Multimodal Transportation 5 (2026) 100323

Contents lists available at ScienceDirect

Multimodal Transportation

journal homepage: www.elsevier.com/locate/multra

Full Length Article

Trust-constrained multi-objective hybrid ensemble framework for spatiotemporal passenger demand prediction for autonomous taxi systems

Mongkut Piantanakulchai [a,*], Adeel Munawar [a,b]

[a] Sirindhorn International Institute of Technology, Thammasat University, Khlong Luang, Pathum Thani, 12120, Thailand
[b] Institute of Economics, Karlsruhe Institute of Technology (KIT), Karlsruhe, 76131, Germany

ARTICLE INFO

Keywords:
Autonomous taxis
Intelligent transportation systems
Spatiotemporal forecasting
Hybrid ensemble learning
Trust-constrained optimization
Multi-objective optimization
Smart mobility

ABSTRACT

Accurate taxi-demand forecasting has become a core requirement for modern smart-city planning and Intelligent Transportation Systems (ITS). As autonomous vehicles, especially Autonomous Taxis (ATs), move toward mainstream deployment, cities need reliable predictions to manage wait times, balance supply with demand, and support efficient fleet operations. This study introduces a trust-constrained hybrid ensemble designed to strengthen AT pick-up demand prediction. The framework merges the distinct advantages of TabNet, a Multi-Layer Perceptron, and Stochastic Gradient Boosting to capture complex spatiotemporal patterns. Instead of assigning equal or heuristic weights, the model applies a multi-objective optimization routine that jointly considers R^2, MAE, and Huber loss within a trust-region solver to improve stability and generalization. The evaluation uses a large-scale dataset of more than 4,700 taxis in Bangkok, enriched with spatiotemporal attributes and synchronized weather variables such as temperature, humidity, rainfall, wind speed, and atmospheric pressure. Simulation results demonstrate that the proposed ensemble consistently outperforms standalone models and equal-weight baselines, achieving a testing RMSE of 9.69, MAE of 5.97, and R^2 of 0.924. These findings highlight the framework's predictive capability and practical potential for improving fleet management, optimizing vehicle deployment, and reducing idle times, thereby supporting data-driven decision-making for efficient and intelligent AT systems.

1. Introduction

The demand for efficient and sustainable smart transportation is being increasingly recognized worldwide as a result of rapid urbanization and increasing population density. In the context of modern smart cities, Intelligent Transportation Systems (ITS) have been at the forefront of identifying solutions to these emerging challenges. Out of the available and groundbreaking solutions, Autonomous Vehicles (AVs) and particularly Autonomous Taxis (ATs) have promised to provide substitutions to conventional mobility by allowing safer, cleaner, and more flexible transportation services (Aldakkhelallah and Simic, 2021; Almihat et al., 2022; Lee et al., 2016).

Most city centers, particularly those with major tourist attractions, commercial districts, or historical sites, experience severe traffic congestion, especially during peak hours. These conditions create localized demand surges, leading to vehicle bunching, extended

* Corresponding author.
E-mail address: mongkut@siit.tu.ac.th (M. Piantanakulchai).

https://doi.org/10.1016/j.multra.2026.100323
Received 6 April 2026; Received in revised form 17 May 2026; Accepted 18 May 2026

passenger waiting times, and inefficient fleet operations. The root cause lies in the imbalance between taxi availability and real time passenger demand, which remains unresolved due to the lack of effective dynamic allocation strategies. This situation creates an opportunity for urban planners and transportation operators to develop adaptive and evidence based solutions that reduce congestion and mitigate negative environmental impacts as travel demand patterns continue to evolve.

Given the fundamental role of ATs in modern urban mobility, accurately forecasting passenger demand is crucial for managing waiting times effectively, aligning supply with demand, and enhancing operational efficiency. Data-driven forecasting models can significantly contribute to these goals by identifying spatial and temporal demand patterns, enabling dynamic fleet allocation, and minimizing vehicle downtime. However, the inherently complex and dynamic nature of urban transportation systems limits the effectiveness of traditional demand forecasting methods, which often rely on static or historical data and lack the flexibility to adapt to real-time changes in traffic conditions and passenger behavior. As a result, these methods result in inefficient fleet deployment, longer travel distances, lower profitability for drivers, and increased carbon emissions. Accurate demand forecasting not only helps reduce fuel consumption and environmental impact but also improves service reliability and customer satisfaction. Unlike traditional approaches, Machine Learning (ML) models offer high adaptability and accuracy, especially when integrated using hybrid ensemble frameworks that combine complementary architectures to capture complex spatiotemporal demand patterns. By leveraging real-time contextual features such as weather conditions, traffic flows, and temporal trends, they provide a foundation for operating intelligent AT systems.

Minimizing cruising distance is one of the most effective strategies for improving driver profitability. However, many taxis spend excessive time cruising in search of passengers (Schaller Consulting, 2004; Qin et al., 2017). This inefficiency leads to excessive fuel consumption, greater wear and tear on vehicles, and increased operational costs. Additionally, prolonged search times can lead to driver fatigue and reduced job satisfaction, underscoring the importance of accurate demand forecasting to optimize dispatching and improve the overall efficiency of taxi services. To address these challenges, this study proposes a trust-constrained (Byrd et al., 1999; Conn et al., 2000) multi-objective hybrid ensemble framework that integrates the strengths of TabNet (Arik and Pfister, 2021), Multi-Layer Perceptron (MLP) (Hornik, 1989), and Stochastic Gradient Boosting (SGB) (Friedman, 2002). TabNet provides hierarchical feature representation through attention-based selection, MLP captures nonlinear dependencies, and SGB reduces residual errors through boosting. The ensemble weights are optimized using a trust-region constrained multi-objective function that maximizes the coefficient of determination (R^2) while minimizing Mean Absolute Error (MAE) and Huber loss. This approach enhances predictive balance, stability, and generalization across temporal segments.

Our methodology utilizes a comprehensive dataset from over 4700 taxis in Bangkok, Thailand (iTIC Foundation, 2021). The data spanned four months from September 2021 to December 2021. To enhance the framework's ability to capture external factors affecting taxi demand, we integrated weather data obtained from the Meteostat platform (Meteostat, 2024), covering parameters such as temperature, humidity, and precipitation, which are known to significantly influence passenger demand patterns in urban environments (Chen et al., 2017). These weather conditions were synchronized with the taxi data to ensure temporal and spatial alignment. The dataset serves as a benchmark for evaluating temporal stability across different daily time segments, enabling a deeper assessment of the framework's robustness.

While ensemble learning has been widely applied in taxi and ride-hailing demand forecasting, weight determination is often based on equal or empirically tuned combinations rather than an explicit constrained multi-objective optimization formulation. In addition, most approaches evaluate performance using individual error metrics without jointly optimizing explanatory power (R^2), absolute deviation (MAE), and robustness to outliers within a unified objective framework. Our findings indicate that the proposed framework addresses these methodological limitations and significantly outperforms individual prediction models across several key performance metrics, including MAE, RMSE, and R^2. The optimized ensemble consistently achieved higher predictive accuracy and improved residual compactness compared with all base learners and the equal-weight ensemble. These improvements demonstrate the effectiveness of the trust-region-based multi-objective optimization in balancing error minimization and model stability. The proposed approach enhances fleet management, reduces passenger wait times, and improves the operational efficiency of autonomous taxi systems in smart cities. With strong generalization capability, the framework enables proactive dispatching and dynamic fleet reallocation, optimizing service availability and minimizing idle time, thereby contributing to sustainability objectives through reduced fuel consumption and improved energy efficiency in alignment with SDG 11.

The main contributions of this study are as follows.

1. We propose a trust-region-based multi-objective hybrid ensemble framework integrating TabNet, MLP, and SGB to enhance accuracy and robustness in ATs demand forecasting.
2. We formulate a multi-objective optimization approach that simultaneously maximizes R^2 while minimizing MAE and Huber loss, leading to improved predictive performance and temporal stability compared with traditional ensemble baselines.

To validate our methodology, we construct a large-scale dataset of more than 4700 taxis operating in Bangkok, enriched with temporally aligned weather data and structured into a spatiotemporal grid format. The results provide practical insights into how such hybrid ensemble learning frameworks can support real-time fleet dispatching, minimize cruising inefficiencies, and facilitate the integration of intelligent prediction models within modern ITS.

The remainder of this paper is organized as follows. The next section presents a review of relevant literature on autonomous taxi demand forecasting and ensemble learning techniques. This is followed by the methodology section, which describes the dataset, preprocessing steps, individual model configurations, and the ensemble formulation. The results and discussion section then presents the performance evaluation, model comparison, and practical implications. Finally, the conclusion summarizes the key findings and outlines potential directions for future research.

2. Literature review

The rapid development of smart cities has led to growing interest in AVs, particularly ATs, as a means to enhance urban mobility. AVs are widely recognized for their potential to reduce traffic congestion, improve safety, and provide flexible transportation services (Campisi et al., 2021). Among these, ATs offer on-demand capabilities that align well with the dynamic and complex nature of urban environments. However, their successful deployment relies heavily on accurate passenger demand forecasting, which remains a challenging task due to spatial and temporal variability in urban mobility patterns (Faisal et al., 2019).

To address these challenges, a number of recent studies have employed ensemble learning and hybrid methodologies aimed at improving prediction accuracy (Munawar and Piantanakulchai, 2025a, 2025b, 2025c). yet most existing frameworks rely on static or heuristic ensemble strategies without optimization across heterogeneous learners. Wu and Levinson (2022) applied Random Forest and linear ensemble models to forecast For-Hire Vehicle (FHV) demand in New York and Chicago, incorporating socioeconomic, temporal, and land use features. However, their approach lacked deeper temporal learning and showed limited generalization across cities. Sarkar et al. (2019) proposed a multi-stage pipeline incorporating OD zone clustering, time binning, and Random Forest modeling using NYC TLC data. While effective in capturing recurring temporal patterns, the model did not incorporate exogenous factors such as weather or events, limiting its adaptability. Similarly, Rajak and Baruah (2020) used multiple statistical and machine learning techniques including SMA, WMA, EMA, Linear Regression, Random Forest, and XGBoost on New York Yellow Taxi data. The ensemble combined these models using fixed strategies without optimization, and the spatial characteristics of the trips were not explicitly modeled.

