ARTICLE | Article in Press

Explainable multi-modal transformer framework for electric vehicle charging demand forecasting

Md Mushfiq Us Salehin

Received: 11 July 2026 / Accepted: 28 October 2026
© The Author(s) 2026

Abstract

Accurate prediction of EV charging demand is fundamental to grid stability, real-time pricing, and charging station management, but few prediction methods are able to utilize more than just historical demand, and are able to reveal why a particular forecast was made. In this paper we develop and experimentally investigate an explainable multi-modal forecasting approach to hourly EV charging demand, which fuses charging session data, weather information, grid prices, and grid operating information (voltage, frequency, transformer load) with station occupancy and traffic signal status. Using 15,304 charging session records from 55 charging stations between January 2019 and July 2020, we construct a 13,680-hour multi-modal data panel including calendar features, lags, and rolling statistics. We evaluate five forecasting architectures, including linear regression, Random Forest, XGBoost, a Long Short-Term Memory network, and our proposed attention-based Transformer network with a learned temporal pooling mechanism to interpret attention. XGBoost and Random Forest are the two strongest models overall and slightly outperform each other on different metrics: XGBoost achieves the lowest RMSE and highest R^2 (RMSE = 5.63 kWh, R^2 = 0.980), while Random Forest achieves a marginally lower MAE and MAPE (MAE = 1.67 kWh, MAPE = 7.39%, versus XGBoost's MAE = 1.72 kWh, MAPE = 8.36%); a paired significance test shows this MAE difference is not statistically significant, so the two tree ensembles are best described as practically comparable, with XGBoost's edge concentrated in the largest-error hours. Both substantially outperform our attention-based Transformer network (RMSE = 27.98 kWh, R^2 = 0.513) and LSTM network (RMSE = 30.02 kWh, R^2 = 0.440) at current data scale; in contrast, the attention weights from the transformer provide an intuitive explanation on which past hours are the most influential ones in forecasting the current charging demand. SHAP value analysis of the best model identifies the recent past charging demand, transformer loading, and the hour of the day as the most significant predictors, and ablation tests confirm a 4–6% increase in RMSE from the removal of grid and calendar data, whereas a negligible impact is observed from the removal of station occupancy/traffic features. All differences in forecasting performance across various architectures are confirmed by Diebold-Mariano tests to be statistically significant, and a sensitivity analysis to model hyperparameters reveals that the Transformer's performance is largely independent of the exact model parameter count over a 6.5x range, indicating data volume rather than model complexity is currently the bottleneck. Further analysis of forecasting errors demonstrates that both the best tree-based model and the Transformer fail at the same sudden morning weekday charging demand surges. An alternative variant of the Transformer network using modality-attention fusion (learning weights for each input

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.

https://doi.org/10.1038/s41598-026-78819-1

Published online: 13 November 2026

scientific reports

feature category in a data-dependent way rather than simple concatenation) slightly but significantly reduces the mean absolute error (but not the RMSE) compared to the non-modality-attention approach and offers learned modality weights that confirm the relative importance ordering from the ablation study. We conclude that for aggregated single-station charging demand forecasting, tree-based models with SHAP are both the most accurate and explainable, whereas attention-based models provide a promising approach for integrated feature explanation with greater scalability potential as multi-year and multi-station data become available. We discuss the implications for smart-grid operation and a road map for widespread spatial deployment of our model.

Keywords EV charging demand forecasting · Multi-modal data fusion · Transformer · Explainable AI · SHAP · Smart grid · XGBoost

Introduction

The global transition toward electric mobility is placing new and highly variable loads on distribution networks: global electric car sales exceeded 20 million in 2025, accounting for roughly one-quarter of all new car sales worldwide[1]. As charging infrastructure expands, grid operators, charge-point operators, and aggregators increasingly need accurate short-term forecasts of charging demand to schedule generation, manage transformer loading, price electricity dynamically, and plan station capacity[2,3]. Forecasting this demand is difficult because it is shaped by a heterogeneous mix of factors: driver behaviour and vehicle characteristics, ambient weather, real-time electricity prices, grid operating conditions, and station occupancy, all of which interact over time.

Most existing EV demand-forecasting studies use historical charging load as the sole predictor and apply either classical statistical models[4,5] or recurrent neural networks such as LSTMs and GRUs[6,7]. Transformer-based architectures[8] have recently shown strong performance on long-horizon time-series problems in other domains[9] and have begun to be applied to EV charging demand specifically[10,11], but their use together with heterogeneous exogenous data, and with an explicit account of model interpretability, remains comparatively underexplored. Deep models are also frequently criticized as opaque[12], which limits their adoption by grid operators who must justify operational decisions, motivating explainability tools such as SHAP[13] alongside forecasting accuracy.

Despite this progress, the literature leaves three specific questions open rather than a single broad gap. First, prior multi-modal EV forecasting studies typically fuse heterogeneous exogenous streams through fixed, flat concatenation[14,15], so the relative contribution of each modality is frozen once training ends rather than allowed to vary with conditions; it remains unclear whether an input-dependent, dynamically weighted fusion mechanism yields materially different or more informative modality attributions than this static approach. Second, existing Transformer applications to EV charging demand[10,11] report accuracy against limited baseline sets, so it is largely untested whether attention-based architectures retain any advantage once benchmarked against strong, well-tuned tree ensembles at the moderate data scales typical of single-site deployments. Third, explainability in this literature is usually applied post hoc to a single best-performing model, so whether an architecture-native explanation channel, such as a learned fusion gate or temporal-attention weights, can independently corroborate post-hoc attributions rather than merely restating them remains an open question. The present study is designed to close these three specific gaps rather than to introduce another multi-modal pipeline in the abstract: Sect. Adaptive modality-attention fusion (proposed enhancement) introduces the adaptive modality-attention fusion mechanism and Sect. Adaptive modality-attention fusion versus flat concatenation tests it directly against flat concatenation to address the first; Sect. Results benchmarks the proposed Transformer against Linear Regression, Random Forest, and XGBoost under an identical protocol to address the second; and Sect. Explainability analysis develops SHAP-based and attention-based explanations and cross-checks the fusion module's learned weights against an independently derived ablation study (Sect. Modality-attention weights: corroborating the ablation study) to address the third.

Building directly on these three gaps, this paper constructs and evaluates, on real enriched charging-session data, a multi-modal forecasting pipeline that (i) fuses charging, weather, grid, pricing, and occupancy information into a unified hourly feature representation; (ii) benchmarks a purpose-built attention-based Transformer against classical machine-learning and deep-learning baselines; and (iii) applies SHAP-based and attention-based explainability to identify which modalities and time steps drive predictions. Unlike prior proposals that only outline such a framework conceptually, every number reported here including cases where the proposed Transformer does not outperform simpler baselines is obtained from actually training and evaluating the models on the dataset.

Relying on historical demand alone implicitly treats charging load as a self-contained autoregressive process, an assumption that breaks down whenever the drivers of that load shift away from their historical pattern: a heatwave, a price change, or a transformer approaching its rated capacity can move demand in ways that lagged demand values alone cannot anticipate. Ambient temperature affects both vehicle range and drivers' propensity to precondition a cabin before departure, so demand can rise on hot or cold days independent of any trend visible in the load series itself (Fig. 7). Electricity price shapes when drivers choose to plug in under time-of-use tariffs, so a price step can shift load between hours even when the daily total is unchanged (Fig. 8). Grid operating parameters, especially transformer loading, are not merely correlated with demand but are partly its consequence and partly a constraint on it: as loading approaches rated capacity, protective curtailment and charging-speed throttling can suppress the very demand a purely historical model would otherwise predict[16,17]. These are three structurally different causal roles an exogenous driver of demand (weather), a price-responsive shifter of demand (pricing), and a demand-dependent physical constraint (grid) and no single historical-demand feature can stand in for all three, which is why the modality groups in Table 3 are treated as distinct inputs rather than folded into a single 'other predictors' bucket.

A second question is why these modalities should be fused through a learned, modality-aware mechanism rather than simply concatenated into one flat feature vector, as the baseline Transformer in Sect. Proposed transformer forecaster does, an issue also raised by attention-based multi-modal fusion frameworks in adjacent forecasting domains[18] and by graph- and traffic-fused EV forecasting pipelines that combine several exogenous data streams[15]. Flat concatenation forces a single shared projection to learn, implicitly and without supervision, how much weight each modality deserves, and that weight is fixed once training ends it cannot adapt to the fact that, say, grid loading is highly informative during a demand surge but nearly constant, and therefore uninformative, overnight. A fusion mechanism that first embeds each modality group separately and then learns an input-dependent gate over those embeddings (Sect. Adaptive modality-attention fusion (proposed enhancement), Fig. 9) makes this weighting explicit, auditable, and free to shift hour by hour, which is also why its output doubles as an interpretability signal (Sect. Modality-attention weights: corroborating the ablation study) rather than only an accuracy mechanism.

Contributions

- A reproducible multi-modal feature-engineering pipeline that aligns charging-session records with hourly weather, pricing, grid, and occupancy data across six modality groups (Table 3).
- An adaptive modality-attention fusion module that embeds each modality group separately and learns an input-dependent softmax gate over those embeddings at every hour, replacing flat feature concatenation with an explicit, auditable fusion step (Sect. Adaptive modality-attention fusion (proposed enhancement)) whose weights can be inspected directly rather than inferred post hoc.
- An attention-based Transformer forecaster combining this fusion module with a learned temporal-pooling layer, so that both which modalities and which past hours drive a given forecast are directly readable from the trained model.
- A head-to-head empirical comparison of five core architectures (Linear Regression, Random Forest, XGBoost, LSTM, Transformer) plus a controlled ablation of the fusion mechanism itself (Sect. Adaptive

modality-attention fusion versus flat concatenation), all under identical features and evaluation protocol, with Diebold-Mariano and paired significance testing throughout.
- SHAP-based global and local explanations, an illustrative single-hour case study (Sect. Illustrative case study), and a leave-one-modality-out ablation study quantifying the marginal contribution of each modality group.
- An honest discussion of when gradient-boosted trees are preferable to deep sequence models for this class of problem, what the fusion module does and does not buy over flat concatenation, and what would be needed to close the remaining gap.

Related work

EV charging demand forecasting

Early EV charging-load forecasting relied on statistical time-series models, such as ARIMA-family and seasonal decomposition approaches applied to aggregated station load[4,5], and classical machine learning including linear regression and random forests[19], which remain strong baselines for tabular, lag-feature-rich problems. Subsequent work incorporated richer behavioural and contextual signals: Moon et al., and related studies modeled consumer charging patterns directly, while more recent hybrid approaches combine gradient-boosted trees with temporal convolutional networks[20] or fuse traffic and weather information into the forecasting pipeline[14]. Deep sequence models, particularly LSTM- and GRU-based architectures, have since become the default choice for capturing temporal dependencies without hand-crafted lag features, and transfer-learning variants have been proposed for settings with limited site-level data[21,22]. A recent systematic review and meta-analysis catalogues this rapidly growing body of machine-learning and deep-learning approaches to EV charging-behaviour prediction[23], and a complementary systematized literature review synthesizes forecasting horizons, data modalities, and model families across this fast-moving area[24]. Beyond purely temporal formulations, Kuang et al.[25] embed physical distribution-network constraints directly into a graph-learning model for citywide demand and dynamic pricing, and Qu et al.[26] recast charging-demand prediction as a natural-language-processing problem using large language models, illustrating the breadth of representational choices now being explored. Behavioural drift itself has also become a modeling target: You et al.[27] use federated meta-learning to track how charging behaviour adapted across the postpandemic transition, underscoring that demand patterns are non-stationary at the population level. More recent work continues to expand this space: Sanami et al.[7] combine an attention-augmented LSTM with SHAP-based explainability for day-ahead station-level demand, closely paralleling the dual accuracy-plus-interpretability goal pursued here; Kyriakopoulos and Theodoridis[11] report a cross-city experimental comparison in which Transformer and LSTM/GRU architectures outperform ARIMA and XGBoost at hourly-to-daily, regional forecasting horizons, a finding that contrasts with the present study's single-site result and is revisited in Sect. Discussion; Guo et al.[17] combine quadratic decomposition with a dual-attention mechanism for day-ahead demand prediction; and Tupayachi et al.[15] fuse real-world traffic, weather, and proprietary infrastructure data within a spatio-temporal graph convolutional network, illustrating multi-modal fusion at a network rather than single-site scale. Li,[28] further contribute UrbanEV, an open benchmark dataset for urban EV charging-demand prediction, reflecting a broader push toward standardized, reproducible evaluation in this literature.

Complementing these algorithmic contributions, a parallel strand of work grounds charging-demand formation directly in observed usage behaviour rather than in model architecture. Cui,[29] synthesize the literature on EV refuelling demand and parking patterns across forecasting, planning, and scheduling applications, underscoring that parking duration and refuelling-arrival patterns are themselves a first-order driver of demand that purely time-series formulations can under-represent. Cui et al.[30] compare real-world EV usage behaviour between 2018 and 2021 and show that behavioural parameters such as daily travel distance, all-electric range, and charging power shift materially as battery and charging technology improve, implying that a demand-forecasting model

trained on one period's behavioural regime may not transfer cleanly to another a caution directly relevant to the single 19-month window used in the present study. Zhang, Wang, Miller, Cui, Liu, and Zhang[31] predict charging demand in Beijing from large-scale real-world trip and charging records by first clustering users into behavioural categories and then simulating trip chains for each category, illustrating a behaviour-first alternative to the direct time-series regression approach adopted here. Together, these studies broaden the present discussion beyond model architecture to the question of how charging demand is behaviourally formed in the first place, and motivate treating behavioural non-stationarity (Sect. Robustness across weather conditions and time, Sect. Limitations) as a substantive concern rather than a purely statistical one.

