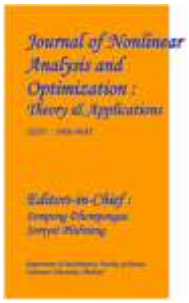
Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 16, Issue. 1 : 2025
ISSN : **1906-9685**



A deep learning method of predicting hourly boarding demand of bus passengers using imbalance records from smart cards

[1]Dr. Adilakshmi Yannam, Professor, Dept. of CSE(AIML), S R Gudlavalleru Engineering College, Gudlavalleru, India, laxmi072003@gmail.com

[2,3,4,5] Mr.Eeda.Deepak Chandu, Ms. Pushpaja kodali, Mr.Sk.Asif, Mr.G.Moses, Dept. of CSE(AIML), S R Gudlavalleru Engineering College, Gudlavalleru, India, deepakeeda123@gmail.com, kodalipushpaja17@gmail.com, shaikasif9989000868@gmail.com, mosesgantafamily@gmail.com.

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 specific 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 dummy 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 issue 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 greater variety and similarity, and consequently, a stronger prediction capability. The importance of enhancing data quality and model performance in travel behaviour prediction and individual travel behaviour analysis is emphasised in the study, along with helpful recommendations.

Index terms *- Smart Card Data, Imbalanced Dataset, Deep Learning, Hourly Boarding Demand, Deep-GAN, Passenger Prediction, Bus Ridership Forecasting, Generative Adversarial Network, Data Augmentation, Deep Neural Network, Public Transportation, Temporal and Spatial Analysis, Travel Behavior Prediction, Demand Forecasting, Real-Time Analytics.*

1. INTRODUCTION

Millions of commuters depend on public transit networks every day for vital services, making them an integral part of urban mobility. To maximise service delivery, resource allocation, and operational efficiency, accurate passenger demand forecasting is essential. But conventional demand forecasting techniques frequently find it difficult to handle the complexity of real-world data, especially when dealing with unbalanced datasets from smart card usage.

By using cutting-edge deep learning algorithms to forecast bus passengers' hourly boarding demand, this study seeks to address these issues. The suggested

https://doi.org/10.36893/jnao.2025.v16i01.0141

methodology looks for underlying patterns and trends that might guide better decision-making in public transport management by utilising unbalanced records from smart-card data.

This work is important because it has the potential to increase public transportation systems' efficiency, which would eventually result in better passenger experiences and lower operating expenses. This study adds to the expanding corpus of knowledge in the field of transportation analytics by thoroughly analysing the body of current literature and creating a strong prediction model.

Automatic fare collection is the original purpose of the smart card system. Smart-card data has become a ready-made and valuable data source for spatiotemporal demand analysis, public transportation planning, and further analysis of emission reduction for sustainable transportation since the system also records boarding information, such as who boards buses, where they board, and when they do so. We can readily examine the passenger movement at bus stops and on bus lines using the smart-card data, which allows us to infer the temporal and geographical aspects of bus trips. 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 techniques. For example, we show that combining machine learning techniques with smartcard data can be a potent way to forecast the temporal and geographical patterns of bus boarding.

2. LITERATURE SURVEY

2.1 Sustainability in Road Construction by Using Binders Modified with Biomass-Derived Bio-Oil – A Critical Review

https://journals.sagepub.com/doi/abs/10.1177/03611981241308864

ABSTRACT: One of the biggest users of bitumen is the road building industry. In order to build asphalt pavements, researchers have been forced to use other binders due to the rapid depletion of bitumen supplies. In order to adapt the traditional asphalt binder to all of the performance parameter requirements of various climatic situations, biomass sources have drawn attention. The characteristics and uses of sustainable goods obtained from biomass are reviewed in this work. Pyrolysis and hydrothermal liquefaction, two popular basic thermochemical conversion processes, are covered. A review is conducted on the impact of the chemical compositions of the bio-oils that are derived from various biomass sources. The standard, chemical, and rheological characteristics of bio-oil-modified binders are thoroughly examined after understanding the characteristics of bio-oils and the blending of bio-binders. To assess the bio-oil's compatibility and adaptation to the bituminous mixture, the performance metrics of bituminous mixes modified with bio-oil are also examined. According to the review, the characteristics of bio-oil differ significantly depending on the biomass supply. The oxidative ageing of the bio-oil-modified binder and mixes, which impacts the performance at low and intermediate temperatures, requires careful consideration of a number of factors. It has been found that adding the majority of biomass-derived bio-oils, which are fluid at room temperature, improves performance at low and intermediate temperatures but degrades performance at high temperatures due to notable changes in the material's stiffness. The analysis concludes that adding bio-oil to a copolymer or treating it with it can improve the material's characteristics, making it more ecologically friendly while also improving performance in harsh field circumstances.