Zhang and Zhao (2021) introduced a Clustering-aided Ensemble Method (CEM) that trained local models on origin-destination pairs from Chinese ride-sourcing data. While it enhanced localized predictions, it demanded significant tuning and lacked cross-regional validation. A more advanced approach was proposed by Ye et al. (2022), who developed the Deep Decomposition Forecasting Model (DDFM) combining Ensemble Empirical Mode Decomposition (EEMD) with LSTM. Their model effectively managed nonlinear and noisy demand patterns, though it was computationally intensive and lacked spatial context. Carson-Bell et al. (2021) applied neural networks, AdaBoost, and Random Forests to forecast ride-hailing demand using spatial and census tract features from Chicago's TLC dataset, but did not include external variables such as weather or mobility trends, potentially limiting model generalizability.

In another line of work, Lai et al. (2019) developed a hybrid LSTM model incorporating spatiotemporal features such as GPS, weather, and POI data from Xiamen and Chengdu via the Didi Chuxing platform. Although the model demonstrated strong spatiotemporal learning capability, it exhibited high sensitivity to hyperparameters and risked overfitting. Sonbhadra et al. (2020) offered two distinct modeling pipelines using New York City taxi data, one based on clustering, time binning, and ensemble learning, and another on spectral feature extraction via Fourier transforms coupled with a deep neural network. These approaches faced challenges in integrating features effectively and yielded less robust results in data-sparse conditions.

A comparative summary of these studies, including datasets, model types, feature usage, and identified limitations, is presented in Table 1. This consolidated view highlights recurring methodological limitations, including insufficient spatial integration, limited inclusion of exogenous features (e.g., weather), and the use of heuristic rather than explicitly optimized ensemble weighting strategies.

Recent advances in deep spatiotemporal learning have produced architectures such as long short-term memory networks (LSTM), gated recurrent units (GRU), spatiotemporal graph convolutional networks (ST-GCN) (Yu et al., 2018), diffusion convolutional recurrent networks (DCRNN) (Li et al., 2018), Graph WaveNet (Wu et al., 2019), and ST-Transformer variants. These models excel when

Table 1
Comparison of related studies on taxi demand forecasting.

Study	Dataset	Model	Features	Limitations / Strengths
Wu and Levinson (2022)	FHV trip data, New York and Chicago	RF, linear ensemble	Socioeconomic, temporal, land use	Lacks temporal sequence modeling; limited generalizability across cities
Sarkar et al. (2019)	NYC TLC trip data, New York	Clustering + RF ensemble	OD zones, time bins	No external features (e.g., weather); spatial clustering used, but no external context
Rajak and Baruah (2020)	Yellow Taxi dataset, New York	SMA, WMA, EMA, LR, RF, XGBoost ensemble	Region, fare, distance, travel time	Spatial context not explicitly modeled; ensemble used fixed (non-optimized) weights
Zhang and Zhao (2021)	Ride-sourcing OD data, China	Clustering-aided Ensemble	OD clusters, time	High tuning complexity; no validation across cities
Ye et al. (2022)	Urban taxi demand data, China	EEMD + LSTM (DDFM)	Decomposed demand time series	High computational cost; spatial and contextual features not integrated
Carson-Bell et al. (2021)	Ride-hailing records, Chicago	NN, AdaBoost, RF	Pickup coordinates, census tract features	External features (e.g., weather) not included
Lai et al. (2019)	Didi Chuxing GPS data, Xiamen and Chengdu	Hybrid LSTM	POIs, weather, temporal, GPS	Performance may decline in regions with sparse contextual data or differing mobility patterns
Sonbhadra et al. (2020)	NYC taxi data, New York	(1) Clustering + binning + ensemble, (2) Fourier-based features + DNN	GPS trajectories, temporal bins, spectral transforms	Lack of integrated feature learning; poor performance in sparse data settings
This Study	GPS and Meteostat data, Bangkok	Trust-constrained hybrid ensemble (TabNet, MLP, SGB)	Spatial, temporal, and weather features	Multi-objective optimization integrating diverse learners; improves stability and generalization over heuristic ensembles

the prediction task is framed over continuous high-resolution time series with well-defined spatial adjacency (e.g., METR-LA, PeMS, NYC TLC). The present study, however, formulates demand forecasting as a structured tabular regression problem at the (grid-cell, hour) level, in which spatial relationships are encoded through grid identifiers and temporal patterns through indicator features and synchronised meteorological variables. Within this regime, ensemble learning over heterogeneous tabular learners has been shown to be both effective and parameter-efficient (Arik and Pfister, 2021; Friedman, 2002), motivating the methodological positioning of the present work.

In contrast to the aforementioned studies, this study proposes a trust-region-based multi-objective hybrid ensemble framework for spatiotemporal passenger demand forecasting in ATs mobility systems. The limitation related to insufficient spatial integration is addressed through a structured grid-based spatial representation combined with temporal demand features. The limited inclusion of exogenous variables is resolved by incorporating weather attributes derived from GPS and Meteostat records in Bangkok. Furthermore, instead of relying on heuristic or equal-weight ensemble strategies, ensemble weights are determined through a trust-region constrained multi-objective optimization that jointly maximizes predictive accuracy (R^2) and minimizes MAE and Huber loss. By explicitly integrating spatial structure, external factors, and optimized ensemble weighting within a unified framework, this study provides a methodologically grounded approach for ATs demand forecasting in ITS applications.

3. Methodology

The framework integrates multiple ML models to exploit their complementary capabilities in handling nonlinear, high-dimensional, and spatiotemporal data. Specifically, the ensemble combines TabNet, MLP, and SGB, forming a hybrid structure that enhances both predictive accuracy and generalization stability.

This methodology is organized into the four primary phases. First, large-scale real-world data were collected from GPS-based taxi mobility records and temporally synchronized with external weather data. Then, a preprocessing pipeline was applied to ensure spatial and temporal consistency across all features, such as grid-based aggregation and normalization. Second, each base model was independently trained on the processed dataset to capture distinctive patterns from the data: TabNet uses attention-based feature selection to extract hierarchical relationships; MLP models nonlinear dependencies among temporal and environmental features; and SGB refines residual errors through gradient-boosted decision trees. Third, the predictions of all base learners were integrated through a hybrid ensemble formulation. Unlike the common practice in various conventional ensembles, where equal or heuristic weight assignments are used, this framework employs trust-constrained multi-objective optimization to derive the optimal weights for the base learners. The optimization maximizes the R^2, while minimizing MAE and Huber loss simultaneously, ensuring a balanced and robust predictive output. Finally, the optimized ensemble was evaluated on a separate test set to assess predictive accuracy, residual compactness, and temporal stability. Finally, this multi-phase design helps ensure the resulting framework not only achieves high predictive performance but, more importantly, maintains reliability across different time segments and demand conditions.

The overall workflow of the framework is illustrated in Fig. 1. Raw GPS mobility data and hourly weather observations enter the preprocessing stage, where they are cleaned, integrated, and aggregated into a structured spatiotemporal dataset at the (grid-cell, hour) level. Three complementary base learners (TabNet, MLP, and SGB) are then trained independently on this dataset. Their validation predictions are passed to the trust-region constrained solver, which optimises the ensemble weights against the hybrid multi-objective loss in Eq. (17). The optimised weights are finally applied to the test predictions to produce the ensemble forecast and the corresponding evaluation metrics.

3.1. Data description and preprocessing

This study utilizes a comprehensive real-world dataset provided by the Intelligent Traffic Information Center Foundation (iTIC Foundation, 2021), comprising GPS-based mobility data collected from more than 4700 taxis operating in Bangkok, Thailand. The data set spans four months, from September to December 2021. Each record contains key attributes such as taxi ID, timestamp, GPS coordinates, speed, direction, and meter status, enabling the extraction of occupancy information and trip activity across different spatial and temporal contexts.

To enrich the mobility data with environmental context, hourly weather observations were obtained from the Meteostat platform (Meteostat, 2024). This dataset includes atmospheric variables such as temperature, humidity, precipitation, wind speed, wind direction, and air pressure. The weather data were temporally synchronized with taxi trajectories to support spatiotemporal modeling under realistic urban operating conditions.

Integrating mobility and environmental datasets ensured a comprehensive representation of demand-driving factors in Bangkok's metropolitan area. The preprocessing pipeline consisted of the following steps:

3.1.1. Data cleaning

Missing values, duplicate entries, and GPS anomalies were filtered out. Taxi coordinates falling outside the Bangkok metropolitan boundary were excluded to maintain spatial relevance (Fig. 2).

3.1.2. Data integration

Taxi mobility data and weather records were merged using timestamp alignment, producing a unified dataset containing both spatiotemporal and environmental attributes.

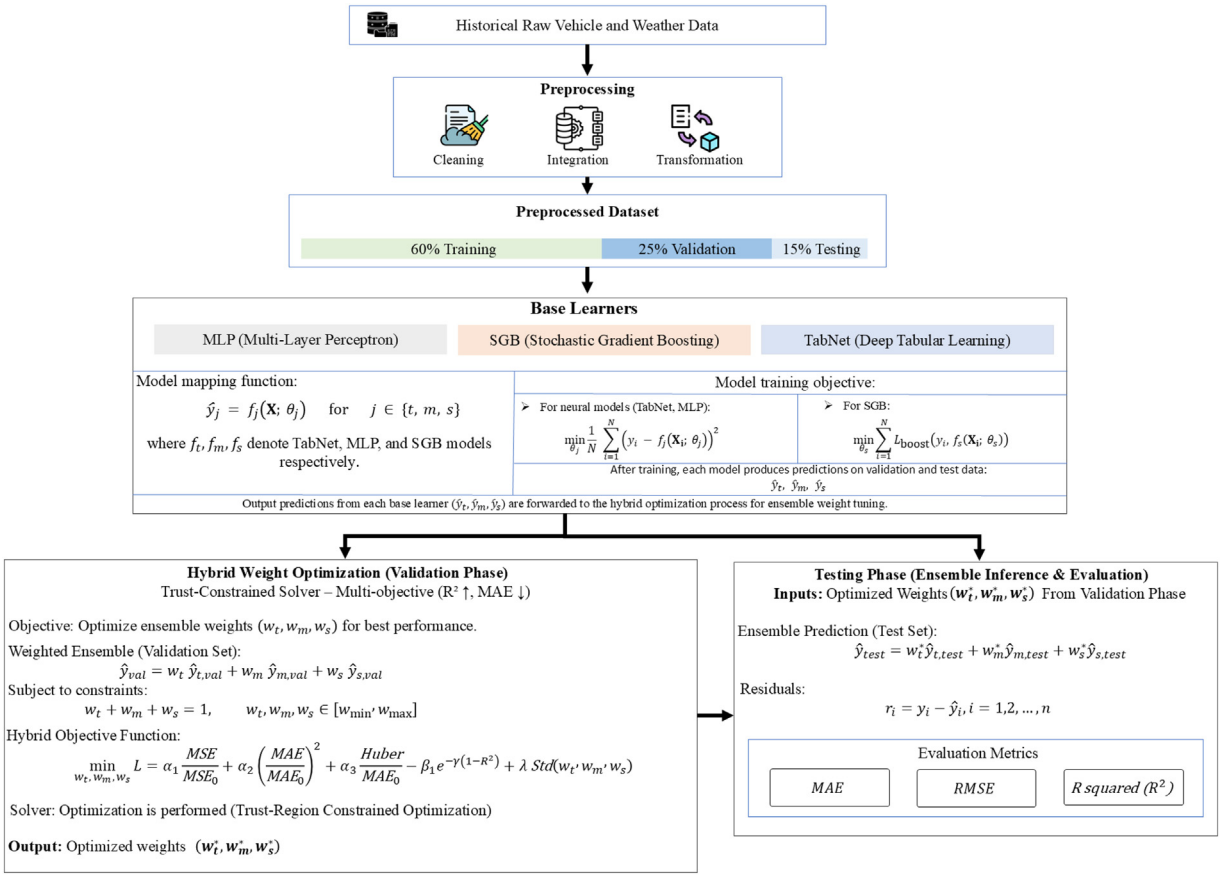

Fig. 1. Proposed trust-constrained hybrid ensemble framework for ATs demand prediction.