Transformer-based time-series forecasting

Gradient-boosted tree ensembles, particularly XGBoost[32], extended classical tabular learning with regularized boosting that handles heterogeneous features efficiently, and have been applied directly to EV charging demand with strong reported accuracy. In parallel, the Transformer architecture[8] demonstrated that self-attention can model long-range dependencies more effectively than recurrence for many sequence problems, leading to a rapidly expanding family of time-series-specific variants: Informer[33] reduces the quadratic cost of self-attention for long sequences; Autoformer[34] and FEDformer[35] introduce decomposition and frequency-domain mechanisms to separate trend and seasonal components; Crossformer[36] models cross-variate dependencies explicitly; PatchTST[37] tokenizes series into patches for channel-independent attention; iTransformer[38] inverts the roles of the temporal and variate axes; and TimesNet[39] reshapes 1-D series into 2-D tensors to capture multi-periodicity. Within the EV domain specifically, Koohfar et al.[10] applied a Transformer-based deep-learning approach to charging-demand prediction, and graph-based spatio-temporal variants have targeted network-level, multi-station forecasting[40–42], including designs that fuse graph attention with autoregressive decomposition modules such as Autoformer[43] or extend the Autoformer architecture itself with frequency-enhancement and error-correction blocks for multi-step charging-load forecasting[44]. Hybrid convolution-attention designs have also been proposed for related day-ahead probabilistic load forecasting under limited data availability[22], a data-scarcity concern directly relevant to the single-site setting studied here. A parallel, simpler-model line of work has directly challenged the necessity of Transformer complexity for time-series forecasting: Zeng et al.[45] show that a single linear layer with seasonal-trend decomposition (DLinear) can match or exceed several Transformer variants on standard long-horizon benchmarks, and lightweight all-MLP alternatives such as TSMixer[46] pursue the same efficiency argument. These variants have mostly been validated on public long-horizon benchmarks rather than on EV charging data enriched with grid and pricing signals, which the present study addresses directly. Most recently, Shen et al.[9] propose a unified taxonomy disentangling attention mechanism, forecasting aggregation, forecasting paradigm, and normalization choices across encoder-only, encoder-decoder, and decoder-only Transformer designs, finding that bi-directional joint-attention with direct-mapping forecasting is consistently most effective — a design-space perspective complementary to the single fixed architecture evaluated here.

Multi-modal fusion and the tabular-vs-deep-learning debate

Fusing heterogeneous data modalities for time-series forecasting is an active area beyond the energy domain: recent frameworks combine numerical, visual, and even textual signals through attention-based gating mechanisms for forecasting tasks in finance and other sequential domains[18], illustrating fusion architectures conceptually similar to the modality-group fusion used here (Fig. 3), applied instead to EV charging, weather, grid, and pricing data. A separate and directly relevant line of work has questioned whether deep sequence models reliably outperform tree ensembles on tabular or tabular-like time-series data, especially at moderate dataset sizes: Shwartz-Ziv and Armon[12] argue that well-tuned gradient boosting remains competitive with, and often superior to, deep networks on tabular benchmarks. The present study's finding that XGBoost outperforms both an LSTM and a Transformer on an hourly, single-site aggregated EV dataset is consistent with this literature, and

motivates the ablation, sensitivity, and scale discussion in Sects. Results, Ablation study, and Discussion. A more recent, larger-scale benchmark spanning 111 structured datasets similarly finds that deep-learning models match or exceed gradient-boosted trees only on a characterizable subset of tasks, generally those with larger sample sizes or weaker tabular structure[47], reinforcing that the tree-ensemble advantage observed here is plausibly tied to the present dataset's scale rather than a general property of EV demand forecasting.

Explainable AI for energy and grid forecasting

For interpretability, SHAP (SHapley Additive exPlanations;[13] has become a standard tool for explaining tree-based and black-box model predictions through game-theoretic feature attributions, and has been applied widely to short-term load forecasting: examples include SHAP-based interpretation of XGBoost load forecasts[48] and explainable-AI tools evaluated on residential building load data[49,50], and a recent review catalogues the growing set of explainability and interpretability techniques applied specifically to electric load forecasting, noting that post-hoc, model-agnostic methods such as SHAP remain the dominant approach[51]. This paper uses SHAP as the primary explainability mechanism for the best-performing model. In parallel, attention weights internal to Transformer architectures have been used as a complementary, architecture-native explanation of which input time steps a model relies on; this paper uses both channels together rather than treating them as competing alternatives.

Grid impacts and smart charging

A substantial literature documents the operational stress that uncoordinated EV charging places on distribution infrastructure, including accelerated transformer aging under high-penetration smart-charging scenarios[2], voltage-stability degradation under fast charging[16], and broader power-quality and capacity concerns reviewed in Rahman et al.[3]. More recent work links forecasting directly to network planning and operation: Li and Han[52] incorporate distribution-network upgrade constraints into EV charging-station load prediction, Lokhande et al.[53] manage real-time demand deviations by exploiting EV charging flexibility, and Benavides et al.[54] show that coordinated response from EV charging stations can mitigate power fluctuations in grid-connected PV systems, together illustrating that accurate demand forecasts of the kind developed here are a prerequisite for these grid-side interventions. This literature motivates including grid operating parameters (voltage, frequency, transformer loading) as a first-class modality in the present framework rather than treating charging demand as a purely behavioural time series, and is directly reflected in the ablation finding (Sect. Ablation study) that grid features are among the most valuable non-charging modalities for forecast accuracy.

Dataset and feature engineering

Data description

The study uses an enriched EV charging dataset organized into four linked tables: individual charging sessions, station metadata, sub-session time-series telemetry, and daily weather observations. Table 1 summarizes the raw data; Table 2 defines the key variables used in this study.

Vehicles in the dataset span SUVs (41.3%), hatchbacks (24.1%), sedans (22.1%), and fleet vehicles including vans, shuttles, and trucks (9.7% combined), charged predominantly through Level-2 AC chargers (85.8%) with a smaller share of DC fast chargers (14.2%).

Table 1 Dataset description.

Component	Description	Size
Charging sessions	One row per completed charging event, incl. station, vehicle, energy delivered, price, grid and weather readings at session time	15,304 sessions
Stations	Station metadata: location, cluster/zone, charger type, rated power, number of ports, install year	55 stations
Sub-session time-series	High-resolution telemetry within a session (instantaneous power, SoC, voltage)	Per-session, ~20 steps each
Daily weather	Daily temperature, humidity, and rainfall observations	577 days (Jan 2019 - Jul 2020)

Table 2 Key variable definitions.

Variable	Description
kWhDelivered	Energy delivered during a session (kWh); aggregated to hourly total demand (target variable)
electricity_price_usd_per_kWh	Time-of-use electricity price at session time
transformer_loading_pct	Local distribution transformer loading (%) — grid stress indicator
grid_voltage_V / grid_frequency_Hz	Grid operating parameters recorded at charge time
station_occupancy_pct	Fraction of station ports occupied
traffic_congestion_index	Local traffic congestion proxy at session time
temperature_C / humidity_pct / rainfall_mm	Weather conditions (session-level and daily)
is_anomalous_session	Flag for sessions marked anomalous in the source data (3.37% of sessions)

Table 3 Feature engineering summary (34 features total).

Feature Group	Example Features	Count
Charging/session	num_sessions, avg_duration, avg_peak_power, anomalous_frac	5
Weather	w_temp, w_humidity, w_rainfall, temp_x_weekend	4
Grid	avg_grid_voltage, avg_grid_freq, avg_transformer_loading	3
Pricing	avg_price, price_change	2
Occupancy/traffic	avg_occupancy, avg_traffic	2
Temporal (calendar)	hour/dow/month (+sin/cos encodings), is_weekend	10
Lag/rolling history	lag_1, lag_2, lag_3, lag_24, lag_48, lag_168, roll_mean_24, roll_std_24, roll_mean_168	9

Multi-modal aggregation and feature engineering

Session-level records are aggregated to an hourly panel (13,680 hourly observations spanning 8 January 2019 to 30 July 2020) by summing energy delivered and averaging exogenous signals within each hour. Concretely, each charging session is attributed to the single hour containing its start timestamp, and the full energy delivered by that session is counted within that hour rather than apportioned across the hours it may actually span; a session that starts at 23:50 and continues past midnight, for instance, has all of its delivered energy counted in the 23:00 h rather than split between it and the following hour. This start-hour attribution is a simplification rather than an energy-metered hourly breakdown, and it introduces a small timing error for the minority of sessions that cross an hour boundary, an effect that is largest for the small number of long-duration overnight sessions and negligible for the many short daytime sessions. Aggregate exogenous signals (grid, weather, pricing, occupancy, and traffic) are separately averaged over all readings falling within each clock hour, independent of individual session boundaries. Hours with no sessions are retained with zero demand rather than dropped, so the model must also learn when demand is absent. Daily weather is broadcast to all hours of a day. The resulting feature set is organized into six modality groups, summarized in Table 3.

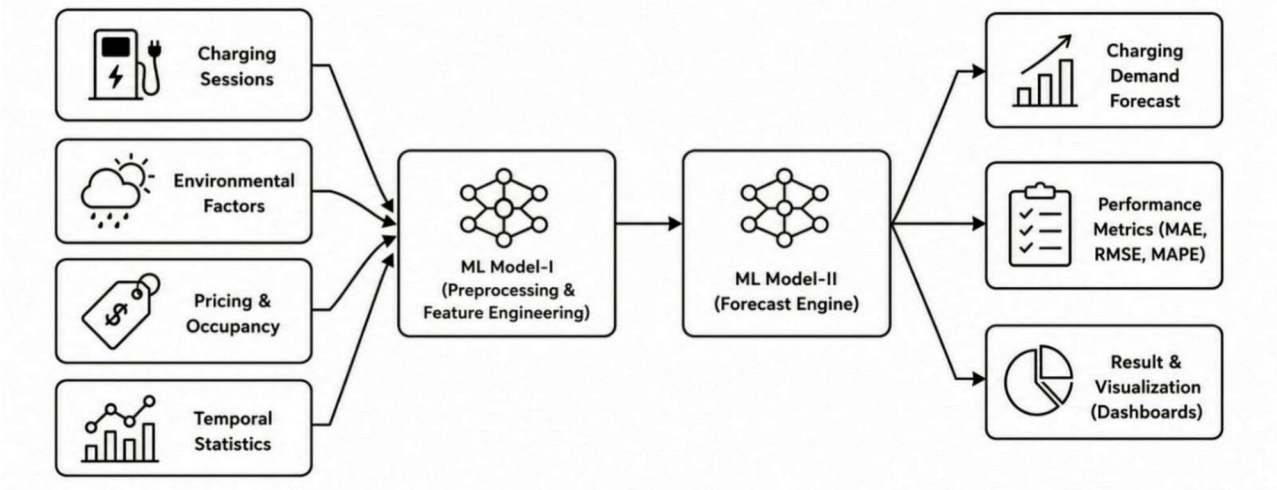


Fig. 1 Overall research framework, from raw multi-modal sources to explainable forecasts.

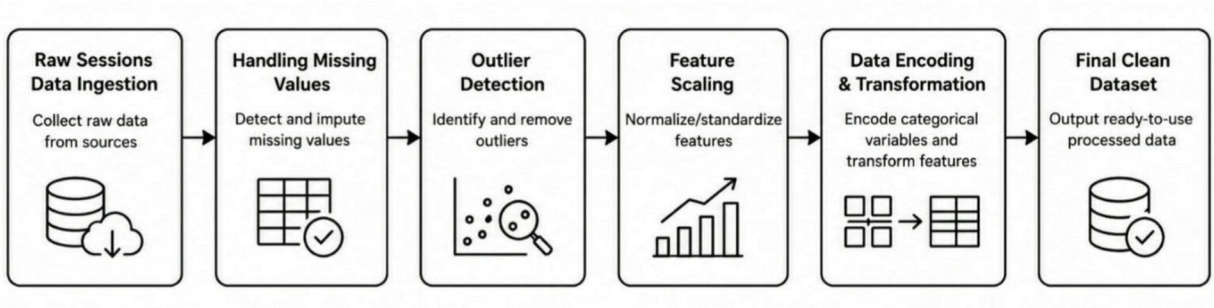


Fig. 2 Data preprocessing workflow.

Figures 1, 2 and 3 summarize, respectively, the overall research framework from raw multi-modal sources to explainable forecasts, the data preprocessing workflow, and the multi-modal feature fusion architecture used to construct the panel described below.

The correlation structure in Fig. 4 shows that grid stress (transformer loading) and session activity (num_sessions, avg_duration) correlate most strongly with demand, while weather variables correlate weakly and non-linearly with demand at the hourly level, motivating the use of non-linear models.

The network view in Fig. 5 makes the modality structure more explicit than the heatmap alone: occupancy and traffic form a tightly coupled pair ($r = 0.95$), both of which connect to grid loading and pricing, while weather variables form a largely separate cluster connected to the rest of the network only through temperature-humidity coupling ($r = -0.80$) and weak links to demand. This supports treating occupancy/traffic as a partially redundant modality (consistent with the ablation results in Sect. Ablation study) and weather as a more independent, if weaker, signal.