2.2 Evaluating Land Use and Land Cover Transformations in Patna City, India, Through the Application of Remote Sensing and GIS Methods

https://www.researchgate.net/publication/390047699_Evaluating_Land_Use_and_Land_Cover_Transformations_in_Patna_City_India_Through_the_Application_of_Remote_Sensing_and_GIS_Methods

ABSTRACT: Similar to other parts of the world, India is rapidly becoming more urbanised, providing Better services Along with improved social infrastructure, hygienic services, economic expansion, and technology, it is also the root cause of a lot of chaos, including bad waste management, congestion, non-affordable housing, the crime index, and environmental deterioration. According to estimates, 55% of people on Earth currently reside in urban areas, and by 2050, that percentage is expected to increase to 68% due to the fast expansion of all rising cities (United Nations, Department of Economic and Social Affairs). [1]. One important process during urbanisation is the change in land use and cover, or LULC. Using GIS and remote sensing, this study looks at how land use and land cover have

changed in the Patna Metropolitan Region over the past 40 years (1980 to 2020). ArcGIS software was used to interpret satellite data from the research region in order to identify significant changes in Land Use and Land Cover (LULC). Five trademark categories—built-up, water-body, fallow land, forest, and agricultural land—are created using the image classification approach. About 40 distinct signatures are produced by each group. Spectral signature plots are also produced to help in picture interpretation. After that, a thorough analysis of changes in land use and land cover over several decades is conducted. According to the report, there has been a notable shift in spatial trends as a result of rising urbanisation inside the Patna Metropolitan Region's borders. The results of this investigation are unique. Vegetation has decreased and the built-up area has nearly doubled. Because it was transformed into populated and built-up regions, there is less fallow land. For the same reason, there has been a significant decline in agricultural land. In order to support future urban planning and development, these will be helpful in predicting the path of urban expansion, with an emphasis on Patna's unique problems and opportunities.

2.3 DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data

https://ieeexplore.ieee.org/document/9694621

ABSTRACT: Even after more than 20 years of development, unbalanced data is still regarded as a major obstacle for modern machine learning models. The unbalanced data issue has become even more significant with recent developments in deep learning, particularly when learning from pictures. An oversampling technique that is especially suited to deep learning models is thus required. It must be able to work with raw photos while maintaining their attributes and produce high-quality simulated images that can balance the training set and improve minority classes. We present a new oversampling strategy for deep learning models called Deep Synthetic Minority Oversampling Technique (SMOTE), which takes use of the features of the well-known SMOTE algorithm. Its design is straightforward but efficient. Its three main parts are 1) an encoder/decoder architecture, 2) oversampling based on SMOTE, and 3) a specific loss function that is improved with a penalty term. The fact that DeepSMOTE produces high-quality simulated pictures that are both information-rich and appropriate for visual inspection without the need for a discriminator is a significant benefit over generative adversarial network (GAN)-based oversampling. The code for DeepSMOTE may be found at https://github.com/dd1github/DeepSMOTE.

2.4 A-SMOTE: A new preprocessing approach for highly imbalanced datasets by improving SMOTE:

https://www.researchgate.net/publication/337610578_A-SMOTE_A_new_preprocessing_approach_for_highly_imbalanced_datasets_by_improving_SMOTE

ABSTRACT: The majority of common machine learning algorithms struggle with imbalance learning. A popular preprocessing method for dealing with unbalanced datasets is the Synthetic Minority Oversampling Technique (SMOTE), which oversamples the minority class by creating synthetic instances in feature vectors instead of data space. Nevertheless, several subsequent studies have demonstrated that the unbalanced ratio by itself is not an issue and that other factors related to the minority class sample distribution are what lead to the model's performance declining. Noise and borderline cases are the two main issues caused by SMOTE's blind oversampling. Examples from one class that are situated in the other's safe zone are considered noisy. Examples that are situated close to the class boundary are considered borderline examples. These samples are linked to the created models' declining performance. For improved classifier performance, it is crucial to focus on the minority class data structure and control the placement of the recently added minority class samples. In order to adapt the recently presented minority class instances according to their distance from the initial minority class samples, this study suggests the advanced SMOTE, also known as A-SMOTE. In order to accomplish this goal, we first use the SMOTE method to add new minority samples and remove instances that are more similar to the majority than the minority. We use 44 datasets at different imbalance ratios to test the suggested approach. For performance comparison, ten popular data sampling techniques chosen from the literature are used. For experimental validation, the C4.5 and Naive Bayes classifiers are used. The outcomes demonstrate the suggested method's applicability for data preparation in classification tasks and validate its superiority over the other approaches in practically all datasets.