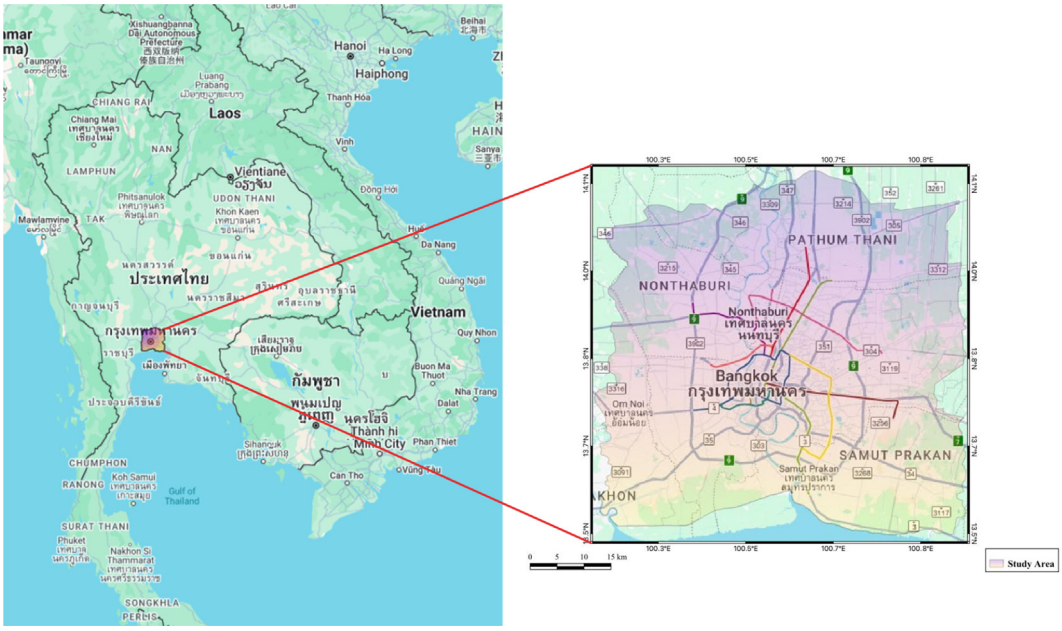

Fig. 2. Study area and spatial boundaries used for autonomous taxi demand analysis in Bangkok, Thailand.

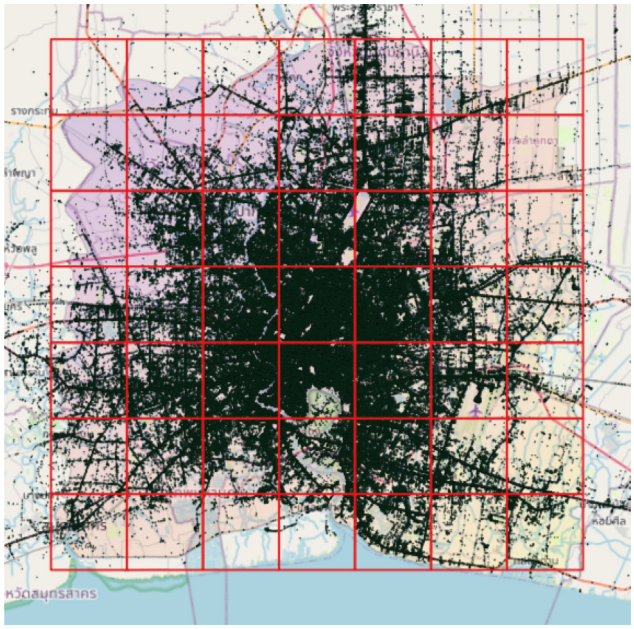

Fig. 3. Spatial grid segmentation and distribution of pickup demand across the study area.

3.1.3. Spatial grid division:

The study area was partitioned into a 7×7 grid, where each cell spans 10 km × 10 km. This grid-based segmentation allows localized aggregation of passenger demand and facilitates feature extraction compatible with the TabNet and ensemble modeling workflow. The 10 km × 10 km cell size was selected to balance operational relevance and statistical reliability: it corresponds approximately to a Bangkok district-level service area, which matches the granularity at which fleet redeployment and idle-vehicle reallocation decisions are made for autonomous taxi systems, while preserving sufficient observations per cell-hour to support stable training of all base learners.

3.1.4. Trip extraction

Pickup and drop-off points were identified by detecting transitions in meter status and analyzing GPS traces, ensuring precise capture of trip initiation and completion events.

3.1.5. Aggregation

Demand values were aggregated hourly within each grid cell, forming a structured spatiotemporal time series of passenger pickups used as the target variable.

3.1.6. Feature normalization

All continuous variables (e.g., temperature, humidity, wind speed, and precipitation) were normalized to improve convergence and comparability across models.

Fig. 2 illustrates the defined study area used for all analyses. Each grid cell was used as a spatial reference for demand aggregation and weather integration, ensuring consistency between mobility and environmental features.

Each processed data point was mapped to its corresponding spatial grid cell. Fig. 3 presents the spatial distribution of aggregated pickup points, highlighting areas of high and low passenger demand intensity across Bangkok.

To capture daily temporal variation, hourly pickup counts were computed over a 24-hour cycle. Fig. 4 depicts the temporal demand trend, indicating morning and evening peak periods corresponding to commuting hours.

Table 2 summarizes the variables derived from both mobility and weather datasets. These include temporal, spatial, and environmental features, along with the target variable representing hourly passenger demand per grid cell.

3.2. Train, validation, and test protocol

The dataset was split sequentially according to chronological order, with the earliest 60% of observations used for training, the next 25% for validation, and the most recent 15% for testing. The three subsets serve strictly distinct roles within the proposed framework. The training set is used exclusively to fit the parameters of each base learner (TabNet, MLP, and SGB). The validation set is used exclusively for ensemble-level operations, namely the computation of the baseline normalisation constants MSE_0 and MAE_0 and the optimisation of the ensemble weight vector (w_t, w_m, w_s) via the trust-region constrained solver. The test set is held out entirely

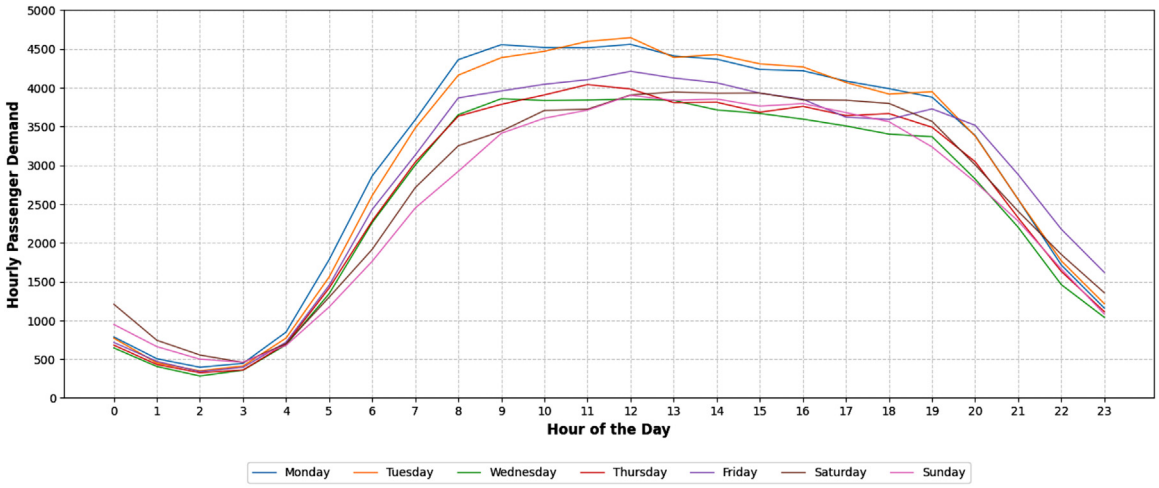

Fig. 4. Hourly passenger pickup trends reflecting typical daily demand variation.

Table 2
Summary of features derived from integrated taxi GPS and weather data.

Feature	Description
day_name	Day of the week
grid_id	Unique identifier for spatial grid
time_slot	Hourly time slot (023)
temp	Temperature (×C)
dwpt	Dew point temperature (×C)
rhum	Relative humidity (%)
prcp	Precipitation (mm)
wdir	Wind direction (degrees)
wspd	Wind speed (m/s)
pres	Atmospheric pressure (hPa)
coco	Encoded weather condition (e.g., clear, cloudy)
demand	Number of taxi pickups per grid cell and hour

and is not accessed during base learner training, weight optimisation, or hyperparameter selection. Once the optimiser converges, the resulting weights are fixed and applied as a static linear combination to the test set predictions. This protocol follows the canonical stacking ensemble convention (Breiman, 1996; Wolpert, 1992) and ensures that all reported test set metrics constitute an unbiased estimate of out-of-sample generalisation.

4. Framework development

This section details the selection, configuration, and mathematical formulation of the three individual models integrated in the proposed hybrid ensemble framework: TabNet, SGB, and a MLP. These models were selected for their complementary strengths in capturing complex, nonlinear, and spatiotemporal relationships within urban passenger demand data.