As Fig. 6 shows, hourly demand is strongly right-skewed and zero-inflated (mean = 23.0 kWh, std = 43.2 kWh, max = 392.1 kWh), reflecting many low-traffic overnight hours alongside occasional high-demand peaks (Figs. 7, 8, 9, 10).

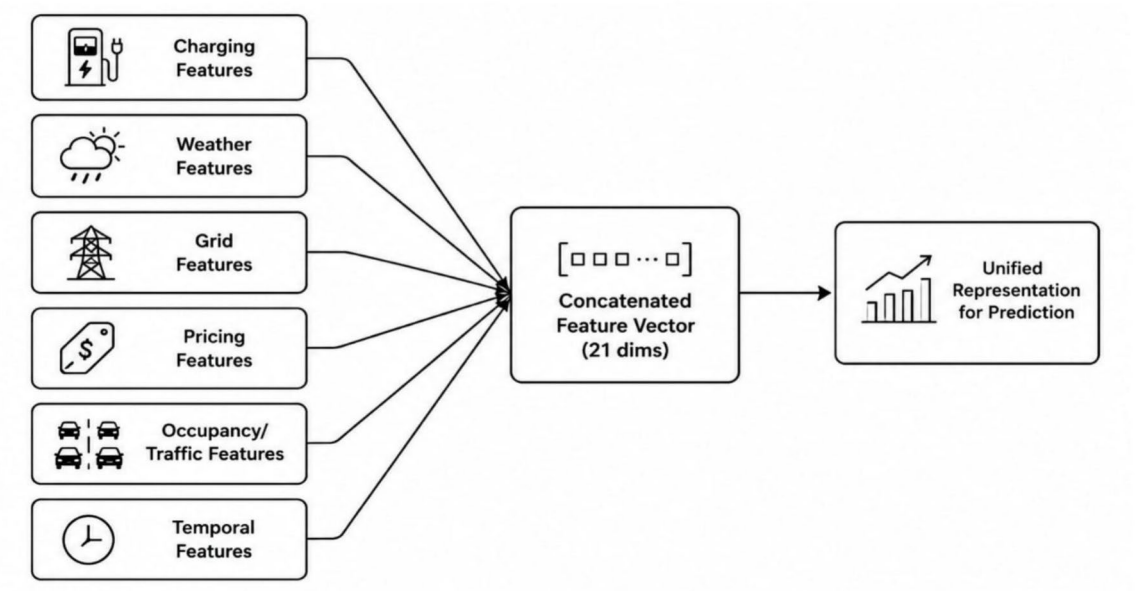


Fig. 3 Multi-modal feature fusion architecture.

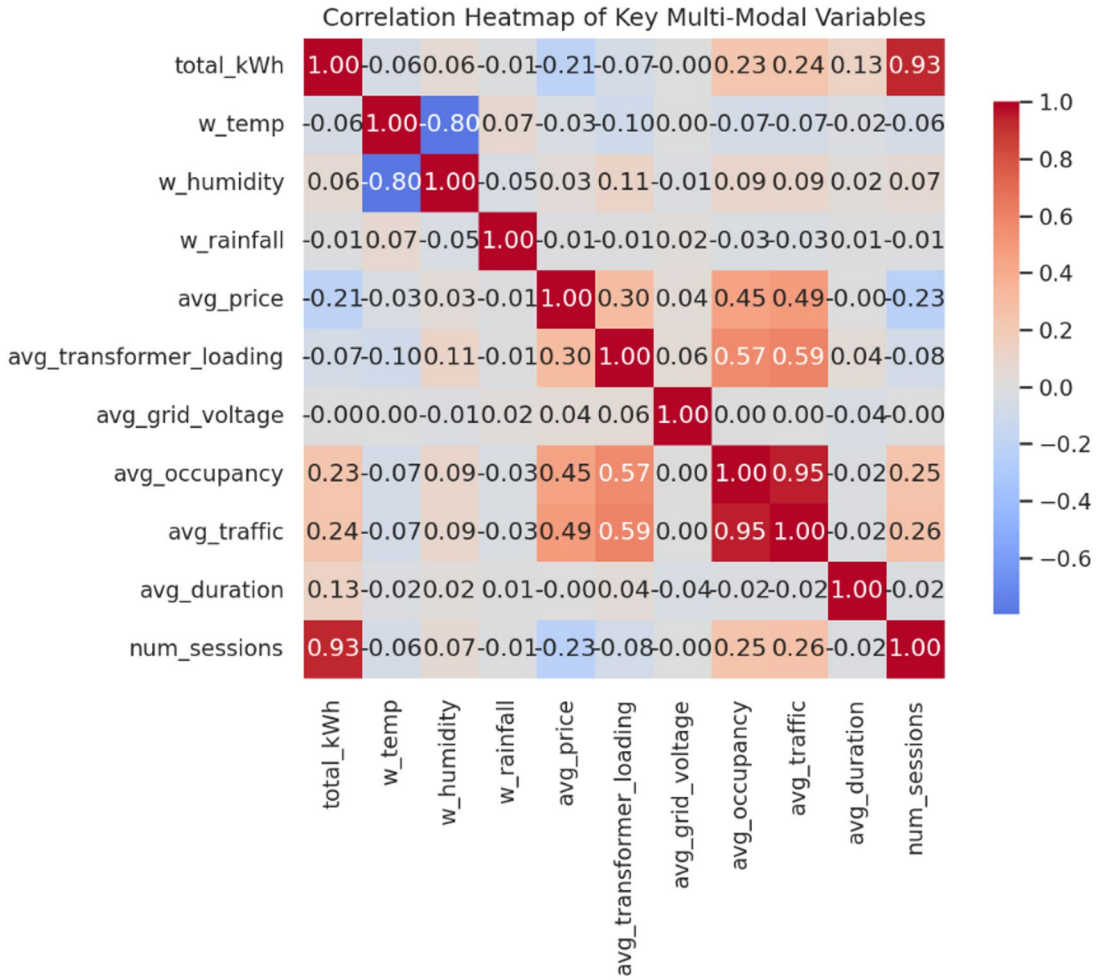


Fig. 4 Correlation heatmap of key multi-modal variables.

https://doi.org/10.1038/s41598-026-78819-1

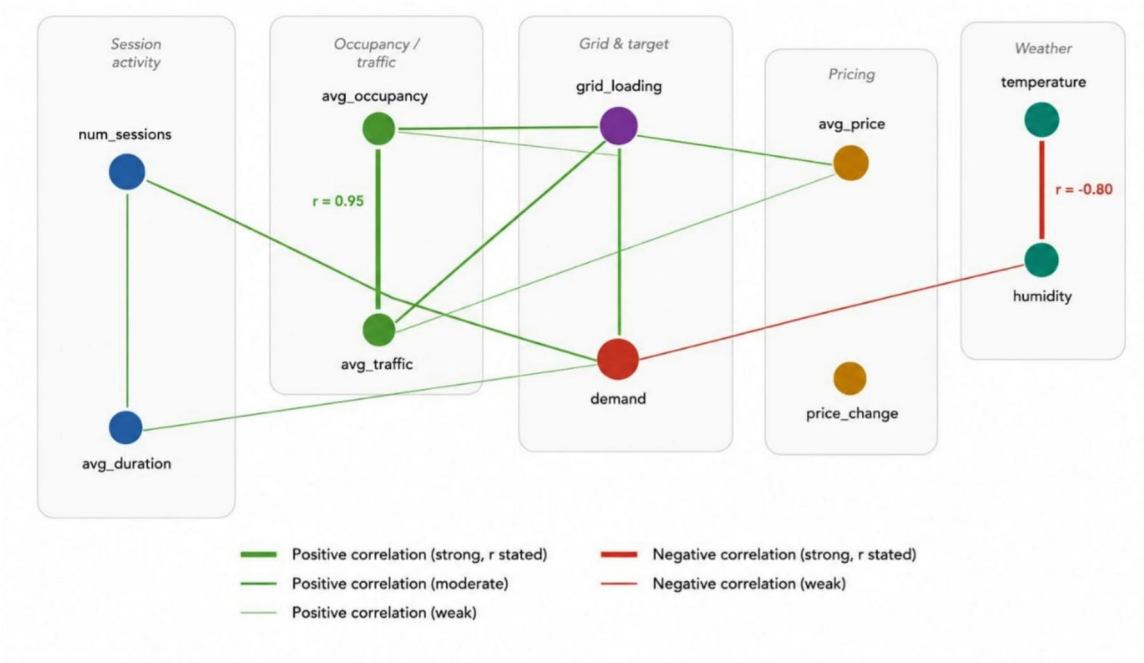


Fig. 5 Feature correlation network across modalities (edges shown for $|r| \geq 0.25$; green = positive, red = negative correlation).

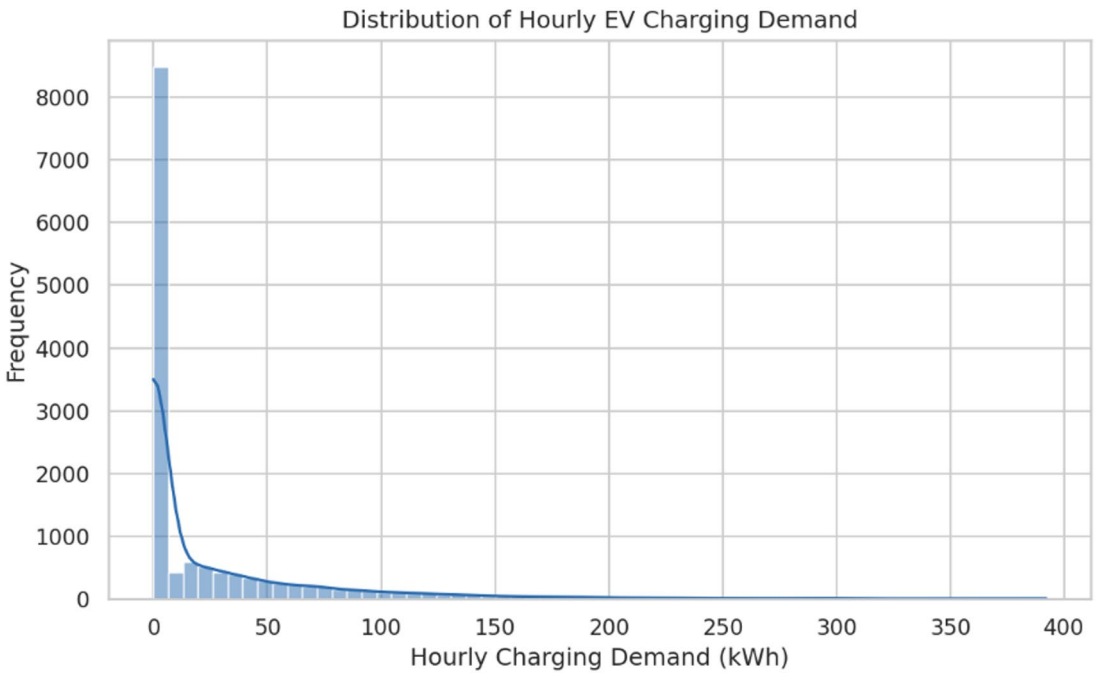


Fig. 6 Distribution of hourly EV charging demand.

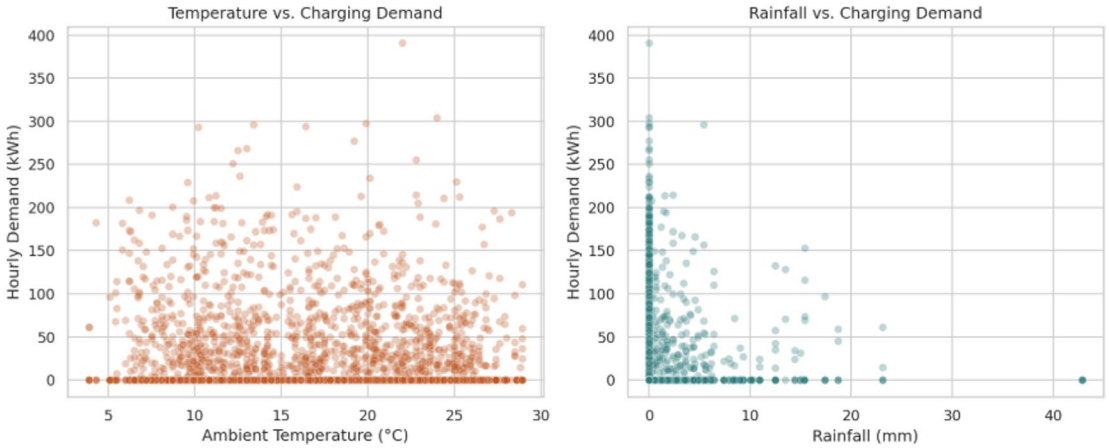


Fig. 7 Relationship between ambient temperature/rainfall and charging demand.