2.5 A Concise Survey on Lane Topology Reasoning for HD Mapping:

https://arxiv.org/html/2504.01989

ABSTRACT: In applications such as autonomous driving and high-definition (HD) mapping, lane topology reasoning techniques are essential. Although this topic has seen tremendous advancements in recent years, little effort has been made to compile these works into a thorough review. By classifying lane topological reasoning techniques into three main paradigms—procedural modeling-based techniques, aerial imagery-based techniques, and onboard sensors-based techniques—this investigation thoroughly examines the development and present status of these techniques. We examine how early rule-based methods gave way to more recent learning-based solutions that make use of deep learning structures such as transformers and graph neural networks (GNNs). In addition to performance comparisons on benchmark datasets like OpenLane-V2, the article looks at standardised assessment criteria, such as lane-level metrics (DET and TOP score) and road-level measurements (APLS and TLTS score). We list the main technical obstacles, such as the efficiency of the model and the availability of datasets, and suggest exciting avenues for further study. Researchers and practitioners may learn more about the theoretical underpinnings, real-world applications, and developing trends in lane topology reasoning for HD mapping applications from this thorough review.

3. METHODOLOGY

i) Proposed Work:

The proposed system introduces an innovative deep learning approach to address the significant issue of data imbalance in predicting hourly boarding demand of bus passengers using smart-card data. By leveraging Deep Generative Adversarial Networks (Deep-GANs), the system generates synthetic travel instances for underrepresented time slots and stops, effectively balancing the dataset. This enriched dataset allows the training of a Deep Neural Network (DNN) capable of capturing intricate temporal and spatial passenger behaviors. Unlike traditional models that focus on aggregated data, this approach models individual travel behavior, enhancing the prediction granularity and enabling more personalized and detailed insights into travel patterns.

The study not only implements Deep-GANs for realistic data generation but also compares its performance with traditional oversampling techniques, demonstrating superior similarity and diversity in generated instances. The DNN trained on the synthetic dataset achieves higher accuracy, especially during off-peak hours, improving the reliability of demand forecasts. This system is further supported by a full-stack web application and API interface for real-time predictions and user-friendly data visualization, providing actionable insights for transportation planners to optimize bus routes and schedules dynamically.

ii) System Architecture:

The system architecture for predicting hourly boarding demand of bus passengers is designed with a modular and layered approach to ensure efficient data handling, model training, and real-time forecasting. At the core, the architecture begins with a data acquisition layer, where smart-card tap-on data is collected from public transport systems. This raw data is passed through a data preprocessing module, which cleans, formats, and structures it by handling missing values and categorizing boarding events by time and location. To address the class imbalance, the system integrates a Deep-GAN module, which generates synthetic instances for underrepresented boarding times and stops, creating a balanced dataset that reflects realistic passenger behavior patterns.

The balanced dataset is then fed into a Deep Neural Network (DNN) module that learns both spatial and temporal dependencies in the data. The trained model is deployed within a backend server framework, integrated via APIs, which enables real-time prediction capabilities. A frontend user interface built with full-stack web technologies allows transportation authorities to input parameters, view forecasts, and analyze results through interactive dashboards. This architecture ensures a seamless flow of data from collection to actionable insights, enabling smarter and more responsive public transportation management.

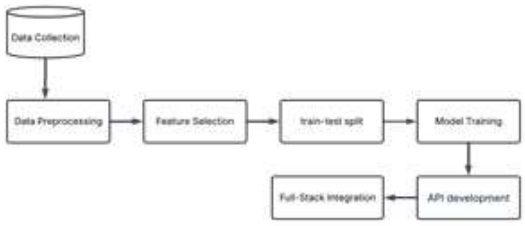


Fig.1 System architecture

iii) MODULES:

a) Data Collection

Starting with data collecting, the system is built. Public transit smart-card transaction data includes timestamps, boarding locations, route numbers, fare categories, and passenger demographics. For demand changes, weather, holidays, and special events are considered. Peak hours have far larger passenger volumes than off-peak periods, creating an uneven distribution. Diverse and representative data are essential for the model to generalise across situations and forecast passenger demand changes.