4.1. Selection of models

The proposed ensemble framework couples the strengths of TabNet, SGB, and MLP. These models are selected to complement each other in terms of their inductive biases, which have proven very successful for structured data. Specifically, TabNet offers attentive, step-wise feature selection and hierarchical representation learning that is optimized for tabular inputs, which enables the model framework to focus on the most relevant temporal, spatial, and environmental attributes (Arik and Pfister, 2021). SGB provides strong performance for structured data by gradient boosting on residuals to sequentially fix residual errors and reduce bias, and improve generalization performance (Friedman, 2001). It sequentially corrects the error of prior models and combines weak learners to form a strong predictive model (Zhang et al., 2022). The MLP is a flexible universal function approximator that captures nonlinear interactions between features by layered representations and backpropagation (Graupe, 2013; Haykin, 1999). The proposed ensemble integrates these complementary learners to exploit attention-driven feature sparsity, residual correction, and deep nonlinear mapping within a unified prediction framework. While recurrent (LSTM, GRU) and graph-based (ST-GCN, Graph WaveNet, DCRNN) models could in principle be incorporated as additional base learners, they are designed for problems with continuous high-resolution temporal sequences and well-defined spatial adjacency. The present formulation operates on a coarse 7×7 grid with hourly

aggregation, where graph convolution offers limited marginal benefit due to the small number of spatial nodes and the heterogeneous topology of Bangkok's road network. The proposed framework is therefore deliberately built upon strong tabular learners; extending it to sequence- and graph-based base models is left as future work.

4.2. Individual model configuration and formulation

4.2.1. TabNet

TabNet is a deep learning architecture specifically designed for tabular data that employs a sequence of decision steps to perform instance-wise feature selection through sparse attention mechanisms (Arik and Pfister, 2021). Unlike conventional feed-forward networks that process all input features simultaneously, TabNet learns to focus selectively on the most informative subset of features at each decision step, enabling both interpretability and efficiency. This makes it particularly suitable for heterogeneous spatiotemporal datasets where temporal, spatial, and environmental features interact in complex and nonlinear ways.

At each decision step i, an *attentive transformer* generates a sparse mask M_i over the input features, while a *feature transformer* processes the masked input to extract higher-level representations. Denoting the input matrix by $X \in \mathbb{R}^{n\times d}$, the hidden state from the previous step by h_{i-1}, and the prior scale by P_{i-1}, the computations are as follows:

$$M_i = \text{Sparsemax}\big(P_{i-1} \odot (h_{i-1} W_a)\big), \tag{1}$$

$$P_i = P_{i-1} \odot (\gamma - M_i), \tag{2}$$

$$h_i = \sigma\big(\text{BN}\big((X \odot M_i) W_f\big)\big), \tag{3}$$

where $M_i \in \mathbb{R}^{n\times d}$ is the attention mask, $\odot$ denotes the Hadamard (elementwise) product, W_a and W_f are learnable weight matrices, $\text{BN}(\cdot)$ is batch normalization, and $\sigma(\cdot)$ is a nonlinear activation (e.g., ReLU). The parameter γ controls feature reuse between steps, ensuring that each decision focuses on complementary aspects of the input space.

After N_d decision steps, the model aggregates the learned representations to produce the final prediction:

$$\hat{y} = \sum_{i=1}^{N_d} h_i W_y, \tag{4}$$

where W_y denotes the projection matrix mapping the decision representations to the output space.

To enforce sparsity and improve interpretability, TabNet applies an additional regularization term to the objective function:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_{\text{sparse}} \sum_{i=1}^{N_d} \sum_{j=1}^{d} -M_{i,j} \log M_{i,j}, \tag{5}$$

where $\mathcal{L}_{\text{task}}$ is the task-specific loss (here, regression loss), λ_{sparse} controls the strength of the sparsity penalty, and the second term penalizes entropy in the attention masks to encourage compact feature selection.

In this study, TabNet was configured as the first base learner within the ensemble framework. The model used three decision steps with compact decision and attention blocks (n_d=8, n_a=8) and a feature reuse coefficient of γ=1.3. The sparsity regularization weight was set to $\lambda_{\text{sparse}}=10^{-4}$ to balance interpretability and performance. Model optimization was performed using the AdamW algorithm with a learning rate of 10^{-3} and a weight decay coefficient of 10^{-4}. This configuration provided stable convergence and interpretable feature selection across spatial, temporal, and weather-related predictors, aligning with recent advances in tabular deep learning (Gorishniy et al., 2021; Wang et al., 2022).

4.2.2. Stochastic gradient boosting (SGB)

The SGB model in this framework applies least-squares boosting, an effective regression technique that reduces bias and enhances predictive accuracy through stage-wise additive modeling. It constructs a sequence of weak learners, typically shallow decision trees, each trained to correct the residual errors of the previous iteration. This iterative refinement enables the model to progressively capture complex nonlinear relationships among features.

The optimization procedure minimizes the loss function using gradient descent (Bentéjac et al., 2021), following the additive formulation introduced by Friedman (Friedman, 2001). The objective function for SGB is expressed as:

$$L(\theta) = \sum_{i=1}^{n} l\big(y_i, F_{m-1}(x_i) + \alpha h_m(x_i)\big), \tag{6}$$

where l denotes the loss function, y_i represents the true target values, $F_{m-1}(x_i)$ is the ensemble prediction at iteration $m-1$, $h_m(x_i)$ is the newly added weak learner, and α is the learning rate controlling update magnitude. The model is updated iteratively as:

$$F_m(x) = F_{m-1}(x) + \alpha h_m(x), \tag{7}$$

where each new learner $h_m(x)$ fits the negative gradient of the loss with respect to the current predictions. This process continues until the specified number of boosting rounds is completed, effectively minimizing residual errors and improving generalization.

In this study, SGB was implemented using the *GradientBoostingRegressor* in `scikit-learn` with $n_{\text{estimators}}$=500, α=0.1, and max_depth=15. A fixed random seed (42) ensured reproducibility. This configuration provided a good balance between bias control and model expressiveness, yielding stable convergence and consistent performance across spatial, temporal, and meteorological predictors, consistent with prior studies on gradient boosting robustness (Bentéjac et al., 2021; Davis et al., 2020; Zhang et al., 2022).

4.2.3. Multilayer perceptron (MLP)

The MLP model employed in this study serves as the deep learning component within the ensemble framework. MLPs are fully connected feedforward neural networks capable of capturing complex nonlinear mappings between features and target variables (Goodfellow et al., 2016; LeCun et al., 2015).

For an input vector $x \in \mathbb{R}^d$, the forward propagation through layer l is expressed as:

$$z^{(l)} = W^{(l)} h^{(l-1)} + b^{(l)}, \qquad h^{(l)} = f\left(z^{(l)}\right), \tag{8}$$

where $W^{(l)}$ and $b^{(l)}$ denote the weights and biases of layer l, $f(\cdot)$ is the activation function (ReLU in this case), and $h^{(0)} = x$. The network output is given by:

$$\hat{y} = W^{(L)} h^{(L-1)} + b^{(L)}. \tag{9}$$

The training objective minimizes the MSE between predicted and true values:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{10}$$

During backpropagation, the gradients of the loss with respect to the parameters are computed as:

$$\frac{\partial \mathcal{L}}{\partial W^{(l)}} = \delta^{(l)} (h^{(l-1)})^\top, \qquad \frac{\partial \mathcal{L}}{\partial b^{(l)}} = \delta^{(l)}, \tag{11}$$

where the layer error term $\delta^{(l)}$ is propagated backward using:

$$\delta^{(l)} = \left(W^{(l+1)}\right)^\top \delta^{(l+1)} \odot f'\left(z^{(l)}\right). \tag{12}$$

For ReLU activation, $f'(z) = 1$ if $z > 0$, otherwise 0. Parameter updates are performed using the Adam optimizer:

$$W^{(l)} \leftarrow W^{(l)} - \eta \frac{\hat{m}^{(l)}}{\sqrt{\hat{v}^{(l)}} + \epsilon}, \tag{13}$$

where η is the learning rate, and $\hat{m}$, $\hat{v}$ are bias-corrected first and second moment estimates (Kingma and Ba, 2015).

In this study, the MLP was implemented in PyTorch as a four-layer feedforward network with hidden dimensions of 64, 32, and 16 neurons, each activated by ReLU. The network was trained for 500 epochs using the Adam optimizer (η=10^{-3}) and MSE loss, with a batch size of 512. This configuration achieved smooth convergence and effectively captured nonlinear dependencies across spatial, temporal, and meteorological predictors, consistent with prior studies on deep feedforward networks for structured regression tasks (Bishop, 2006; Hornik, 1989).

4.3. Hybrid ensemble formulation and optimization

The three base learners (TabNet, MLP, and SGB) were integrated through a trust-constrained (Byrd et al., 1999; Conn et al., 2000) multi-objective hybrid ensemble framework designed to enhance framework robustness and generalization. This ensemble approach leverages the complementary strengths of each base learner. Fig. 1 illustrates the overall workflow, where framework outputs from individual learners are combined and optimized during the validation phase before final inference.

4.3.1. Weighted ensemble structure

Let $\hat{y}_t$, $\hat{y}_m$, and $\hat{y}_s$ denote the predicted outputs from TabNet, MLP, and SGB, respectively. The ensemble prediction for the i-th sample is defined as a convex combination of these outputs:

$$\hat{y}^{(i)}_{\text{ens}} = w_t \hat{y}^{(i)}_t + w_m \hat{y}^{(i)}_m + w_s \hat{y}^{(i)}_s, \tag{14}$$

subject to the constraints:

$$w_t \geq 0, \quad w_m \geq 0, \quad w_s \geq 0, \quad w_t + w_m + w_s = 1. \tag{15}$$

The non-negativity constraint ensures interpretable and proportional weight allocation among the base models, while the summation constraint enforces convexity. Base learners remain fixed after training; only the ensemble weights (w_t, w_m, w_s) are optimized during the validation phase.

4.3.2. Robust loss integration

The Huber loss (Huber, 1964) is included to improve resilience against outliers by combining the quadratic sensitivity of MSE for small errors with the linear robustness of MAE for large deviations:

$$\text{Huber} = \begin{cases} \frac{1}{2}(y_i - \hat{y}_i)^2, & \text{if } |y_i - \hat{y}_i| \leq \delta, \\ \delta\left(|y_i - \hat{y}_i| - \frac{1}{2}\delta\right), & \text{otherwise}, \end{cases} \tag{16}$$

where δ defines the transition threshold. This loss function provides stable gradient propagation during optimization and mitigates the effect of outlier-induced bias in the ensemble learning process.