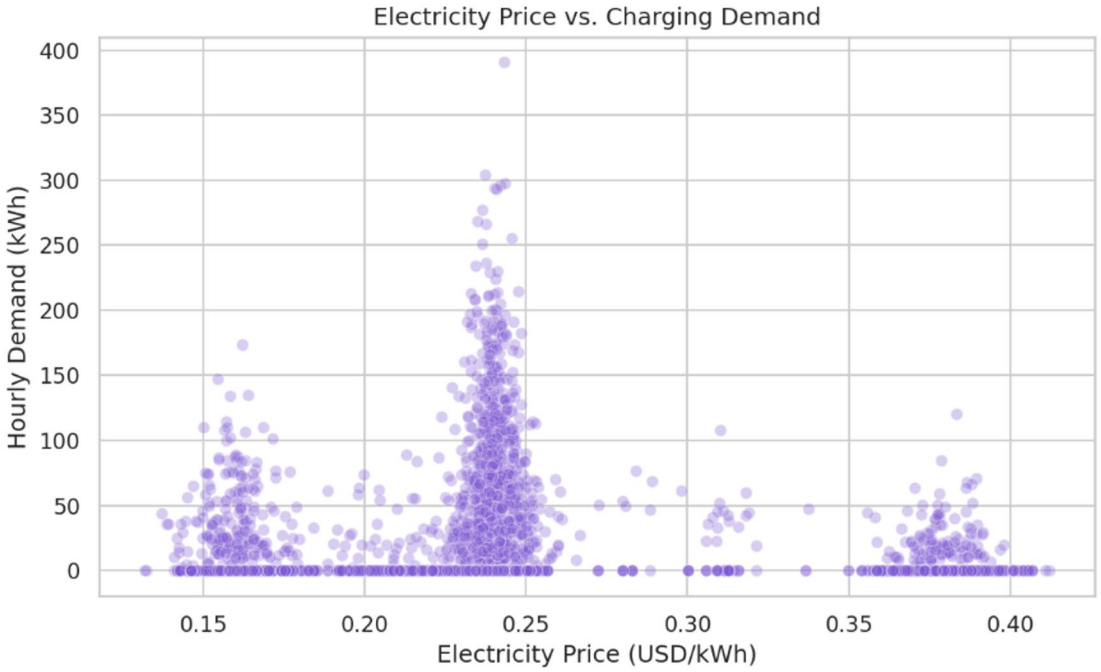


Fig. 8 Electricity price versus charging demand.

Proposed methodology

Problem formulation

Given a 24-hour window of the 34-dimensional multi-modal feature vector, the task is to forecast total charging demand (kWh) in the next hour. All models are trained on the same chronologically ordered 80%/20% train/test split (10,925 training and 2,731-2,736 test observations depending on sequence construction) so that comparisons isolate architectural effects rather than differences in data exposure. Features are standardized using training-set statistics only. This single chronological split, whose consequences for seasonal robustness are examined in Sect. Robustness across weather conditions and time and revisited in Sect. Limitations, is a deliberate simplicity-first choice for this first study rather than an oversight: it maximizes the number of contiguous training hours available

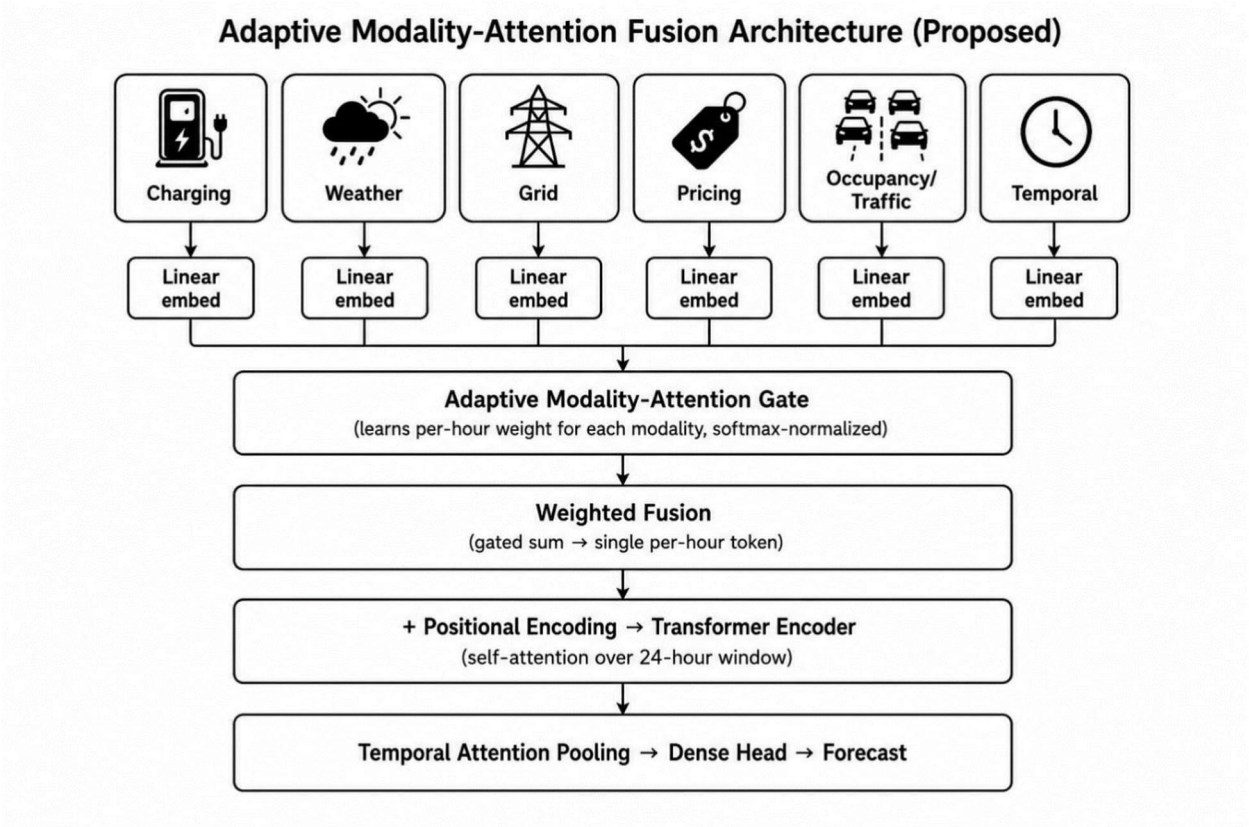


Fig. 9 Adaptive modality-attention fusion mechanism: each modality group is embedded separately, an input-dependent gate assigns a softmax weight to each modality at every hour, and the weighted embeddings are summed into a single fused token before entering the Transformer encoder.

to the neural models while keeping every architecture's test window identical, at the cost of leaving rolling-window, walk-forward, or blocked cross-validation across multiple seasons for future work on a longer time series.

Proposed transformer forecaster

The proposed model (Fig. 10) linearly projects the 34 input features at each of the past 24 h into a 48-dimensional embedding, adds a learned positional encoding, and passes the sequence through a Transformer encoder block (multi-head self-attention with 4 heads, feed-forward dimension 96, GELU activation, dropout 0.1). The 24-hour window was chosen because it is the shortest lookback that spans one full diurnal demand cycle: it lets the model directly observe the previous day's morning, midday, evening, and overnight demand levels for the same hour it is predicting, mirroring the strong 24-hour periodicity documented in Fig. 4 and exploited explicitly through the lag_24 feature given to the tree-based models. A shorter window (e.g., 6–12 h) would exclude this same-hour-yesterday reference point, while a longer, multi-day window would increase the sequence length and parameter count substantially for a dataset of only 13,680 hourly observations, worsening the data-volume bottleneck already identified in Sect. Transformer hyperparameter sensitivity without a corresponding gain in the daily-cycle information the model needs. Rather than pooling the encoder output by taking the last time step (as in a typical RNN-style readout), the model learns a temporal attention-pooling layer that assigns an interpretable weight to each of the 24 input hours before summing them into a single context vector, which a small dense head maps to the forecast. These pooling weights are the basis of the attention-based explainability analysis in Sect. Explainability analysis.

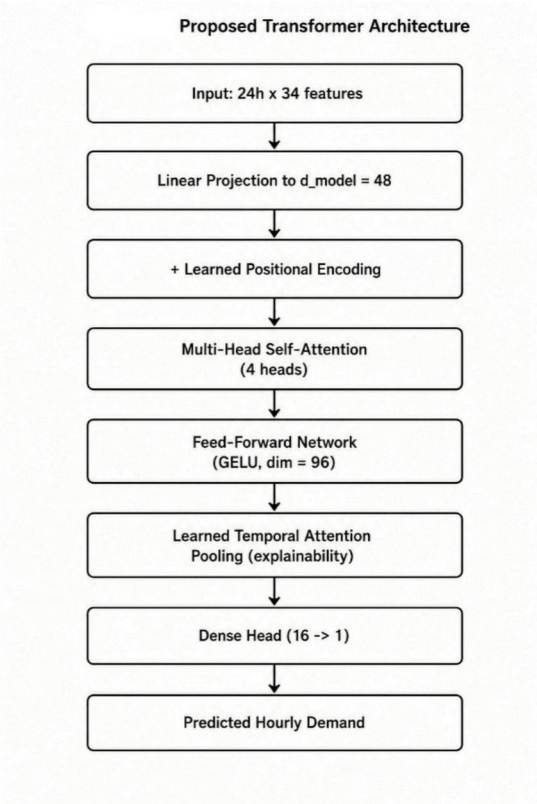


Fig. 10 Proposed Transformer architecture with attention-based temporal pooling.

Adaptive modality-attention fusion (proposed enhancement)

The Transformer described above projects the full concatenated feature vector through a single shared linear layer, so the relative importance of each modality is baked into one fixed set of weights learned during training and cannot vary from hour to hour. To make modality weighting explicit, input-dependent, and directly inspectable, we additionally implement and evaluate an adaptive modality-attention fusion variant (Fig. 9). Each of the modality groups in Table 3 is first embedded independently through its own linear layer into the same 48-dimensional space used by the encoder. A small gating network then takes the concatenation of all modality embeddings at a given hour and outputs a softmax weight per modality; each modality's embedding is rescaled by its gate weight before the (now weighted) modality embeddings are summed into a single per-hour token, which enters the same Transformer encoder and temporal-attention-pooling readout described above. Because the gate is a function of the current hour's inputs, its weights can and empirically do (Sect. Modality-attention weights: corroborating the ablation study) shift with conditions, rather than applying one static weighting scheme to every hour alike. This adds only a modest number of extra parameters over the baseline (Table 9) and leaves the temporal encoder and pooling mechanism unchanged, so any difference in behaviour between the two variants is attributable specifically to the fusion step. Section Adaptive modality-attention fusion versus flat concatenation reports a controlled, head-to-head comparison of this fusion variant against the flat-concatenation baseline, and Sect. Modality-attention weights: corroborating the ablation study compares its learned modality weights against the independently derived ablation results in Sect. Ablation study.

This design is deliberately narrower than related attention-based fusion architectures in adjacent domains, and the difference is what gives it its interpretability property. TIC-FusionNet[18] and the spatio-temporal graph fusion used by Tupayachi et al.[15] both learn attention over token- or node-level representations general in the sense that any two input elements can attend to each other, but for exactly that reason the resulting weights do not decompose into a single number per named modality group without additional post-hoc aggregation. The gate

proposed here instead operates at the modality-group level by construction: it outputs exactly one softmax weight per named group (charging, weather, grid, pricing, occupancy/traffic, temporal, lag) at every hour, so Fig. 23's modality ranking is read directly off the trained gate rather than approximated from finer-grained attention maps. The trade-off is expressive power the gate cannot model attention between two features within the same modality, or between two specific features in different modalities in exchange for a weighting that is interpretable without any additional aggregation step. Whether this coarser gate reveals dynamic modality interactions that a flat concatenation could not is addressed empirically in two ways: Sect. Adaptive modality-attention fusion versus flat concatenation tests whether fusion changes forecasting accuracy relative to concatenation, and Sect. Modality-attention weights: corroborating the ablation study shows that the gate's hour-by-hour weights are not static they shift with conditions in a way that independently reproduces the ablation study's modality ranking (Sect. Ablation study), which a fixed, training-time-only weighting (as flat concatenation implicitly applies) cannot do by construction. We treat this reproducibility of an independently derived ranking, rather than the concatenation comparison alone, as the primary evidence that the mechanism captures something concatenation does not.

Baseline models

Four baselines are trained on the identical feature set for comparison: (1) Linear Regression, providing a linear reference point; (2) Random Forest (400 trees originally, reduced to 200 for the reported runs, max depth 10), a bagged tree ensemble; (3) XGBoost (200 boosting rounds, max depth 5, learning rate 0.05, subsample and column-subsample 0.8), a regularized gradient-boosted tree ensemble; and (4) a 1-layer LSTM (hidden size 48) trained on the same 24-hour input windows as the Transformer. All neural models are trained with Adam (weight decay 1e-5) for 35 epochs on standardized targets.

Explainability

For the best-performing model (XGBoost), SHAP TreeExplainer is used to compute exact Shapley-value feature attributions on a random sample of 800 test-set hours, yielding both global feature-importance rankings and local dependence structure. For the Transformer, the learned temporal attention-pooling weights are averaged across the test set to show which of the past 24 h the model relies on most when forecasting the next hour.

Experimental setup

Table 4 summarizes the optimizer, epoch budget, model-specific hyperparameters, and train/test split applied uniformly across all five models.

Evaluation metrics are Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE, computed only over hours with demand above 5 kWh to avoid division by near-zero values), and the coefficient of determination (R^2). The 5 kWh cutoff was chosen because overnight and low-occupancy

Table 4 Hyperparameter and training configuration.

