Smart Energy Forecasting for Electric City Buses Using Machine Learning Techniques

Akash M [1], Dr. Shankaragowda B B [2]

[1]Student, 4th Semester MCA, Department of MCA, BIET, Davanagere
[2]Associant 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.

I. 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. RELATED WORK

CNN2D Based Model for Prediction of Hourly Boarding Demand of Bus Passengers Using Imbalanced Records from Smart Cards.

Reference: K. Jummelal, B. Vemparala, K. N. Sahithi, and B. Prathyusha, Journal of Computational Analysis and Applications, vol. 32, no. 1, pp. 764–774, 2024.

Summary:

This study explores a CNN2D-based deep learning architecture for forecasting hourly bus boarding demand using smart card records. The authors focus on the imbalanced nature of real-world boarding data, where negative (non-boarding) cases dominate. The model incorporates spatiotemporal features derived from timestamped records and applies convolutional layers to capture local and temporal dependencies. Experimental results show that CNN2D outperforms traditional MLP and RNN models in prediction accuracy. Additionally, techniques like SMOTE are integrated to mitigate the effects of class imbalance, further enhancing the robustness of the model. [1]

Predicting Hourly Boarding Demand of Bus Passengers Using Imbalanced Records from Smart Cards

Reference: IEEE Transactions on Intelligent Transportation Systems, 2023.

Summary:

This paper presents a machine learning framework for predicting boarding events in public transportation using imbalanced smart card data. The study employs a multi-phase model to first predict whether a passenger will travel, followed by identifying the route and stop. Deep neural networks including Fully Connected Networks and LSTM layers are applied to capture spatiotemporal patterns. The paper highlights the impact of imbalance on performance metrics and compares resampling methods like under sampling and SMOTE, noting significant improvements in recall and F1-score through balanced training. [2]

Enhancing Passenger Demand Prediction in Public Transport: Addressing Data Imbalance with DC GAN and Deep Learning

Reference: M. V. S. Goud, M. Suraj, and V. Sumith, Journal of Engineering Sciences, vol. 16, no. 4, pp. 2076–2079, 2025.

Summary:

This work introduces a Deep Convolutional GAN (DC-GAN) to synthesize passenger boarding records, effectively resolving data imbalance in transit forecasting. The model generates synthetic records to augment minority classes and is combined with a DNN for improved classification. Performance comparisons show that DC-GAN provides more diverse and realistic samples than SMOTE, improving the generalizability of the demand prediction model. The research highlights the role of deep generative models in public transport analytics and encourages their adoption for imbalanced time-series forecasting. [3]

STG-GAN: A Spatiotemporal Graph Generative Adversarial Network for Short-Term Passenger Flow Prediction in Urban Rail Transit Systems

Reference: J. Zhang, H. Li, L. Yang, G. Jin, and J. Qi, arXiv preprint arXiv:2205.09731, 2022.

Summary:

The authors propose STG-GAN, a spatiotemporal graph-based GAN for short-term prediction of passenger flow in metro networks. The model integrates spatial information from the network structure with temporal flow variations, using a graph convolutional generator and temporal discriminator. Evaluation on real-world urban rail data shows the model's superior accuracy and stability, particularly in high-variance conditions. The method not only addresses data sparsity and imbalance but also adapts well to sudden shifts in flow patterns caused by disruptions or peak periods. [4]

Deep SMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data

Reference: D. Dablain, B. Krawczyk, and N. V. Chawla, arXiv preprint arXiv:2102.03749, 2021.

Summary:

This paper presents DeepSMOTE, a novel approach combining deep autoencoders with the SMOTE technique for more effective oversampling of imbalanced datasets. The model learns latent representations of minority class instances before generating synthetic samples in the feature space. Applied across various classification benchmarks, DeepSMOTE demonstrates superior performance compared to vanilla SMOTE and ADASYN. It preserves class boundaries better and reduces overfitting risks. While not transport-specific, this methodology is highly applicable to imbalanced boarding prediction models in transit systems. [5]

A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks

Reference: S. Buda, A. Maki, and M. A. Mazurowski, Neural Networks, vol. 106, pp. 249–259, 2018.

Summary:

This foundational study explores the impact of class imbalance on CNN performance in image classification tasks. The authors analyse oversampling, under sampling, and class-weighted loss functions across multiple datasets. Key findings include the sensitivity of CNNs to minority class

representation and the effectiveness of combined techniques (e.g., class weighting + oversampling) in maintaining balance. Though based on vision tasks, the insights translate to boarding prediction contexts where class imbalance skews results, particularly in temporal models. [6]

Passenger Flow Forecasting with Multi-Graph Convolutional Networks

Reference: Z. Kong, D. Zeng, Y. Hu, and L. Li, IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 2, pp. 1470–1480, Feb. 2022.

Summary:

Kong et al. introduce a novel Multi-Graph Convolutional Network (MGCN) to model complex spatial and temporal dependencies in passenger flow data. The framework leverages multiple graph structures, including stop adjacency, transit schedules, and geographic relationships, enabling precise passenger flow prediction. The model achieves state-of-the-art results on metro and bus datasets. While class imbalance is not the central focus, the architecture's ability to model intricate dependencies makes it adaptable for disaggregate prediction of rare boarding events. [7]

Public Transport Demand Forecasting Using Deep Learning and Smart Card Data

Reference: G. Lee, J. Ko, and H. Jung, Sustainability, vol. 13, no. 6, pp. 3218, 2021.