4.3.3. Hybrid objective function

The optimization of ensemble weights (w_t, w_m, w_s) was guided by a hybrid objective function that simultaneously considers multiple learning criteria to achieve a trade-off between accuracy, robustness, and stability. The function integrates normalized mean and absolute errors, a loss term (Huber), an exponential trust component related to R^2, and a regularization penalty controlling weight dispersion. It is defined as:

$$\min_{w_t, w_m, w_s} L = \alpha_1 \frac{\text{MSE}}{\text{MSE}_0} + \alpha_2 \left(\frac{\text{MAE}}{\text{MAE}_0} \right)^2 + \alpha_3 \frac{\text{Huber}}{\text{MAE}_0} - \beta_1 e^{-\gamma(1-R^2)} + \lambda \, \text{Std}(w_t, w_m, w_s), \tag{17}$$

where MSE and MAE denote the mean squared and mean absolute errors on the validation dataset, each normalized by their respective baseline values MSE_0 and MAE_0 obtained before optimization. The Huber term represents the robust Huber loss (Huber, 1964), defined in Eq. (16), which combines the advantages of MAE and MSE by maintaining smooth gradients for small residuals while reducing the influence of outliers. The coefficient of determination R^2 acts as a reward term, encouraging higher explanatory power of the ensemble model. Finally, the $\text{Std}(w_t, w_m, w_s)$ term denotes the standard deviation among the ensemble weights, functioning as a trust-regularization component that prevents any single learner from dominating the ensemble and promotes balanced contributions among TabNet, MLP, and SGB. The baseline values MSE_0 and MAE_0 are computed once, prior to optimisation, as the average of the validation-set errors produced by the three independently trained base learners (TabNet, MLP, and SGB). They serve as fixed scaling constants that bring the heterogeneous loss components to a comparable order of magnitude and remain unchanged throughout the optimisation procedure. Consequently, no adaptive information from individual validation samples is leaked into the optimisation beyond what is already implied by the standard role of validation data in supervised learning (Breiman, 1996; Wolpert, 1992).

The hyperparameters α_1, α_2, and α_3 control the relative importance of squared, absolute, and robust loss components, while β_1 and γ modulate the exponential trust factor that penalizes poor model fit when R^2 is low. The final term, scaled by λ, enforces smoothness in the learned weight distribution, thereby avoiding oscillatory convergence or degenerate solutions. This composite design enables simultaneous optimization of predictive fidelity and model reliability. The inclusion of both loss normalization and trust-region regularization ensures that the ensemble learns stable, interpretable weights while minimizing sensitivity to noise and outlier distortions during training.

4.3.4. Trust-region constrained solver

To solve Eq. (17), a trust-region constrained optimization algorithm was employed (Byrd et al., 1999). This method iteratively updates the ensemble weight vector within a dynamically defined neighborhood, known as the trust region that restricts the magnitude of parameter changes, thereby ensuring stable convergence under nonlinear and bound-constrained conditions (Conn et al., 2000).

At iteration k, the algorithm constructs a local quadratic approximation of the objective function (Byrd et al., 1999; Conn et al., 2000):

$$m_k(p) = f_k + g_k^T p + \frac{1}{2} p^T B_k p, \tag{18}$$

where f_k is the current objective value, g_k is the gradient vector, B_k represents the Hessian approximation, and p is the trial step constrained by $\|p\| \leq \Delta_k$, with Δ_k denoting the trust-region radius. The step p_k is accepted when the actual reduction sufficiently matches the predicted reduction; otherwise, the trust-region radius is reduced to refine the local approximation.

The solver was initialized with uniform weights $[w_t, w_m, w_s] = [1/3, 1/3, 1/3]$ and bounded within $[0.1, 0.1, 0.1] \leq [w_t, w_m, w_s] \leq [0.9, 0.9, 0.9]$. Convergence was achieved when both the gradient norm ($\|\nabla f_k\| < 10^{-6}$) and relative step size ($\|p_k\| < 10^{-5}$) satisfied termination criteria, yielding optimized ensemble weights for final framework integration.

4.3.5. Ensemble workflow and functioning

The ensemble optimization was performed exclusively during the validation phase. After each base learner produced its validation predictions $(\hat{y}_{t,\text{val}}, \hat{y}_{m,\text{val}}, \hat{y}_{s,\text{val}})$, these outputs were linearly combined according to Eq. (14). The trust-region constrained solver then iteratively adjusted the ensemble weights (w_t, w_m, w_s) to minimize the hybrid multi-objective loss in Eq. (17) while satisfying bound and normalization constraints.

Once the optimization converged, the resulting weights (w_t^*, w_m^*, w_s^*) were applied to generate ensemble predictions on unseen data, whereas the base model parameters remained unchanged. This decoupled optimization ensured that the ensemble captured cross-model complementarities without altering the internal representations of individual learners. Throughout this procedure, the test set is held strictly separate. It is not used during base learner training, validation-based weight optimisation, or any intermediate selection step. The optimised weights (w_t^*, w_m^*, w_s^*) are therefore fixed before being applied to the test set predictions, ensuring that all reported test-set metrics provide an unbiased estimate of out-of-sample generalisation.

By explicitly constraining the optimization within the trust-region bounds and incorporating multi-objective regularization, the proposed framework achieved stable weight convergence. The resulting hybrid ensemble operates as a meta-learner that dynamically integrates heterogeneous learning paradigms attention-based feature selection, gradient boosting, and deep nonlinear mapping within a unified and reproducible architecture.

4.4. Performance metrics

To assess the predictive performance of the proposed ensemble and its baseline models, three standard regression metrics were employed: MAE, RMSE, and R^2. These metrics jointly capture framework accuracy, error dispersion, and variance explanation capability, providing a balanced evaluation of predictive quality.

The MAE quantifies the average magnitude of absolute deviations between predicted and observed values:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{19}$$

where y_i and $\hat{y}_i$ denote the actual and predicted values, respectively, and n is the number of observations. MAE offers an intuitive measure of typical prediction errors and assigns equal weight to all deviations.

The RMSE measures the square root of the mean squared error:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{20}$$

which penalizes large deviations more strongly due to the squaring operation, making it more sensitive to outliers and suitable for evaluating models where large errors are undesirable.

The MSE measures the average of the squared deviations between predicted and observed values:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{21}$$

where y_i and $\hat{y}_i$ denote the actual and predicted values, respectively. MSE penalizes larger errors more heavily due to the squaring operation and is commonly used to assess overall prediction accuracy and model variance.

The coefficient of determination (R^2) indicates the proportion of variance in the observed data explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{22}$$

where $\bar{y}$ is the mean of the observed target values. A higher R^2 value reflects a stronger framework fit and greater explanatory power.

Together, these complementary metrics provide a rigorous basis for quantifying predictive accuracy, assessing residual distribution, and comparing the performance of individual learners and the optimized ensemble (Chicco et al., 2021).

4.5. Theoretical contribution vs. engineering contribution

To clarify the methodological positioning of the present framework relative to conventional weighted ensembles, AutoML pipelines, and standard hyperparameter optimisation, the theoretical and engineering aspects of the contribution are summarised separately below.

Theoretical contribution. The methodological novelty of the present work lies in the formulation of a constrained multi-objective ensemble weight optimisation that integrates four design elements within a single objective in Eq. (17): normalised MSE for sensitivity to large deviations, squared normalised MAE for absolute-error reduction, Huber loss for robustness against outliers, and an exponential R^2 reward term that penalises poor explanatory power, with an additional weight-dispersion regulariser to discourage degenerate solutions. These components are aggregated into a single objective and minimised under simplex and bound constraints by a trust-region solver, providing a principled mechanism for adaptive ensemble weight selection that simultaneously accounts for accuracy, error dispersion, and robustness within a unified formulation.

Engineering contribution. Beyond its theoretical novelty, the framework provides an end-to-end engineering pipeline for spatiotemporal autonomous taxi demand prediction. This includes the integration of three complementary base learners with distinct inductive biases (TabNet for attention-driven feature selection, MLP for deep nonlinear mapping, and SGB for residual-based boosting), the combination of large-scale GPS-based mobility records with synchronised meteorological variables on a grid-based spatial structure, and a deployment-oriented workflow suitable for real-time fleet allocation in autonomous taxi systems.

Differentiation from related approaches. The proposed framework differs from related methodologies in three principal ways:

- *Conventional weighted ensembles* typically employ equal, fixed, or empirically tuned weights and optimise a single error metric. The proposed framework instead derives weights through a multi-objective trust-region optimisation that jointly accounts for accuracy, robustness, and explanatory power.
- *AutoML pipelines* automate model selection and hyperparameter tuning across large search spaces. The present framework operates at the ensemble-combination layer with a fixed set of deliberately chosen complementary learners, preserving interpretability and reproducibility without large-scale automated search.
- *Hyperparameter optimisation methods* tune base-learner parameters to minimise validation loss. The proposed framework leaves base-learner parameters fixed after independent training and optimises only the ensemble combination weights, thereby decoupling base-learner expressiveness from ensemble-level robustness.

This positioning makes explicit the methodological scope of the present study and clarifies its contribution relative to existing optimisation and ensemble-learning paradigms.

5. Results and discussion

5.1. Simulation setup

In this study, the proposed trust-constrained multi-objective hybrid ensemble framework is simulated and tested using Python 3.11 on a computing environment with NVIDIA RTX 3050 GPU acceleration. Each base learner was trained separately on the same preprocessed data and then combined using the proposed ensemble optimization procedure. Evaluation metrics of both the base learners and the ensemble variants include MAE, RMSE, and R^2, as defined in Eqs. (19)(21). These three metrics jointly capture prediction accuracy, error dispersion, and variance explanation, forming a comprehensive evaluation framework.

5.2. Overall performance

Table 3 presents a comparison of all models on the validation and testing sets. The proposed optimised ensemble consistently achieved the strongest predictive performance, yielding the lowest MAE of 5.97, the lowest RMSE of 9.69, and the highest R^2 of 0.924 on the testing set, with comparable values on the validation set (MAE 6.01, RMSE 9.82, R^2 0.920). Among the individual base learners, the deep MLP delivered the best stand-alone result on the testing set (MAE 6.62, RMSE 10.27, R^2 0.915), followed by the SGB (MAE 6.45, RMSE 10.82, R^2 0.905), with TabNet showing the largest residuals (MAE 8.00, RMSE 12.40, R^2 0.876). The equal-weight ensemble already improved upon the strongest base learner (MAE 6.19, RMSE 9.96, R^2 0.918), confirming that simple averaging exploits complementary error patterns; however, the optimised ensemble obtained through the trust-region multi-objective procedure provided a further consistent reduction across all error dimensions.