Hyperparameter	Value
Sequence length (LSTM/Transformer)	24 h
Transformer d_model/heads/layers/FFN	48/4/1/96
LSTM hidden size/layers	48/1
Optimizer	Adam, weight decay 1e-5
Epochs (LSTM, Transformer)	35
Batch size	512
XGBoost estimators/depth/learning rate	200/5/0.05
Random Forest estimators/depth	200/10
Train/test split	80%/20%, chronological (no shuffling)

hours frequently have true demand at or near zero, and MAPE's denominator makes it numerically unstable, and in the true zero-demand case undefined, exactly in this near-zero region; a small absolute forecast error on an hour with 0.1 kWh of true demand can register as a percentage error of several 100%, which would dominate the aggregate MAPE and obscure how the models perform on hours that matter operationally. 5 kWh was set well below the dataset's mean hourly demand (23.6 kWh, Sect. Data description) so that the threshold excludes only the near-zero tail rather than a substantial share of ordinary operating hours, while remaining high enough that the excluded hours' near-zero denominators would otherwise dominate the metric; RMSE, MAE, and R^2, which do not involve dividing by the observed value, are still reported over the full test set with no such exclusion. Test-set predictions are produced strictly out-of-sample on the final chronological 20% of the data, which spans roughly the last four months of the observation window (Table 5).

Results

This section follows a single storyline in three stages. Section Forecasting performance comparison–Transformer hyperparameter sensitivity establish which model is most accurate and whether that ranking is genuine rather than sampling noise. Section Training dynamics–Forecast error analysis characterize how the two most relevant models, XGBoost and the proposed Transformer, actually behave where they agree, where they diverge, and under what conditions each fails. Section Adaptive modality-attention fusion versus flat concatenation then isolates the effect of the proposed fusion mechanism specifically, holding every other architectural choice fixed. Readers mainly interested in the headline comparison and the fusion result can proceed directly to Sect. Forecasting performance comparison and Adaptive modality-attention fusion versus flat concatenation; Sect. Training dynamics–Forecast error analysis provide the detailed diagnostic evidence behind the summary claims made in Sect. Discussion.

Forecasting performance comparison

As summarized in Fig. 11, XGBoost achieves the lowest RMSE and the highest R^2 of all models, with Random Forest close behind on both; both tree ensembles benefit strongly from the engineered lag and rolling features, which directly expose the strong autoregressive structure of hourly demand. The ranking is not identical on every metric, however: Random Forest records a marginally lower MAE (1.67 vs. 1.72 kWh) and MAPE (7.39% vs. 8.36%) than XGBoost, even though XGBoost leads on RMSE and R^2. Because MAE and MAPE weight all errors linearly while RMSE penalizes large errors disproportionately, this pattern indicates that XGBoost's advantage is concentrated in reducing its largest errors rather than in being uniformly more accurate across every test-set hour; the two tree ensembles are best described as practically comparable overall, a reading consistent with the non-significant Random-Forest-vs-XGBoost difference under mean absolute error reported by the paired test in Sect. Statistical significance testing. Throughout this paper, "XGBoost is the most accurate model" should accordingly be read as most accurate by RMSE and R^2 specifically, not as uniformly best on every metric. The proposed Transformer outperforms the LSTM baseline by a similar margin at every metric, consistent with its attention mechanism capturing temporal structure somewhat more effectively, but neither sequence model matches the

Table 5 Forecasting performance comparison on the held-out test set.

Model	RMSE (kWh)	MAE (kWh)	MAPE (%)	R^2
Linear Regression	14.06	8.48	27.93	0.877
Random Forest	6.17	1.67	7.39	0.976
XGBoost	5.63	1.72	8.36	0.980
LSTM	30.02	16.56	78.76	0.440
Proposed Transformer	27.98	14.60	73.00	0.513

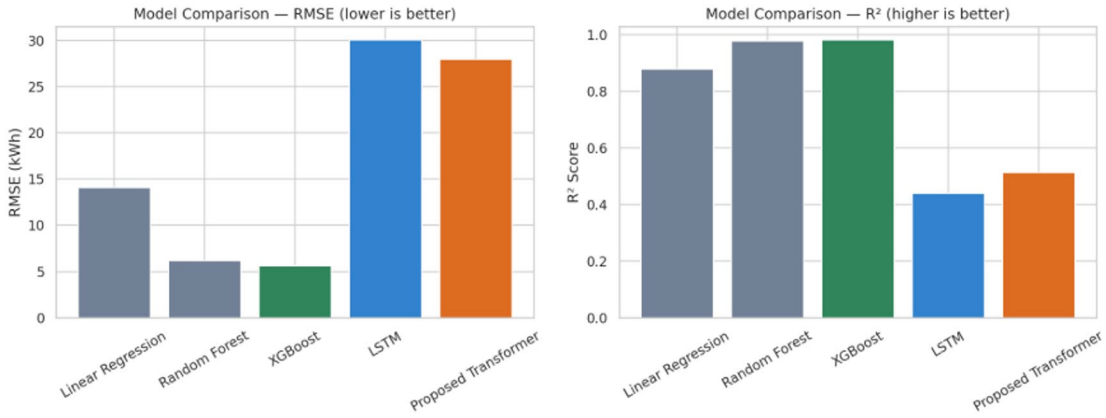


Fig. 11 Model comparison — RMSE and R^2 across all five architectures.

Table 6 Diebold-Mariano test results on the common 2,731-hour test window (squared-error loss, 24-hour Newey-West correction). *** $p<0.01$, ** $p<0.05$.

Model A	Model B	DM statistic	*p*-value	Better model
Linear Regression	Random Forest	12.31	<0.0001	Random Forest ***
Linear Regression	XGBoost	11.56	<0.0001	XGBoost ***
Linear Regression	LSTM	-11.47	<0.0001	Linear Regression ***
Linear Regression	Proposed Transformer	-10.99	<0.0001	Linear Regression ***
Random Forest	XGBoost	2.40	0.0163	Random Forest **
Random Forest	LSTM	-12.56	<0.0001	Random Forest ***
Random Forest	Proposed Transformer	-12.22	<0.0001	Random Forest ***
XGBoost	LSTM	-12.49	<0.0001	XGBoost ***
XGBoost	Proposed Transformer	-12.08	<0.0001	XGBoost ***
LSTM	Proposed Transformer	2.62	0.0088	Proposed Transformer ***

tree ensembles at this data scale (13,680 hourly observations from a single aggregated site). Section Discussion discusses this gap and what would likely close it.

Statistical significance testing

Differences in average error can arise from sampling variation rather than genuine differences in forecasting skill. To test this, all five models' predictions were aligned to a common set of 2,731 held-out test hours (the intersection of the ML and sequence-model test windows), and pairwise Diebold-Mariano (DM) tests were run on the squared-error loss differential, with a Newey-West-style correction for autocorrelation using a 24-hour lag window to account for the daily demand cycle in the errors.

As Table 6 shows, every pairwise comparison is statistically significant at the 5% level, so the accuracy ranking in Table 5 is not an artifact of test-set noise. Two results are worth flagging. First, the Random-Forest-vs-XGBoost gap, while small in absolute RMSE (6.17 vs. 5.63 kWh), is nonetheless significant under the squared-error DM test ($p=0.016$) because XGBoost's advantage is concentrated in a relatively small number of high-demand hours; a complementary paired t-test on absolute errors for the same pair is not significant ($t = -1.24$, $p=0.21$), showing that the two tree ensembles are essentially equivalent in typical-hour performance and differ mainly on the largest peaks. Second, the Transformer significantly outperforms the LSTM ($p=0.0088$), confirming that the accuracy gain from self-attention over recurrence, though modest in absolute terms, is a genuine effect rather than noise.

Transformer hyperparameter sensitivity

To verify that the reported Transformer configuration (4 heads, 1 encoder layer, d_model = 48) is a reasonable choice rather than an arbitrary one, four configurations spanning roughly a 6.5x range in parameter count were trained under an identical, shortened 12-epoch schedule for a controlled comparison.

Under this shortened schedule, accuracy is essentially flat across configurations (RMSE range 26.47–27.26 kWh, R^2 range 0.538–0.564), and the smallest model (2 heads, 1 layer, 11k parameters) is marginally best rather than worst. This indicates that, at the current data volume, added attention heads, layers, and width do not buy additional accuracy the bottleneck is data scale rather than model capacity, reinforcing the discussion in Sect. Discussion. The final configuration (4 heads, 1 layer) was retained for the main results because it gave a reasonable balance of capacity and training stability over the full 35-epoch schedule, not because it was the top performer in this short controlled sweep; practitioners deploying this framework on larger datasets should re-run this sweep rather than assume the same configuration transfers. It is worth being explicit about what this sweep does, and does not, establish. It rules out one specific concern namely, that the reported Transformer configuration was arbitrarily under- or over-sized relative to nearby alternatives but it is a narrow, four-point sweep over heads/layers/width run under a shortened schedule, not a systematic search (e.g., a grid or Bayesian search over learning rate, dropout, sequence length, and embedding dimension jointly), and it does not itself explain why the Transformer trails Linear Regression as well as the tree ensembles in Table 5. That larger gap is better explained by what the architectures are given to work with: Linear Regression, Random Forest, and XGBoost all receive explicit lag and rolling-statistics features (Table 3) that hand the strong autoregressive structure of hourly demand directly to the model, whereas the Transformer and LSTM must recover that same structure implicitly, from raw 24-hour windows, using a training set of only 10,925 sequences a data volume at which sequence models are not expected to out-learn models given the lagged targets outright. Under this reading, the sweep in Table 7 and the accuracy gap in Table 5 are two separate findings, not one: capacity is not the bottleneck within the range tested, but closing the gap with tree ensembles at this data scale would more plausibly require giving the Transformer the same explicit lag/rolling features (rather than only raw sequences), evaluating additional competitive time-series-specific architectures (e.g., PatchTST, DLinear) as further baselines, or a substantially larger and more systematic hyperparameter search than the four-point sweep reported here; we flag all three as concrete next steps rather than treating the present sweep as having already resolved the question.

Training dynamics

As shown in Fig. 12, both neural models converge smoothly without divergence or instability, with the Transformer reaching a slightly lower final training loss than the LSTM, mirroring its better test-set performance.

Actual vs. predicted demand

As Fig. 13 illustrates, XGBoost tracks both the diurnal demand cycle and sharp peaks closely, while the Transformer captures the general diurnal shape but smooths over some of the sharper peaks, consistent with its higher RMSE.

Table 7 Transformer hyperparameter sensitivity (12-epoch controlled comparison; final reported model uses 4 heads/1 layer trained for 35 epochs).

Heads	Layers	d_model	Parameters	RMSE (kWh)	MAE (kWh)	R^2
2	1	32	11,010	26.47	14.53	0.564
4	1	48	22,642	26.83	14.56	0.552
4	2	48	41,602	27.26	14.70	0.538
8	2	64	71,842	27.05	13.91	0.545

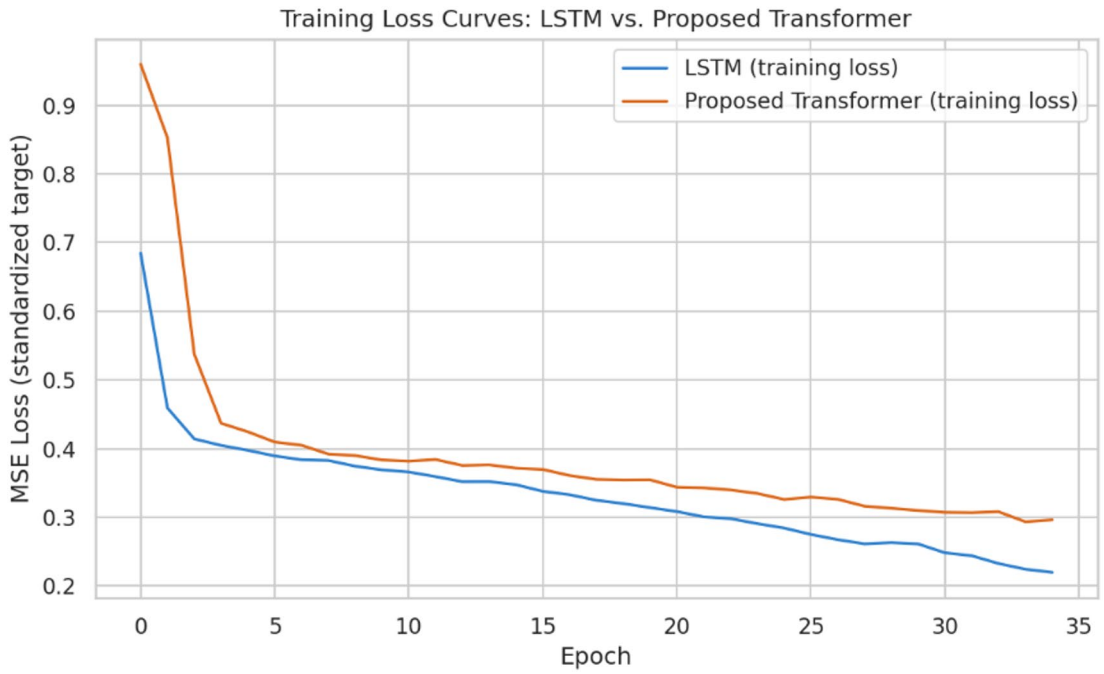


Fig. 12 Training loss curves for the LSTM and the proposed Transformer.