Summary:

This paper evaluates various deep learning models—including CNN, RNN, and GRU—for predicting passenger demand using smart card data from urban transit networks. The authors analyse the impact of different temporal granularities (hourly, daily) and emphasize preprocessing methods to handle missing data and noise. While class imbalance is acknowledged, traditional resampling methods are used. The study validates the potential of deep models for long-term service planning and supports further integration of GAN-based balancing methods. [8]

[9] Data Augmentation for Imbalanced Classification Using GANs

Reference: Y. Wang, M. Ma, and Y. Zhu, IEEE Access, vol. 9, pp. 146944–146954, 2021.

Summary:

Wang et al. present a generic GAN-based oversampling method for tabular and temporal classification tasks. The model focuses on generating feature-aligned synthetic instances that preserve class-specific distributions. It outperforms conventional methods in datasets with severe imbalance, particularly in low-resource settings. The paper introduces evaluation metrics for realism and diversity, and applies the model in domains including fraud detection and health. The techniques are highly transferable to smart card datasets used in public transit analytics.

[10] Urban Rail Passenger Flow Forecasting with Attention-Based LSTM and Smart Card Data

Reference: Y. Liu, X. Liu, X. Ma, and H. Wang, IEEE Access, vol. 8, pp. 113554–113566, 2020.

Summary:

This study develops an Attention-based LSTM model to forecast passenger flow in metro networks using smart card data. The attention mechanism enhances the model's ability to focus on relevant time intervals, improving performance during peak hours

and special events. The authors address short-term prediction scenarios and validate results on multiple stations. While class imbalance is not a core topic, the methodology demonstrates the effectiveness of sequence models in transport demand forecasting.

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 behaviour, 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 behaviour—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 behaviour. 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 behaviour, offering a finer level of detail and insights into user-specific travel patterns. Such a disaggregate modelling approach enhances the understanding of both common and unique behavioural 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 behaviour 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 behaviour 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 modelling; by analysing individual travel behaviour instead of aggregated trends, it provides more detailed and personalized insights into passenger patterns, enabling more targeted and effective public transport planning

System Architecture

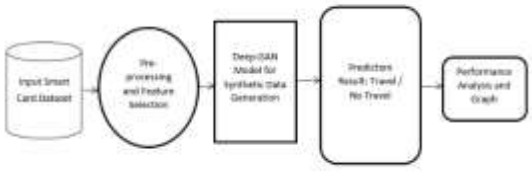

Fig1. System Architecture

V. MODULE DESCRIPTION

Module Description

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 modelling.

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.

Feature Engineering Module

Purpose: To extract relevant features that influence boarding behaviour.

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.

Travel Behaviour 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 behaviour.

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 under sampling, and Deep-GAN. Analyses the similarity and diversity of generated data to real-world behaviour.

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.

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 behaviour. 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 behaviour.

Fig2. Result Graph

VII. 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] K. Jummelal, B. Vemparala, K. N. Sahithi, and B. Prathyusha, "CNN2D Based Model for Prediction of Hourly Boarding Demand of Bus Passengers using Imbalanced Records from Smart Cards," Journal of Computational Analysis and Applications, vol. 32, no. 1, pp. 764–774, 2024.

[2] "Predicting Hourly Boarding Demand of Bus Passengers Using Imbalanced Records From Smart Cards," IEEE Transactions on Intelligent Transportation Systems, 2023.

[3] M. V. S. Goud, M. Suraj, and V. Sumith, "Enhancing Passenger Demand Prediction in Public Transport: Addressing Data Imbalance with DC GAN and Deep Learning," Journal of Engineering Sciences, vol. 16, no. 4, pp. 2076–2079, 2025.

[4] J. Zhang, H. Li, L. Yang, G. Jin, and J. Qi, "STG-GAN: A Spatiotemporal Graph Generative Adversarial Network for Short Term Passenger Flow Prediction in Urban Rail Transit Systems," arXiv preprint arXiv:2205.09731, 2022.

[5] D. Dablain, B. Krawczyk, and N. V. Chawla, "DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data," arXiv preprint arXiv:2102.03749, 2021.

[6] S. Buda, A. Maki, and M. A. Mazurowski, "A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks," Neural Networks, vol. 106, pp. 249–259, 2018.

[7] Z. Kong, D. Zeng, Y. Hu, and L. Li, "Passenger Flow Forecasting with Multi-Graph Convolutional Networks," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 2, pp. 1470–1480, Feb. 2022.

[8] G. Lee, J. Ko, and H. Jung, "Public Transport Demand Forecasting Using Deep Learning and Smart Card Data," Sustainability, vol. 13, no. 6, pp. 3218, 2021.

[9] Y. Wang, M. Ma, and Y. Zhu, "Data Augmentation for Imbalanced Classification Using GANs," IEEE Access, vol. 9, pp. 146944–146954, 2021.

[10] Y. Liu, X. Liu, X. Ma, and H. Wang, "Urban Rail Passenger Flow Forecasting with Attention-Based LSTM and Smart Card Data," IEEE Access, vol. 8, pp. 113554–113566, 2020.