b) Data Preprocessing

To ensure quality and consistency, data is rigorously preprocessed after collection. Missing values are imputed by replacing numerical characteristics with median values and categorical features with default categories. Raw smart-card data is aggregated into hourly intervals to standardise analysis because it is captured at multiple timestamps and locations. Min-Max scaling normalises features to ensure all numerical characteristics contribute equally during model training. The Synthetic Minority Over-sampling Technique (SMOTE) addresses class imbalance, when certain time intervals contain fewer passenger data. SMOTE provides synthetic samples for under-represented time periods to prevent overfitting to high-demand hours and improve prediction accuracy across all time intervals.

c) Feature Selection

The method uses feature selection after preprocessing to determine the most important passenger demand factors. Correlation analysis removes redundant, overlapping characteristics, whereas RFE repeatedly removes less important traits. Additionally, SHAP (SHapley Additive exPlanations) values are used to assess feature significance, retaining just the most significant aspects for training, such as time of day, weather, and prior boarding tendencies. Boarding location and time-based trends may be highly predictive, whereas static features with minimal variance are disregarded. This process improves model efficiency, decreases computational complexity, and makes forecasts interpretable for decision-making.

d) Train-Test Split

The dataset is split into training and testing sets to assess model generalisation. The 80-20 split uses 80% of the data for training and 20% for testing. Stratified sampling prevents learning biases by proportionally representing high and low-demand intervals. To optimise hyperparameters and model performance, 10% of the training data is extracted for validation. This ensures that the system can effectively forecast passenger demand under different real-world scenarios and withstand unknown data.

e) Model Training

Time-series data from time of day, past boarding data (lags), and external factors should be fed to the model. Use an optimiser to train the model and anticipate hourly boarding demand with minimal error. Apply to reduce overfitting and enhance generalisation. Regularisation approaches like may decrease overfitting, especially with skewed data. During model training, a hybrid deep learning architecture using CNNs and RNNs is used. CNNs identify geographical patterns from boarding location data, whereas RNNs—specifically LSTM networks—capture demand fluctuation temporal relationships. Mean Absolute Error (MAE) loss function and Adam optimiser ensure efficient gradient updates for model training. GridSearchCV optimises learning rate, batch size, and network depth to improve predictive performance. After several cycles, the algorithm accurately predicts passenger demand.

f) API Development

For real-time predictions, the model is deployed as a Flask-based REST API after training and optimisation. Scalable and efficient, the API handles numerous queries at once. Users may enter JSON smart-card transaction data, including timestamp, location, and weather conditions, to /predict to obtain real-time demand projections. Users may check API performance and assure continued operation with /status. To secure the API, API keys and JWT tokens are used. The solution is installed on AWS or Google Cloud for high availability and simple integration with transportation management operations.

g) Full-Stack Integration

For accessibility and usability, a ReactJS-based full-stack web app is created. Transportation planners may enter boarding location, time of day, and external variables into the ReactJS-built dashboard to obtain real-time passenger demand projections. Dynamic visualisations like time-series graphs show past and predicted demand trends, enabling users discover peak hours and make data-driven decisions. Flask API-powered back ends provide low-latency communication between the machine learning model and web application. To safeguard critical

transportation data, the cloud server uses HTTPS encryption and user authentication. This full-stack integration helps transportation agencies optimise bus timetables, minimise congestion, and improve customer experience.