Relative to the strongest base learner (MLP), the optimised ensemble reduced RMSE by approximately 5.7% (from 10.27 to 9.69) and MAE by approximately 9.8% (from 6.62 to 5.97) on the testing set, while R^2 increased from 0.915 to 0.924, corresponding to an absolute gain of 0.009 (relative improvement of about 1.0%). These improvements highlight the benefit of the trust-constrained multi-objective optimisation, which adaptively balances bias and variance contributions from each base learner. Consequently, the ensemble captures the complementary strengths of TabNet's attention-driven feature sparsity, MLP's nonlinear feature mapping, and SGB's residual error correction, producing a robust and interpretable predictive framework.

Fig. 5 illustrates the residual and absolute error distributions. The optimized ensemble exhibits the narrowest and most symmetric residual spread around zero, indicating minimal systematic bias. The absolute error boxplots further confirm the ensemble's tighter interquartile range and reduced variance, highlighting improved consistency and robustness.

5.3. Optimised ensemble weights and stability analysis

The trust-region constrained optimiser converged to the following weight vector on the validation set: $w_t^* = 0.1000$, $w_m^* = 0.4270$, and $w_s^* = 0.4730$, with optimality measure 8.20×10^{-9} and constraint violation 0.00. The SGB receives the largest share of the ensemble prediction, followed by MLP, while TabNet acts as a smaller corrective component at the lower bound of the feasible region. The boundary location of w_t reflects the fact that, with a sufficiently strong gradient-boosted learner already capturing most residual structure, attention-based feature selection contributes complementary diversity rather than dominant predictive volume.

To verify that the optimised weights are not artefacts of a single validation realisation, a bootstrap resampling experiment was conducted. Ten independent resamples of the validation set were drawn, and the trust-region optimisation was rerun on each resample with identical bounds, constraints, and initial conditions $w_0 = (1/3, 1/3, 1/3)$. The resulting mean and standard deviation of each weight component, together with the joint deep contribution $(w_t + w_m)$, are reported in Table 4.

Table 3
Comparative performance of all models on validation and testing sets.

Model	Validation Set				Testing Set			
	MSE	RMSE	MAE	R^2	MSE	RMSE	MAE	R^2
TabNet	156.20	12.50	8.02	0.871	153.64	12.40	8.00	0.876
Deep MLP	109.60	10.47	6.69	0.909	105.53	10.27	6.62	0.915
SGB	117.43	10.84	6.44	0.903	117.16	10.82	6.45	0.905
Ensemble (Equal)	101.87	10.09	6.24	0.916	99.19	9.96	6.19	0.918
Ensemble (Optimized)	**96.41**	**9.82**	**6.01**	**0.920**	**93.92**	**9.69**	**5.97**	**0.924**

Table 4
Optimised ensemble weights and stability across 10 bootstrap resamples of the validation set.

Component	Optimised Value	Bootstrap Mean	Bootstrap Std.
w_t (TabNet)	0.1000	0.1897	0.0742
w_m (MLP)	0.4270	0.3792	0.0285
w_s (SGB)	0.4730	0.4311	0.0486
$w_t + w_m$ (joint deep contribution)	0.5270	0.5689	0.0486

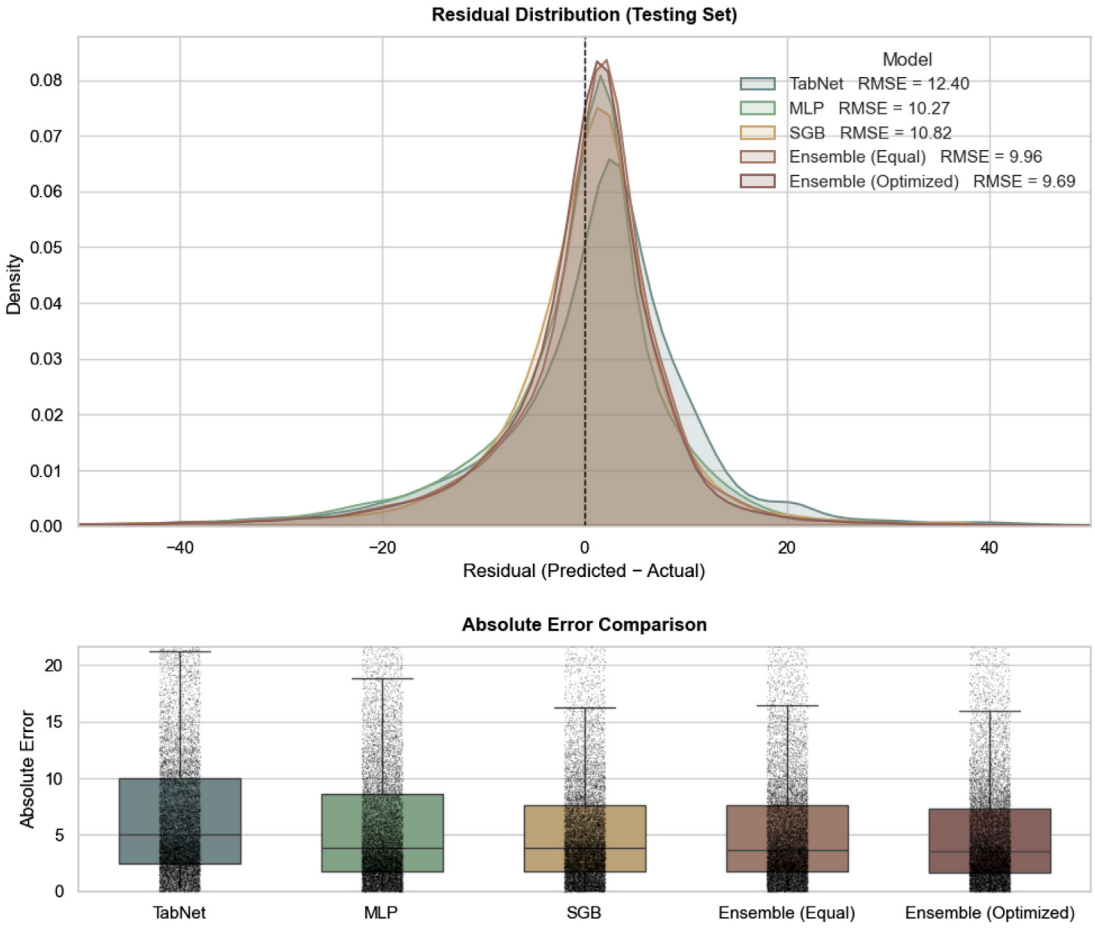

Fig. 5. Residual distribution (top) and absolute error comparison (bottom) across models on the testing set.

The MLP weight is highly stable across resamples (coefficient of variation $\approx 7.5\%$), and the SGB weight is also stable (coefficient of variation $\approx 11.3\%$). The TabNet weight exhibits somewhat higher relative variability, which is methodologically expected because TabNet and MLP are both deep learners and partially overlap in capturing nonlinear feature interactions; the optimiser can therefore redistribute weight between them across resamples without altering the overall deep contribution. A more representative stability indicator is the combined deep contribution ($w_t + w_m$), which inherits the variability of w_s through the unit-sum constraint and therefore exhibits a standard deviation of 0.0486 (coefficient of variation $\approx 8.5\%$). This confirms that the macro-structure of the ensemble (deep block versus boosting block) is highly reproducible across validation realisations, even when individual deep weights shift between resamples.

The interpretability of the weight distribution is consistent with the underlying error structure of the base learners. The SGB refines residual local patterns through stage-wise additive modelling and contributes the largest share. The MLP captures global nonlinear dependencies among temporal and meteorological features. TabNet, while contributing a smaller absolute weight, brings hierarchical attention-based feature reasoning that increases ensemble diversity and supports robustness under shifting demand conditions. To verify that the framework is not over-tuned to a specific choice of coefficients in the hybrid objective of Eq. (17), a sensitivity analysis was performed by sweeping the principal coefficients across a grid of 81 combinations covering values around the nominal setting. The resulting test-set RMSE varies only between 9.68 and 9.88, MAE between 5.97 and 6.13, and R^2 between 0.921 and 0.924, while the optimised weight ordering ($w_s > w_m > w_t$) is preserved across all combinations. The narrow performance range and the stable weight structure confirm that the framework is robust to coefficient choice and that the reported performance is a stable property of the hybrid objective formulation rather than the result of fine-tuning.

5.4. Statistical significance of ensemble improvements

To rigorously confirm that the predictive gains of the proposed optimised ensemble over the base learners and the equal-weight baseline are not artefacts of a specific test partition, the Diebold-Mariano (DM) test (Diebold and Mariano, 1995) with the Harvey small-sample correction was applied to the paired absolute errors on the held-out test set ($n \approx 16{,}000$). The DM test is the standard

Table 5
Statistical significance of test-set MAE improvements achieved by the optimised ensemble over each baseline (Diebold-Mariano test with Harvey small-sample correction).

Comparison	Mean MAE reduction	% reduction	DM statistic	**DM p-value**
Optimised vs TabNet	2.0312	25.39	−42.40	$< 1 \times 10^{-16}$
Optimised vs MLP	0.6508	9.83	−21.67	$< 1 \times 10^{-16}$
Optimised vs SGB	0.4802	7.45	−15.14	$< 1 \times 10^{-16}$
Optimised vs Equal-weight Ensemble	0.2236	3.61	−15.34	$< 1 \times 10^{-16}$