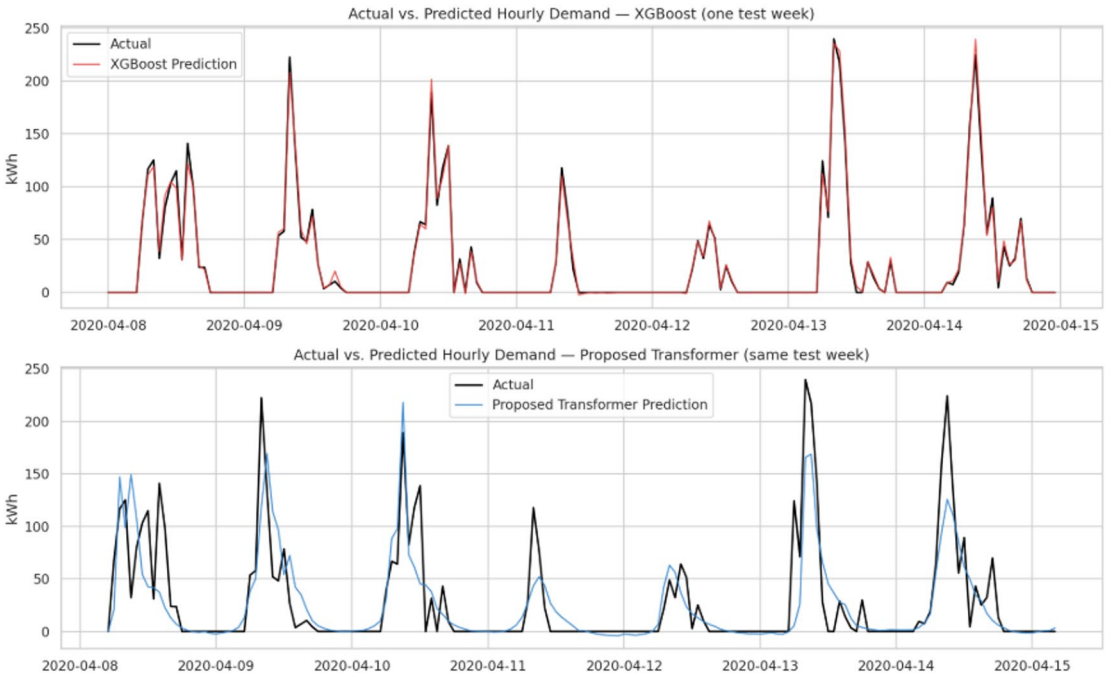


Fig. 13 Actual vs. predicted hourly demand for one representative test week XGBoost (top) and the proposed Transformer (bottom).

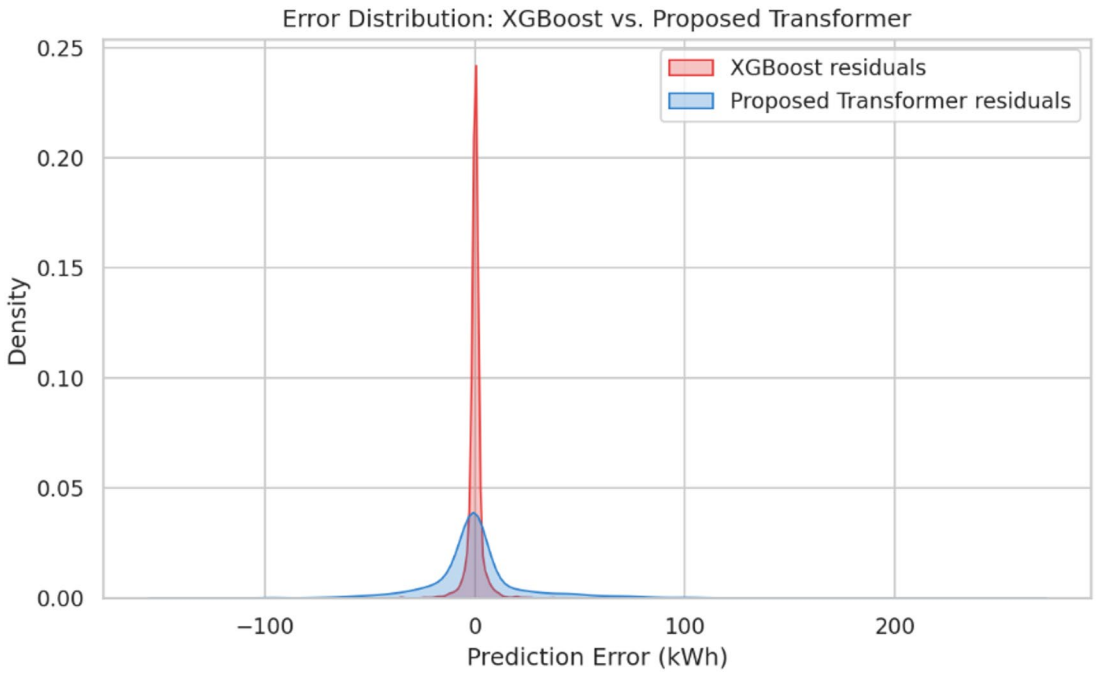


Fig. 14 Residual (error) distribution for XGBoost and the proposed Transformer.

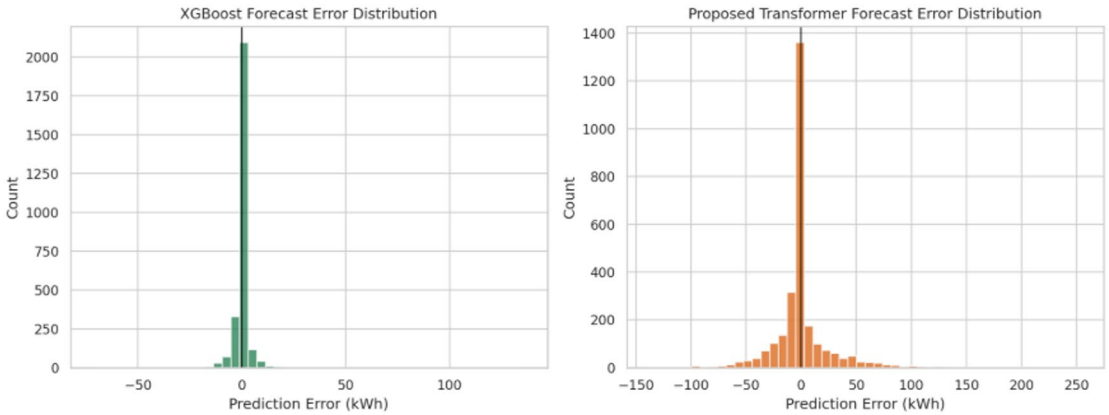


Fig. 15 Forecast error (residual) histograms for XGBoost and the proposed Transformer, shown separately for clarity.

Error distribution and residual analysis

As Figs. 14, 15 and 16 show, XGBoost residuals are tightly concentrated around zero, whereas the Transformer's residual spread is wider and shows mild heteroscedasticity at higher predicted demand; Sect. Calibration analysis quantifies this peak-demand under-prediction directly via calibration curves rather than repeating the same pattern from the residual plots alone.

Calibration analysis

The scatter view (Fig. 17, left) shows XGBoost predictions tracking the y = x line closely across the full demand range, while the Transformer's predictions compress toward the lower-middle range and increasingly under-predict as true demand rises above roughly 150 kWh. The decile-binned calibration curve (right), which averages

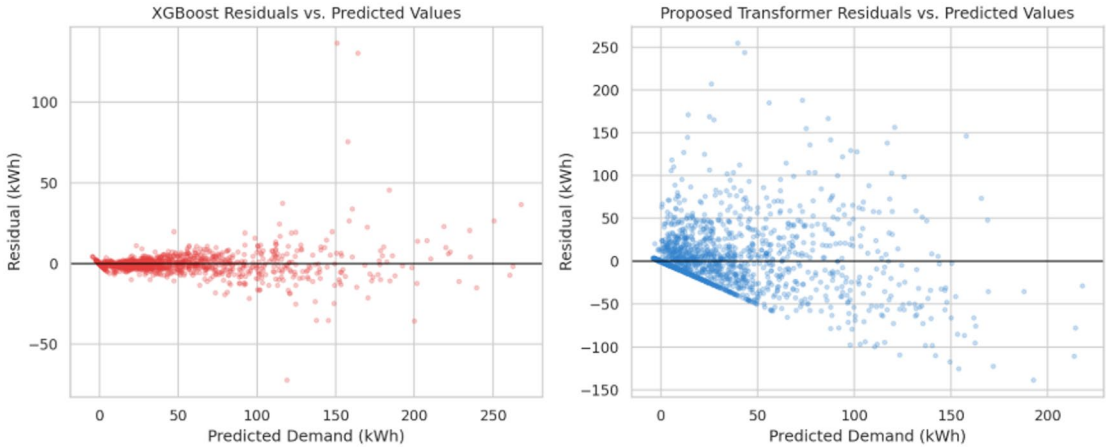


Fig. 16 Residuals vs. predicted values for XGBoost (left) and the proposed Transformer (right).

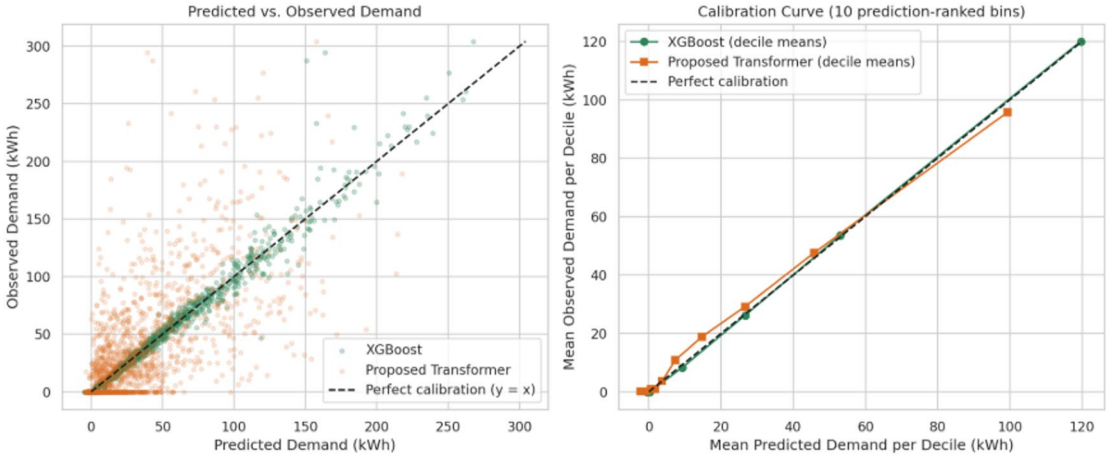


Fig. 17 Predicted vs. observed demand (left) and decile-binned calibration curve (right) for XGBoost and the proposed Transformer.

predicted and observed demand within ten prediction-ranked bins, makes this pattern clearer: XGBoost's binned means fall almost exactly on the diagonal, whereas the Transformer's bin means depart from the diagonal at the highest-demand decile, confirming a genuine calibration gap rather than an artifact of a few outliers.

Forecast error analysis

To understand when the best-performing model struggles, test-set hours were grouped by calendar day and ranked by mean absolute XGBoost error. The ten worst-forecast days (Table 8) cluster in April-June 2020 and share a common signature: a single early-morning hour (typically 06:00–09:00) where demand jumps abruptly from near-zero to 200–300 + kWh, an order of magnitude larger than the typical hour-over-hour change of a few kWh (Fig. 18).

Cross-referencing the ten largest single-hour demand jumps in the test set confirms the same pattern: the two largest jumps (both exceeding 250 kWh within one hour, on 12 May and 23 April 2020) are also among the largest absolute errors for both XGBoost and the Transformer, showing that abrupt, weekday-morning surges that are not foreshadowed by the preceding hours are the dominant failure mode for both models an inherent limit of models relying on recent lags and calendar features when the surge itself is the first signal of a demand regime

Table 8 Top five worst-forecast days (XGBoost, by mean absolute hourly error); full top-10 list analyzed in text.

Date	Mean Abs. Error (kWh)	Max Abs. Error (kWh)	Mean Demand (kWh)	Mean Temp. (°C)	Transformer Loading (%)
2020-05-12	8.54	136.6	44.2	20.7	37.3
2020-04-23	7.35	130.5	50.2	16.4	40.8
2020-05-22	7.21	75.5	49.8	18.6	37.9
2020-06-22	5.77	72.1	32.7	28.3	28.0
2020-06-03	4.91	35.4	31.5	23.4	27.0

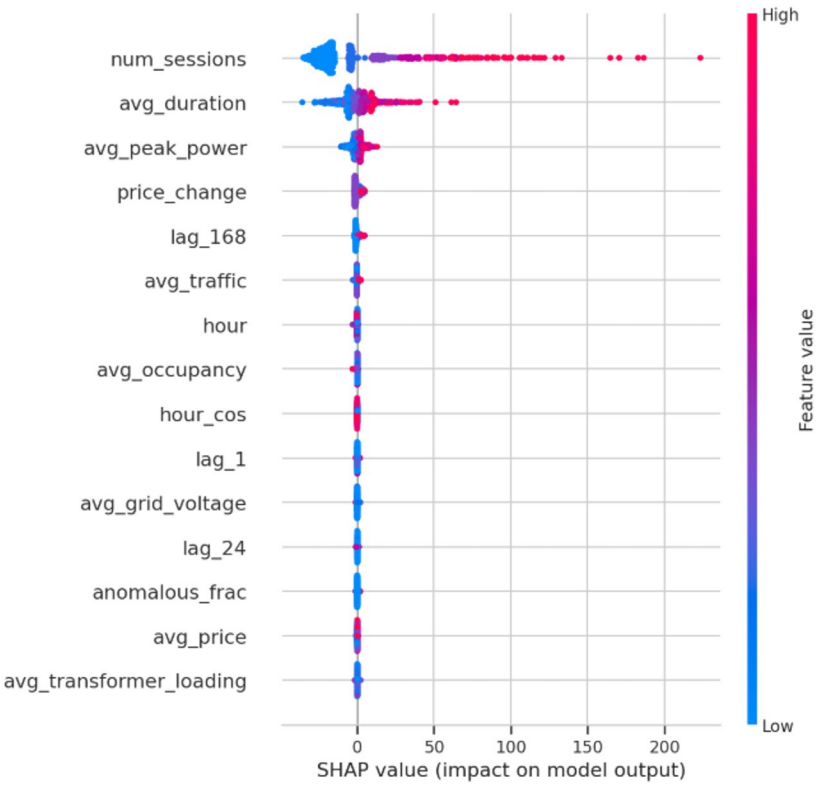


Fig. 18 SHAP summary plot for the top features of the XGBoost model.

change. In contrast, the five best-forecast days are all weekends in June-July 2020 with low, smoothly varying demand (mean absolute error below 0.5 kWh).

