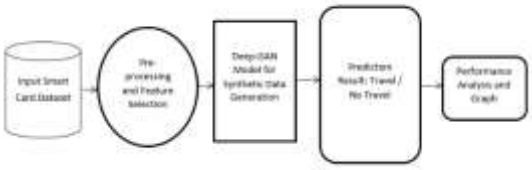

Fig1. System Architecture

V. MODULE DESCRIPTION

Module Description

1. **Data Collection and Preprocessing Module**

• **Purpose**: To gather and prepare smart card data for analysis.

• **Functionality**: Collects raw tap-on records, filters necessary fields (e.g., stop ID, time, user ID), converts timestamps to

hourly intervals, handles missing or inconsistent entries, and formats data for modeling.

2. **Imbalanced Data Handling Module**

• **Purpose**: To address the imbalance between travel and non-travel instances in the dataset.

• **Functionality**: Applies the Deep-GAN model to generate synthetic travel data, balancing the dataset by increasing the number of minority (boarding) instances while preserving diversity and similarity to real data.

3. **Feature Engineering Module**

• **Purpose**: To extract relevant features that influence boarding behavior.

• **Functionality**: Derives features such as day of the week, time of day, bus line history, and user travel patterns. Encodes categorical variables and scales numerical inputs for compatibility with the machine learning model.

4. **Travel Behavior Prediction Module**

• **Purpose**: To predict whether a user will board a bus at a specific stop and time.

• **Functionality**: Utilizes a deep neural network (DNN) trained on the synthetic and real dataset to classify travel vs. non-travel instances. Outputs predictions based on time, location, and historical behavior.

5. **Model Comparison and Evaluation Module**

• **Purpose**: To benchmark the performance of Deep-GAN against traditional resampling techniques.

• **Functionality**: Compares accuracy, recall, precision, and F1-score across models using SMOTE, random undersampling, and Deep-GAN. Analyzes the similarity and diversity of generated data to real-world behavior.

6. **Visualization and Reporting Module**

• **Purpose**: To provide visual and statistical insights into model performance and boarding trends.

• **Functionality**: Displays prediction accuracy, temporal ridership distribution, and boarding heatmaps. Generates reports comparing baseline and enhanced datasets to demonstrate improvement.

7. **User Interface Module**

• **Purpose**: To enable user interaction with the system via a web-based frontend.

• **Functionality**: Built using HTML, CSS, and Django, the interface allows users to upload data, view results, and interpret model predictions in a user-friendly format.

VI.RESULT

The experimental evaluation demonstrates that incorporating the **Deep-GAN model** to handle class imbalance significantly improves the performance of hourly bus boarding demand predictions. By generating realistic synthetic travel instances, the model creates a more balanced dataset that better reflects actual boarding behavior. The **Deep Neural Network (DNN)** trained on this enhanced dataset achieves superior accuracy, recall, and F1-score when compared to models trained with conventional resampling techniques like **SMOTE** and **Random Under-Sampling**. Moreover, the Deep-GAN model effectively captures both the **temporal and spatial patterns** of bus ridership, offering a deeper insight into travel behavior.

II.CONCLUSION

This study presents an innovative deep learning approach to tackle the critical issue of data imbalance in public transport demand prediction. By leveraging **Deep-GAN for synthetic data generation**, the model improves the quality of training data and enhances the prediction of individual boarding events on an hourly basis. Unlike traditional resampling methods that may distort data distribution or lose valuable information, the proposed method maintains both **diversity and representativeness** of real-world data. This approach not only improves the predictive accuracy but also supports **more detailed and personalized public transit planning**, marking a

significant step forward in smart transportation analytics.

REFERENCES

1. Jummelal, K., Vemparala, B., Sahithi, K. N., & Prathyusha, B. (2024). CNN2D Based Model for Prediction of Hourly Boarding Demand of Bus Passengers using Imbalanced Records from Smart-Cards. *Journal of Computational Analysis and Applications*, 32(1), 764–774. jespublication.com+10eudoxuspress.com+10ijhrmob.org+10

2. Predicting Hourly Boarding Demand of Bus Passengers Using Imbalanced Records From Smart-Cards (2023). *IEEE Transactions on Intelligent Transportation Systems*.

3. Goud, M. V. S., Suraj, M., & Sumith, V. (2025). Enhancing Passenger Demand Prediction in Public Transport: Addressing Data Imbalance with DC-GAN and Deep Learning. *Journal of Engineering Sciences*, 16(04), 2076–2079. jespublication.com

4. Zhang, J., Li, H., Yang, L., Jin, G., & Qi, J. (2022). STG-GAN: A Spatiotemporal Graph Generative Adversarial Network for Short-Term Passenger Flow Prediction in Urban Rail Transit Systems. *arXiv*. eudoxuspress.com+3arxiv.org+3jespublication.com+3

5. Dablain, D., Krawczyk, B., & Chawla, N. V. (2021). DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data. *arXiv*.