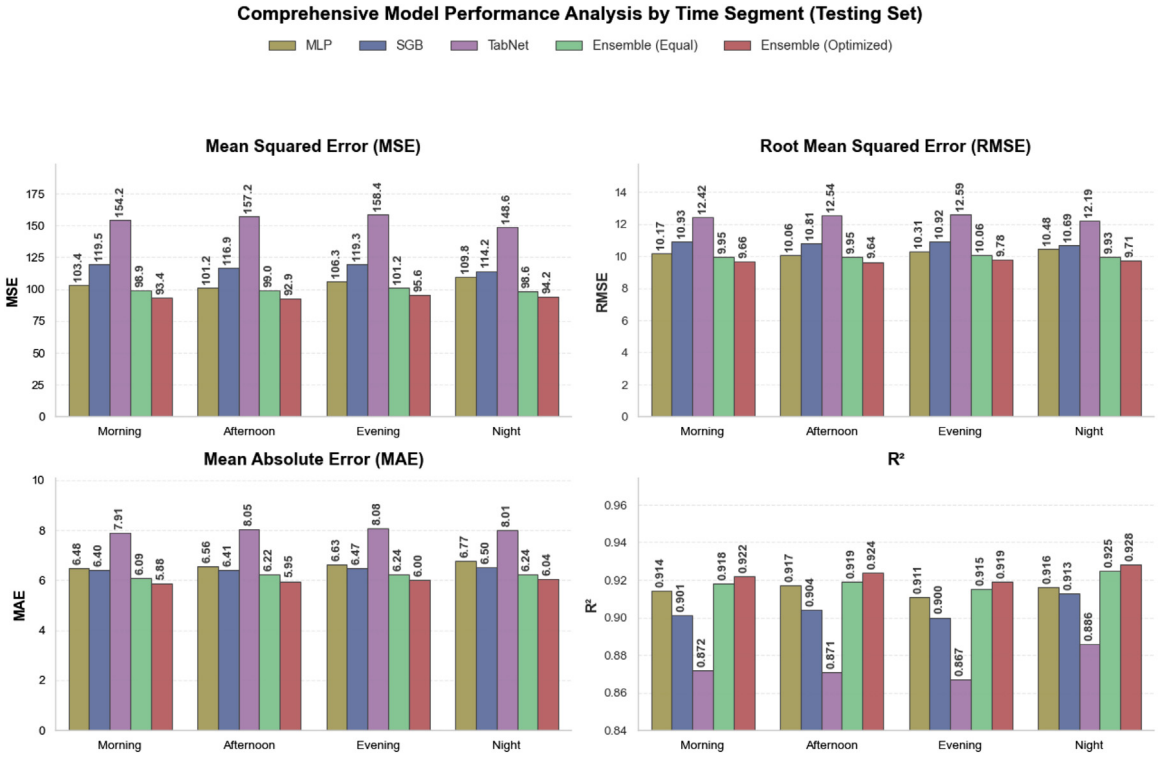

Fig. 6. Comprehensive performance comparison across time segments on the testing set.

procedure for assessing the statistical significance of differences in forecast accuracy and is widely used in time-series and forecasting research. Negative DM statistics indicate that the optimised ensemble produces lower absolute errors than the corresponding baseline.

The results are summarised in Table 5. All DM statistics are strongly negative, indicating that the optimised ensemble consistently produces lower absolute errors than every baseline, and all p-values are several orders of magnitude below any conventional significance threshold ($\alpha = 0.05$ or even $\alpha = 0.001$). The largest gains are observed against TabNet (25.39% MAE reduction) and MLP (9.83%), while even the smallest improvement, namely the 3.61% reduction over the equal-weight ensemble, remains highly statistically significant. These results confirm that the improvements attributable to the trust-region multi-objective optimisation are statistically robust rather than chance occurrences.

5.5. Temporal segment analysis

To evaluate temporal stability, framework performance was analysed across four daily segments: morning (06:00–11:59), afternoon (12:00–16:59), evening (17:00–21:59), and night (22:00–05:59). Fig. 6 presents the comparative MSE, RMSE, MAE, and R^2 values across these time windows. The optimised ensemble maintained stable accuracy throughout all segments, with the testing-set MAE varying within the narrow range of 5.88–6.04 and R^2 ranging between 0.919 and 0.928, outperforming every individual base learner and the equal-weight ensemble in each segment.

The proposed optimised ensemble consistently outperformed all baseline models across different time segments. Among the base learners, the SGB emerged as the strongest individual model in all four segments (testing-set MAE between 6.40 and 6.50), followed by the MLP (MAE between 6.48 and 6.77), with TabNet exhibiting the largest residuals (MAE between 7.91 and 8.08). Despite this variation in base-learner strength, the optimised ensemble achieved a further reduction of approximately 0.4–0.5 MAE units in every segment relative to the best base learner, confirming that the trust-region multi-objective optimisation extracts complementary infor-

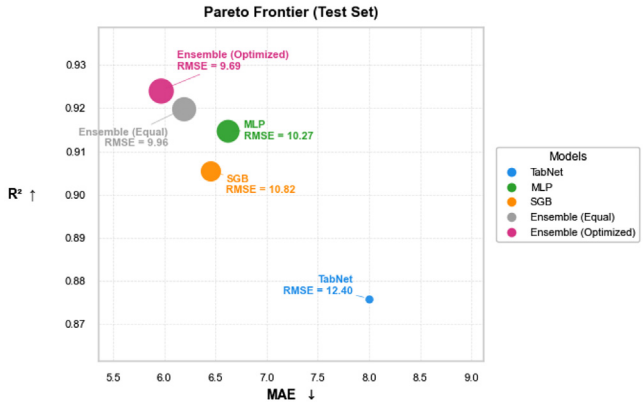

Fig. 7. Pareto frontier illustrating trade-offs between MAE and R^2 for all models (testing set).

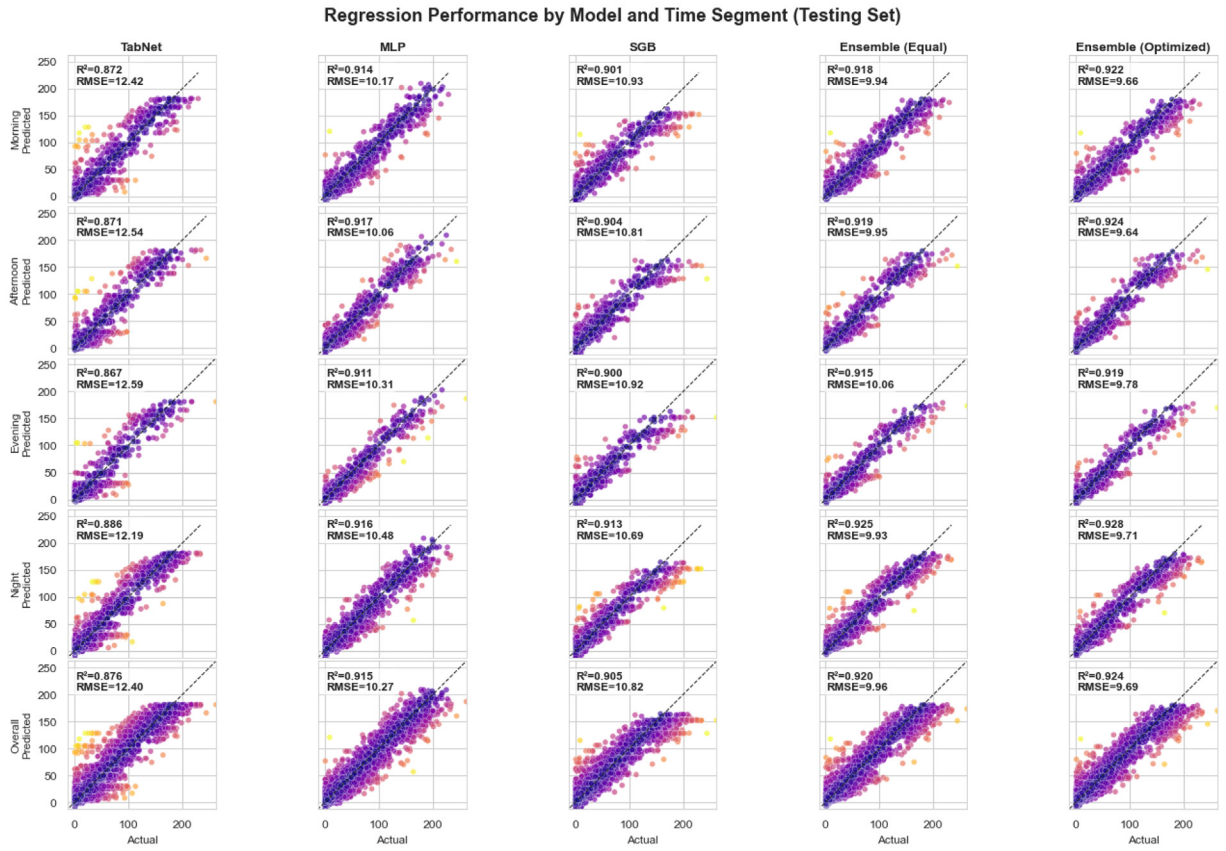

Fig. 8. Regression performance of all models across daily time segments (testing set).

mation from each model rather than simply favouring the dominant one. The framework demonstrated particular robustness during high-volatility periods, namely the morning and evening peaks, when fluctuations in passenger demand and weather conditions are most pronounced. This resilience emphasises the model's ability to adapt to changing demand dynamics while effectively integrating the complementary learning behaviours of its individual base learners.

5.6. Pareto and stability analysis

The trade-off between predictive error and explanatory power was further examined through a Pareto frontier of MAE (minimization) versus R^2 (maximization), as shown in Fig. 7. The optimized ensemble occupies the top-left region of the frontier, indicating its dominance over all other models in multi-objective space.

5.7. Regression fit and predictive alignment

Fig. 8 shows scatterplots of actual versus predicted demand values for each model and time segment. The optimized ensemble demonstrates the highest alignment with the diagonal reference line, confirming its superior fit and reduced heteroscedasticity. This alignment is particularly prominent in high-demand intervals, where other models tend to underestimate passenger pickups.

The observed compact clustering and reduced residual spread validate the hybrid optimization strategy's effectiveness in improving predictive coherence among the base learners. This demonstrates that the trust-constrained multi-objective formulation not only enhances overall accuracy but also ensures temporal consistency and interpretability.

5.8. Discussion

The findings validate that hybrid ensemble learning captures the nonlinear and multi-factorial dynamics of passenger demand in ATs systems. The MLP and SGB models provide localized pattern recognition and gradient-based residual correction, while TabNet brings in feature-level interpretability through attention-driven sparsity. The proposed trust-constrained optimizer then harmonizes these complementary traits to achieve a well-balanced ensemble that minimizes bias and variance. The results are theoretically based on the principles of ensemble generalization. An ensemble of models with diverse error patterns helps to reduce overall prediction error through compensatory interactions. The optimized ensemble achieves a balance by adjusting weights, ultimately finding a Pareto-efficient solution that balances mean absolute error (MAE) and R^2. A systematic sensitivity analysis on spatial grid resolution, examining the trade-off between spatial granularity, demand sparsity, and predictive accuracy across multiple cell sizes, is identified as a valuable extension of the present study.

In practical terms, the proposed framework shows applicability in real-world ITS and smart mobility planning. With forecasts that accurately capture demand fluctuations between time of day, this allows for effective fleet allocation, reduces idle times, and improves reliability in passenger service. Besides, the stability observed across the temporal segments has confirmed that the model is ready to be integrated into dynamic, real-time dispatch systems supporting autonomous mobility networks.

6. Conclusion

This study proposed a trust-constrained hybrid ensemble learning framework that integrates TabNet, SGB, and MLP for passenger demand forecasting in ATs systems. The ensemble was further refined through a multi-objective optimization strategy where the contribution of each base learner was adaptively determined by leveraging complementary strengths of attention-based tabular reasoning, residual error correction, and nonlinear representation learning.