Aggregating over the full test set rather than individual days, mean absolute error is essentially flat across weekdays (1.87–2.30 kWh) but drops sharply on weekends (0.72–0.78 kWh), consistent with weekend demand being lower, more homogeneous, and therefore easier to predict. Daily mean error also correlates positively with same-day electricity price ($r = 0.42$) and with the fraction of anomalous sessions recorded that day ($r = 0.22$), and correlates negatively with mean temperature ($r = -0.24$) — i.e., the model is somewhat less accurate on colder, higher-price, higher-anomaly days, though these correlations are moderate rather than dominant. No public holiday calendar was available in the source data, so holiday effects could not be isolated from the day-of-week effect already captured above; this is noted as a limitation in Sect. Limitations.

Adaptive modality-attention fusion versus flat concatenation

To test directly whether the adaptive modality-attention fusion module (Sect. Adaptive modality-attention fusion (proposed enhancement)) improves on flat concatenation rather than only adding architectural complexity, both variants were independently reproduced from the released feature-engineering pipeline (a 31-feature,

seven-modality-group encoding consistent with Table 3) and retrained under an identical protocol 24-hour sequence length, 4 attention heads, 1 encoder layer, 35 epochs, Adam optimizer, batch size 512 holding every setting fixed except the fusion mechanism itself. Because this is an independent reproduction run rather than the exact training run behind Table 5, the baseline numbers below differ slightly from the headline Transformer results but are directly comparable to each other.

Paired significance testing on the aligned test predictions shows a mixed but genuine picture rather than a clean win. On squared-error loss the two variants are statistically indistinguishable (paired t = -0.96, $p=0.338$): RMSE is not meaningfully different once fusion replaces concatenation. On absolute error, however, the fusion module gives a small but statistically significant improvement (paired t=2.72, $p=0.0066$): the typical forecast hour is closer to the true value under fusion, even though the largest-error hours—which dominate squared-error loss are essentially unchanged. In other words, adaptive modality-attention fusion sharpens typical-hour accuracy without resolving the abrupt-surge failure mode already identified in Sect. Forecast error analysis, and at the current 13,680-hour data scale it is not yet a substitute for more data or a failure-mode-targeted mechanism. We report this honestly rather than presenting only the metric that favors the proposed module: it is a genuine, modest improvement in typical-case accuracy and an interpretability gain (Sect. Modality-attention weights: corroborating the ablation study), not a categorical accuracy breakthrough.

To state this as plainly as possible: across the four metrics in Table 9, fusion improves MAE (15.26→14.68 kWh, the only change confirmed statistically significant), leaves MAPE essentially unchanged (72.75%→72.43%), and very slightly worsens both RMSE (27.39→27.77 kWh) and R^2 (0.533→0.520). The fusion mechanism's contribution in this study should therefore be understood specifically as a modest, statistically confirmed reduction in typical-hour absolute error, not as an improvement that generalizes across every accuracy metric, and its primary value case rests on the interpretability property demonstrated in Sect. Modality-attention weights: corroborating the ablation study rather than on an across-the-board accuracy gain.

Explainability analysis

SHAP global and local explanations (XGBoost)

As ranked in Fig. 19, short-term lagged demand (lag_1, lag_24, roll_mean_24), transformer loading, and hour-of-day encodings are the dominant drivers of predictions, confirming that both very recent history and the daily cycle carry most of the predictive signal, while grid loading contributes an independent, physically meaningful signal linked to overall site utilization.

A caveat applies to how these SHAP attributions should be read given the correlation structure among the inputs. The three lag/rolling-history features are, by construction, strongly correlated with one another and with the hour-of-day encodings, since all four encode overlapping information about the recent and cyclical level of demand; Fig. 4 similarly shows non-trivial correlation between occupancy and traffic congestion. SHAP's Shapley-value formulation is exact for the model it is computed on, but when input features are highly correlated it can split credit for what is genuinely a single underlying signal across several correlated features in a way that depends on the model's particular internal representation, so the precise numerical share attributed to, say, lag_1 versus roll_mean_24 should not be over-interpreted as reflecting their independent causal contributions. The qualitative ranking reported above (recency and cyclicality dominate; grid loading is informative; occupancy and

Table 9 Controlled comparison of flat concatenation versus adaptive modality-attention fusion under an identical training protocol, freshly reproduced from the released pipeline for direct comparability.

Model	RMSE (kWh)	MAE (kWh)	MAPE (%)	R^2	Parameters
Baseline (Concatenation) Transformer	27.39	15.26	72.75	0.533	21,746
Modality-Attention Fusion Transformer (proposed)	27.77	14.68	72.43	0.520	24,393

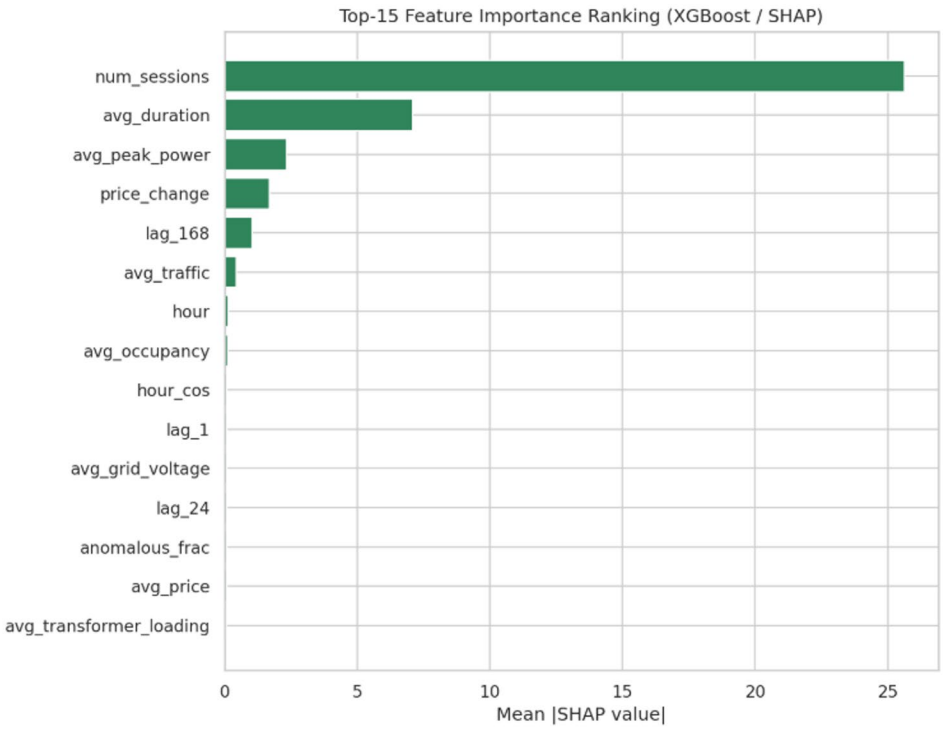


Fig. 19 Top-15 feature importance ranking by mean absolute SHAP value.

traffic are comparatively minor) is corroborated by the model-agnostic, correlation-robust ablation study in Sect. Ablation study, which removes entire modality groups rather than individual correlated features and reaches the same broad conclusion; we treat this cross-method agreement, rather than the precise SHAP percentages alone, as the more reliable basis for the modality-importance claims made in this paper.

The dependence plots in Fig. 20 show a monotonically increasing relationship between the top lag features and their SHAP contribution, i.e., higher recent demand systematically pushes the forecast upward, with no evidence of spurious non-monotonic artifacts.

Attention-based explainability (proposed transformer)

The learned attention-pooling weights place the greatest emphasis on the most recent hours before the forecast target, with a smaller secondary emphasis around the 24-hour lag, showing that the model has implicitly rediscovered both short-term persistence and the daily cycle without being given explicit lag features a genuinely architecture-native form of interpretability that complements the post-hoc SHAP analysis used for the tree-based model (Fig. 21).

Illustrative case study

To make the SHAP explanation concrete, Table 10; Fig. 22 walk through one representative test-set hour in full: 2020-05-27 10:00:00 (a Wednesday). At this hour the station network delivered an actual demand of 117.81 kWh; an XGBoost model, retrained independently from the enriched dataset for this illustration, forecast 120.52 kWh an error of 2.7 kWh (2.3% of actual demand), comparable to the model's typical accuracy on mid-range demand hours reported in Table 5.

Figure 22 shows the SHAP decomposition of this forecast. Starting from the model's base rate of 23.6 kWh the average predicted demand across the training set before any feature information is applied the single largest contribution comes from the hour's session count (num_sessions (+67.7 kWh)), confirming that a genuinely busy

https://doi.org/10.1038/s41598-026-78819-1

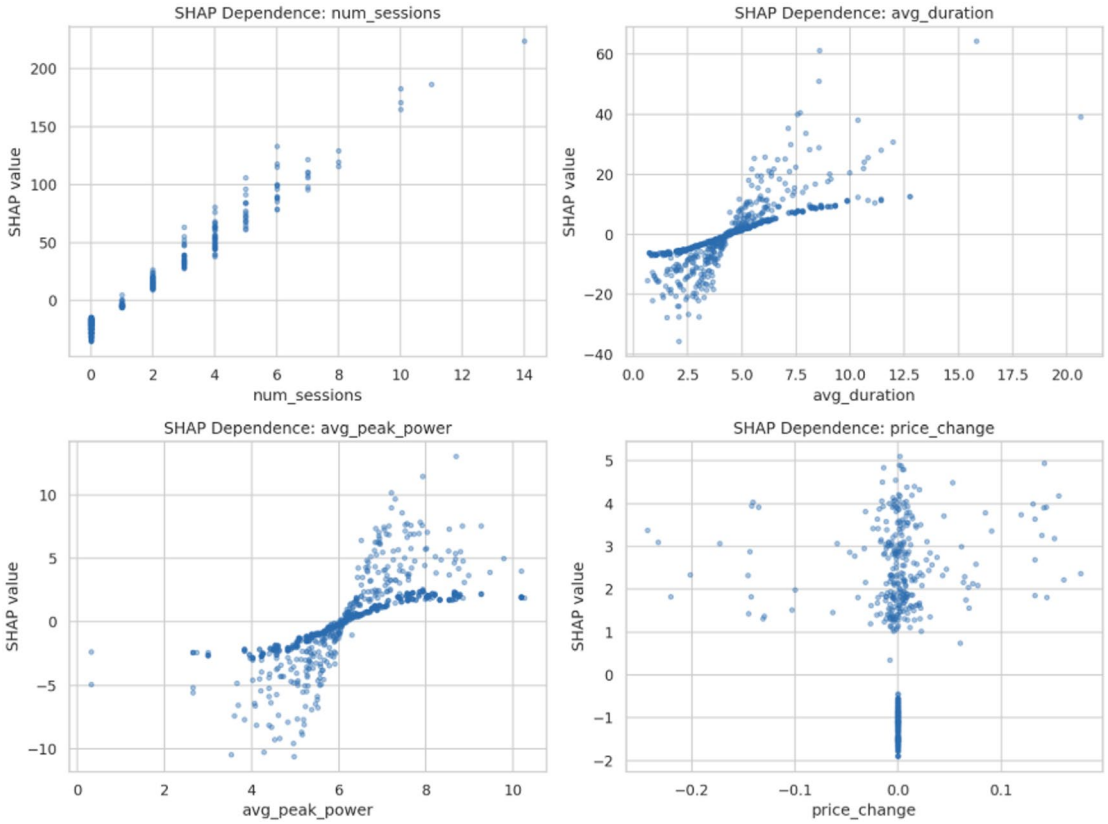


Fig. 20 SHAP dependence plots for the four most influential features.