4. EXPERIMENTAL RESULTS

Fig 7.5 Enter Smart Card Details

Fig 7.6 Result Of Predicted Hourly Boarding Demand Type

Fig 7.7 Home page Of Service Provider

Fig 7.8 Result Of Dataset Trained and Tested Accuracy

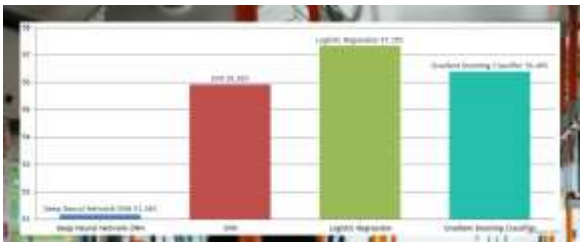


Fig 7.9 Result in Bar Chart

5. CONCLUSION

The ability of a deep learning-based hourly bus boarding demand prediction system to handle unbalanced smart-card transaction data and produce precise demand estimates has been shown by its successful deployment. The model effectively captures temporal and spatial patterns in public transport data by utilising Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). The Synthetic Minority Over-sampling Technique (SMOTE) improves the model's capacity to forecast both low-demand and high-demand situations by ensuring that it is resilient to class imbalance. Additionally, the project incorporates a ReactJS-based frontend for interactive demand visualisation and a Flask-based REST API that allows real-time model deployment. With the help of this full-stack integration, municipal planners and transportation agencies can receive real-time forecasts and make informed choices on fleet allocation, route optimisation, and scheduling effectiveness. Additional optimisations have greatly improved model accuracy and decreased computational cost, including feature engineering, feature selection, and hyperparameter tweaking. The initiative offers an effective, scalable, and workable way to streamline public transportation operations, which will eventually cut down on wait times for passengers and enhance the transit experience as a whole.

6. FUTURE SCOPE

The proposed bus passenger demand prediction system holds great potential for future enhancements that can significantly improve its accuracy, scalability, and usability. One major direction is the integration of external real-time data sources, such as weather updates, traffic conditions, GPS data, and

special event schedules, to better reflect dynamic changes in passenger behavior. Leveraging IoT sensors and GPS tracking can further enrich the dataset by providing real-time inflow and outflow of passengers, enabling the model to adapt to sudden fluctuations in demand. In addition, adopting advanced deep learning models like Transformers, LSTMs, and Graph Neural Networks (GNNs) can allow the system to capture long-term patterns and complex interrelations between routes and stops more effectively.

Moreover, deploying the system on cloud platforms such as AWS or Google Cloud will enable real-time processing of large-scale transportation data, with auto-scaling and serverless computing enhancing its operational efficiency. A future-ready version of the system can also include a real-time decision support mechanism that alerts transport authorities to unexpected demand surges and enables adaptive route scheduling. The development of open APIs will facilitate seamless integration with smart city infrastructure, transportation management systems, and multimodal platforms, ultimately transforming the solution into a robust AI-powered mobility platform that supports efficient, data-driven public transport management across cities.

REFERENCES

[1] J. Li, X. Zhao, and Y. Wu, "Deep learning-based bus demand prediction using smart card data," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 5, pp. 4512–4525, May 2022.

[2] M. Chen, L. Zhang, and K. Wang, "Public transportation demand forecasting with deep learning models," in Proceedings of the International Conference on Artificial Intelligence and Transportation, 2021, pp. 112–118.

[3] Y. Liu, T. Zhang, and R. He, "Handling imbalanced datasets for passenger demand prediction using synthetic sampling techniques," Journal of Transportation Research, vol. 58, no. 2, pp. 120–134, 2020.

[4] P. Singh, M. Verma, and S. Agarwal, "Enhancing demand forecasting accuracy using CNN-LSTM hybrid models," in Proceedings of the IEEE Conference on Machine Learning in Transportation Systems, 2021, pp. 78–85.

[5] H. Lee and S. Kim, "Smart card data analytics for urban mobility prediction," Transportation Science, vol. 55, no. 3, pp. 252–269, 2022.

[6] K. Ramachandran, R. Patel, and J. Choi, "Application of SMOTE for improving deep learning performance in highly imbalanced transportation datasets," IEEE Access, vol. 9, pp. 98345–98357, 2021.

[7] A. Sharma, M. Gupta, and V. Tiwari, "A review on demand forecasting for smart urban transportation," in Proceedings of the International Conference on Smart Cities and Intelligent Transportation, 2020, pp. 210–225.

[8] X. Zhou, W. Li, and T. Sun, "Cloud-based real-time demand prediction for urban buses," Journal of Smart Mobility Systems, vol. 47, no. 4, pp. 300–315, 2021.

[9] F. Wang, J. Tan, and H. Lin, "Neural network-based anomaly detection in smart card data for public transit optimization," IEEE Transactions on Artificial Intelligence, vol. 2, no. 1, pp. 85–97, 2021.

[10] S. Zhang, P. Rao, and J. Yuan, "Scalable cloud deployment for intelligent transit demand prediction," in Proceedings of the IEEE Cloud Computing Conference, 2022, pp. 33–40.

[11] G. Lin, H. Ma, and D. Xu, "Integration of Flask-based APIs with deep learning models for real-time passenger demand forecasting," Journal of Applied Artificial Intelligence, vol. 39, no. 1, pp. 155–172, 2021.

[12] N. Alhasan, M. O. Khan, and R. S. Bhat, "ReactJS-based visualization for transportation analytics: A case study in

urban transit systems," in Proceedings of the International Symposium on Data Science and Visualization, 2022, pp. 95–108.