Simulations performed on different validation and testing datasets demonstrate that the proposed optimized ensemble has outperformed the individual models as well as the baseline equal-weight ensembles. The optimized framework achieved the lowest MAE of 5.97, lowest RMSE of 9.69, and the highest R^2 of 0.924, which reflects a balanced improvement in accuracy and generalization capability. Temporal analysis further revealed that while the SGB emerged as the strongest individual base learner across all time segments, the proposed ensemble framework consistently delivered superior predictive accuracy and stability throughout all time segments, particularly in the morning and evening peaks when passenger behavior and weather vary largely. Residual and absolute error analyses confirmed its narrower error distribution and higher robustness, establishing the ensemble as the Pareto-optimal solution across all evaluated performance dimensions.

These results emphasize that the ensemble generalizes well under nonstationary conditions in an urban environment and can be a reliable tool for ITS and real-time fleet management. Accurate short-term demand forecasts from the framework may enable dynamic resource allocation and reduce passenger waiting times and inefficient vehicle utilization; it is one of the key enablers of efficient and sustainable smart mobility ecosystems.

Future research might focus on enriching the framework by considering contextual factors in real time-such as traffic congestion and social event data-to enhance responsiveness in fluctuating urban conditions. Integrating environmental objectives, such as minimization of emissions or energy use, would better align forecasting frameworks with green and sustainable transportation policies. Besides that, extending ensembles to federated or distributed learning environments could provide a path to scalable, privacy-preserving implementations across many smart cities for the next generation of adaptive, low-carbon mobility intelligence. In addition, while the present study focuses on tabular ensemble learning, future research will extend the framework to incorporate deep spatiotemporal base learners such as LSTM, GRU, ST-GCN, DCRNN, and Graph WaveNet. Such extensions will become particularly valuable when finer-grained spatial graphs and longer continuous trajectory sequences are available, allowing the trust-region multi-objective optimisation layer proposed here to be applied across a richer family of heterogeneous learners. A further valuable extension lies in moving from the present point-forecasting setting toward probabilistic and quantile forecasting, where dedicated models such as quantile regression or distributional ensembles can be coupled with the trust-region multi-objective optimisation layer to produce calibrated demand intervals and to support evaluation through additional metrics such as pinball loss in their native setting.

Acknowledgment

This study was supported by Thammasat University Research Fund, Contract No. TUFT 031/2568. The authors gratefully acknowledge this support.

Declaration of Generative AI and AI-Assisted Technologies

During the preparation of this manuscript, the authors used ChatGPT (OpenAI) for language editing and clarity improvement. The authors reviewed and take full responsibility for the final content.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Mongkut Piantanakulchai reports financial support was provided by Thammasat University Research Fund. Mongkut Piantanakulchai reports a relationship with Thammasat University that includes: funding grants. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Mongkut Piantanakulchai: Writing – review & editing, Validation, Supervision, Resources, Methodology, Investigation, Conceptualization. **Adeel Munawar:** Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization.

References

Aldakkhelallah, A., Simic, M., 2021. Autonomous vehicles in intelligent transportation systems. In: Human Centred Intelligent Systems: Proceedings of KES-HCIS 2021 Conference. Springer, pp. 185–198.

Almihat, M.G.M., Kahn, M., Aboalez, K., Almaktoof, A.M., 2022. Energy and sustainable development in smart cities: an overview. Smart Cities 5 (4), 1389–1408.

Arik, S., Pfister, T., 2021. Tabnet: attentive interpretable tabular learning. Proc. AAAI Conf. Artif. Intell. 35 (8), 6679–6687.

Bentéjac, C., Csörgő, A., Martínez-Muñoz, G., 2021. A comparative analysis of gradient boosting algorithms. Artif. Intell. Rev. 54, 1937–1967.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer, New York.

Breiman, L., 1996. Stacked regressions. Mach. Learn. 24 (1), 49–64.

Byrd, R.H., Hribar, M.E., Nocedal, J., 1999. An interior point algorithm for large-scale nonlinear programming. SIAM J. Optim. 9 (4), 877–900.

Campisi, T., Severino, A., Al-Rashid, M.A., Pau, G., 2021. The development of the smart cities in the connected and autonomous vehicles (CAVs) era: from mobility patterns to scaling in cities. Infrastructures 6 (7), 100.

Carson-Bell, D., Adadevoh-Beckley, M., Kaitoo, K., et al., 2021. Demand prediction of ride-hailing pick-up location using ensemble learning methods. J. Transp. Technol. 11 (02), 250.

Chen, D., Zhang, Y., Gao, L., Geng, N., Li, X., 2017. The impact of rainfall on the temporal and spatial distribution of taxi passengers. Plos One 12 (9), e0183574.

Chicco, D., Warrens, M.J., Jurman, G., 2021. The coefficient of determination r-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerjComput. Sci. 7, e623.

Conn, A.R., Gould, N.I.M., Toint, P.L., 2000. Trust Region Methods. SIAM.

Davis, J., Devos, L., Reyners, S., Schoutens, W., 2020. Gradient boosting for quantitative finance. J. Comput. Finance 24 (4), 1–40.

Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. J. Bus. Econ. Stat. 13 (3), 253–263.

Faisal, A., Kamruzzaman, M., Yigitcanlar, T., Currie, G., 2019. Understanding autonomous vehicles. J. Transp. Land Use 12 (1), 45–72.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29 (5), 1189–1232.

Friedman, J.H., 2002. Stochastic gradient boosting. ComputStat. Data Anal. 38 (4), 367–378.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press.

Gorishniy, Y., Rubachev, I., Khrulkov, V., Babenko, A., 2021. Revisiting deep learning models for tabular data. In: Advances in Neural Information Processing Systems, Vol. 34, pp. 18932–18943.

Graupe, D., 2013. Principles of Artificial Neural Networks, Vol. 7. World Scientific.

Haykin, S., 1999. Neural Networks: A Comprehensive Foundation, 2nd edition Prentice Hall, Upper Saddle River, NJ.

Hornik, K., 1989. Multilayer feedforward networks are universal approximators. Neural Netw. 2 (5), 359–366. doi:10.1016/0893-6080(89)90020-8.

Huber, P.J., 1964. Robust estimation of a location parameter. Ann. Math. Stat. 35 (1), 73–101. doi:10.1214/aoms/1177703732.

iTIC Foundation, 2021. The Intelligent Traffic Information Center Foundation (iTIC). https://itic.longdo.com/opendata/. Accessed: October 2025.

Kingma, D.P., Ba, J., 2015. Adam: a method for stochastic optimization. In: International Conference on Learning Representations (ICLR).

Lai, Y., Zhang, K., Lin, J., Yang, F., Fan, Y., 2019. Taxi demand prediction with LSTM-based combination model. In: 2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom). IEEE, pp. 944–950.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444. doi:10.1038/nature14539.

Lee, E.-K., Gerla, M., Pau, G., Lee, U., Lim, J.-H., 2016. Internet of vehicles: from intelligent grid to autonomous cars and vehicular fogs. Int. J. Distrib. Sens. Netw. 12 (9), 1550147716665500.

Li, Y., Yu, R., Shahabi, C., Liu, Y., 2018. Diffusion convolutional recurrent neural network: data-driven traffic forecasting. In: International Conference on Learning Representations (ICLR).

Meteostat, 2024. Meteostat: Free historical weather data. https://meteostat.net. [Online; accessed 2024].

Munawar, A., Piantanakulchai, M., 2025a. Centralized agent-based model for efficient dispatching of autonomous taxis in smart cities. Int. J. Transp. Sci. Technol. doi:10.1016/j.ijtst.2025.07.001.

Munawar, A., Piantanakulchai, M., 2025b. Explainable AI to enhance demand forecasting model for autonomous taxis in smart cities. In: 2025 11th International Conference on Engineering, Applied Sciences, and Technology (ICEAST). IEEE, pp. 77–80.

Munawar, A., Piantanakulchai, M., 2025c. Machine learning-driven passenger demand forecasting for autonomous taxi transportation systems in smart cities. Expert Syst. 42 (3), e70014.

Qin, G., Li, T., Yu, B., Wang, Y., Huang, Z., Sun, J., 2017. Mining factors affecting taxi drivers' incomes using GPS trajectories. Transp. Res. C: Emerg. Technol. 79, 103–118.

Rajak, S., Baruah, U., 2020. An ensemble model for predicting passenger demand using taxi data set. In: Machine Learning, Image Processing, Network Security and Data Sciences: Second International Conference, MIND 2020, Silchar, India, July 30–31, 2020, Proceedings, Part II 2. Springer, pp. 336–346.

Sarkar, A., Sonbhadra, S.K., Agarwal, S., 2019. Predictive analytics using ensemble models. In: 2019 IEEE Students Conference on Engineering and Systems (SCES). IEEE, pp. 1–5.

Schaller Consulting, 2004. The New York City Taxicab Fact Book. Schaller Consulting, Brooklyn, NY.
Sonbhadra, S.K., Agarwal, S., Syafrullah, M., Adiyarta, K., 2020. An application of ensemble and deep learning models in predictive analytics. In: 2020 IEEE 7th International Conference on Industrial Engineering and Applications (ICIEA). IEEE, pp. 574–582.
Wang, S., Sun, Z., Lin, K., Zhang, J., Zhou, M., 2022. TabTransformer: tabular data modeling using contextual embeddings. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, pp. 9247–9255.
Wolpert, D.H., 1992. Stacked Generalization, Vol. 5. Elsevier, pp. 241–259.
Wu, H., Levinson, D., 2022. Ensemble models of for-hire vehicle trips. Front. Future Transp. 3, 876880.
Wu, Z., Pan, S., Long, G., Jiang, J., Zhang, C., 2019. Graph wavenet for deep spatial-temporal graph modeling. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI), pp. 1907–1913.
Ye, R., Xu, Z., Pang, J., 2022. DDFM: a novel perspective on urban travel demand forecasting based on the ensemble empirical mode decomposition and deep learning. In: Proceedings of the 5th International Conference on Big Data Technologies, pp. 373–379.
Yu, B., Yin, H., Zhu, Z., 2018. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI), pp. 3634–3640.
Zhang, X., Zhao, X., 2021. A clustering-aided ensemble method for predicting ridesourcing demand in chicago. arXiv preprint arXiv:2109.03433.
Zhang, Z., Zhu, X., Liu, D., 2022. Model of gradient boosting random forest prediction. In: 2022 IEEE International Conference on Networking, Sensing and Control (ICNSC). IEEE, pp. 1–6.