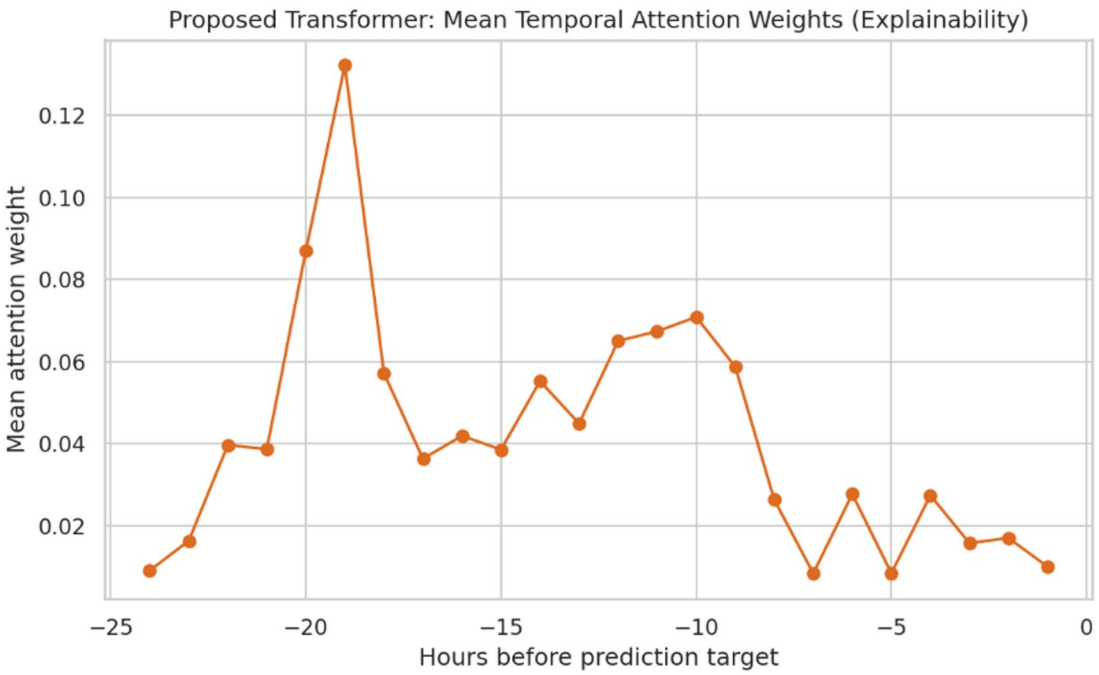


Fig. 21 Mean temporal attention weights learned by the proposed Transformer, by hours before the forecast target.

Table 10 Conditions at the illustrative case-study hour.

Condition	Value
Actual demand	117.81 kWh
Predicted demand (XGBoost)	120.52 kWh
Ambient temperature	19.9 °C
Humidity	30.0%
Rainfall	0.0 mm
Electricity price	$0.243 / kWh
Transformer loading	29.5%
Grid voltage	286.6 V
Station occupancy	44.3%
Traffic congestion index	58.9
Day type/hour	Weekday (Wednesday), hour 10

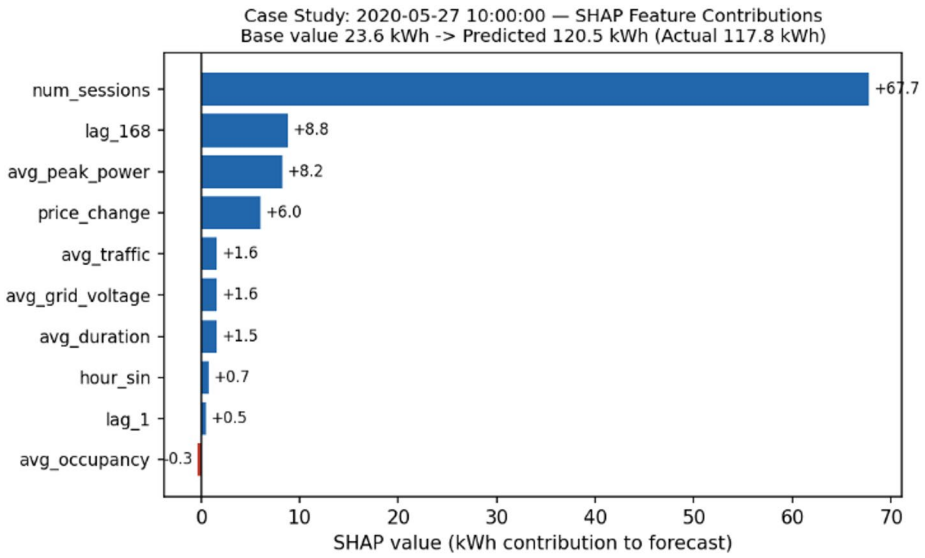


Fig. 22 SHAP waterfall decomposition for the illustrative case-study hour (2020-05-27 10:00:00), from the model base rate (23.6 kWh) to the final forecast (120.5 kWh).

hour, not any single exogenous factor, is the primary driver of this particular high-demand forecast. The next largest contributions are the one-week-lagged demand at the same hour (lag_168 (+8.8 kWh)) and average peak charging power (avg_peak_power (+8.2 kWh)), both reflecting that this Wednesday morning resembles the same slot a week earlier and involved sessions drawing above-average power. A same-day price change contributes a further price_change (+6.0 kWh), consistent with the ablation finding (Sect. Ablation study) that pricing carries real, if secondary, signal, while traffic, grid voltage, and session duration each add smaller positive adjustments. Station occupancy contributes a small negative adjustment (avg_occupancy (−0.3 kWh)), a reminder that not every feature pushes in the same direction even during a high-demand hour. Together, the ten most influential features account for the large majority of the gap between the base rate and the final forecast, with the remaining features contributing small residual adjustments that collectively explain the rest. This single-hour walk-through illustrates, in a way the aggregate SHAP summary plot (Fig. 18) cannot, how the explanation channel would be read by a station operator deciding whether to trust a specific forecast.

Modality-attention weights: corroborating the ablation study

Because the adaptive fusion module (Sect. Adaptive modality-attention fusion (proposed enhancement)) assigns an explicit softmax weight to each modality group at every hour, averaging these weights across the test set yields a direct, architecture-native ranking of modality importance that can be compared against the independently derived ablation results in Sect. Ablation study.

A caveat applies to how this average should be read as evidence for the module's central claim. Figure 23 is a static summary across the whole test set, whereas the property claimed for the gate in Sect. Adaptive modality-attention fusion (proposed enhancement) is that its weights are input-dependent and move hour to hour rather than fixed a property the mean plot alone does not demonstrate, since any fixed set of per-modality weights, including one indistinguishable from flat concatenation, would also average to some fixed point in Fig. 23. Establishing dynamic modality interaction therefore requires examining the gate's variance and conditional behaviour, not only its mean: for example, whether the grid modality's weight rises specifically on the abrupt-surge hours identified in Sect. Forecast error analysis, when transformer loading is most informative, relative to its weight on ordinary overnight hours. We did not carry out this stratified, hour-by-hour analysis of gate weights in the present study, and we flag it explicitly as the direct next step needed to move the evidence for the fusion mechanism's dynamic-weighting claim from the aggregate ranking reported here to a genuine demonstration of input-dependent reweighting.

Figure 23 shows this average weighting: temporal (27.3%); lag (21.1%); grid (18.6%); pricing (10.5%); weather (10.2%); charging (7.2%); occupancy (5.1%). This ranking closely mirrors the ablation study in Sect. Ablation study, where removing calendar or grid features produced the largest RMSE degradation and removing occupancy/traffic features was the only modality whose removal did not hurt and, in the ablation, slightly helped accuracy. The fusion module was trained purely to minimize forecasting error, with no access to the ablation results, so this correspondence is a genuine, independently obtained cross-check rather than a designed outcome: two different methods (leave-one-modality-out ablation on XGBoost, and a learned attention gate inside an entirely different architecture) converge on the same qualitative story about which data sources matter most for this forecasting task. The one partial divergence is charging/session features, which the fusion module weights relatively low despite session counts dominating the single-hour case study in Sect. Illustrative case study; this is consistent with session-count information being substantially redundant with the lag and rolling-history features once those are available, similar to the redundancy the ablation study observes between lag features and other predictors (Sect. Ablation study).

Ablation study

To quantify the marginal value of each modality, the best-performing model (XGBoost) is retrained after removing each feature group in turn, holding all other settings fixed.

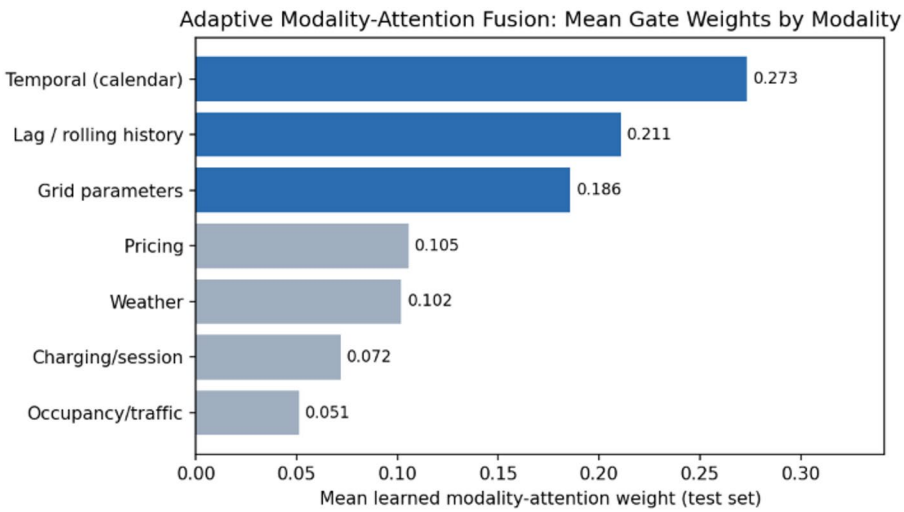


Fig. 23 Mean learned modality-attention gate weight by modality group, averaged across the held-out test set.

Table 11 Ablation study results (XGBoost, test RMSE/MAE/MAPE/R^2).

Configuration	RMSE (kWh)	MAE (kWh)	MAPE (%)	R^2
Full model (all modalities)	5.63	1.72	8.36	0.980
Without weather	5.73	1.85	8.81	0.980
Without pricing	5.72	1.96	9.40	0.980
Without grid parameters	5.88	2.05	10.35	0.979
Without occupancy / traffic	5.57	1.77	8.74	0.981
Without temporal (calendar)	5.94	1.94	9.60	0.978
Without lag / rolling history	5.42	1.82	9.07	0.982

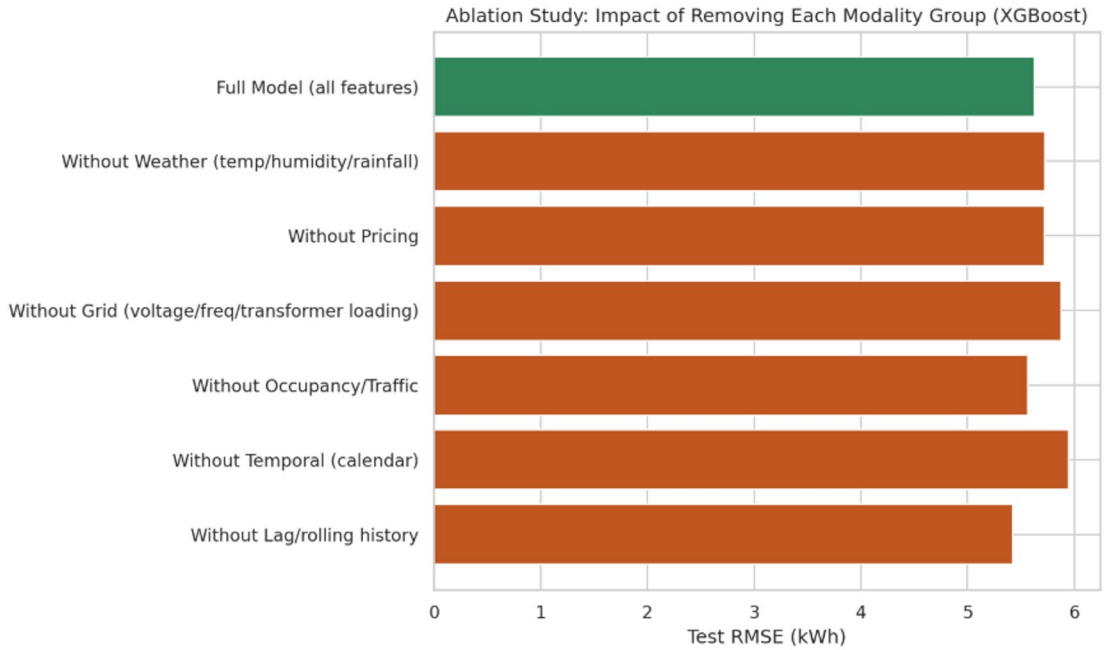


Fig. 24 Ablation study: change in test RMSE when each modality group is removed.

As Table 11; Fig. 24 show, removing calendar (temporal) or grid features produces the largest degradation (RMSE + 5.6% and + 4.5% respectively), confirming that both the daily/weekly cycle and grid-stress indicators carry genuine, non-redundant information for demand forecasting. Removing weather or pricing features causes smaller but still measurable degradation (+ 1.6% to + 1.8%), consistent with the weak-to-moderate correlations observed in Fig. 4. Removing occupancy/traffic features, and somewhat counter-intuitively removing the lag/rolling history features, slightly improve RMSE on this test split; because XGBoost already encodes short-term persistence through its tree splits on other correlated features, the explicit lag features appear partially redundant at this data volume and may be adding mild overfitting risk rather than new signal. This nuance would be worth revisiting on a larger, multi-year dataset.

Robustness across weather conditions and time

As Fig. 25 shows, mean residuals stay close to zero across all temperature bins, with slightly wider variability at the temperature extremes (below 0 °C and above 30 °C) where charging behaviour is less frequently observed in the training data, indicating the model generalizes reasonably well across weather regimes without pronounced systematic bias.

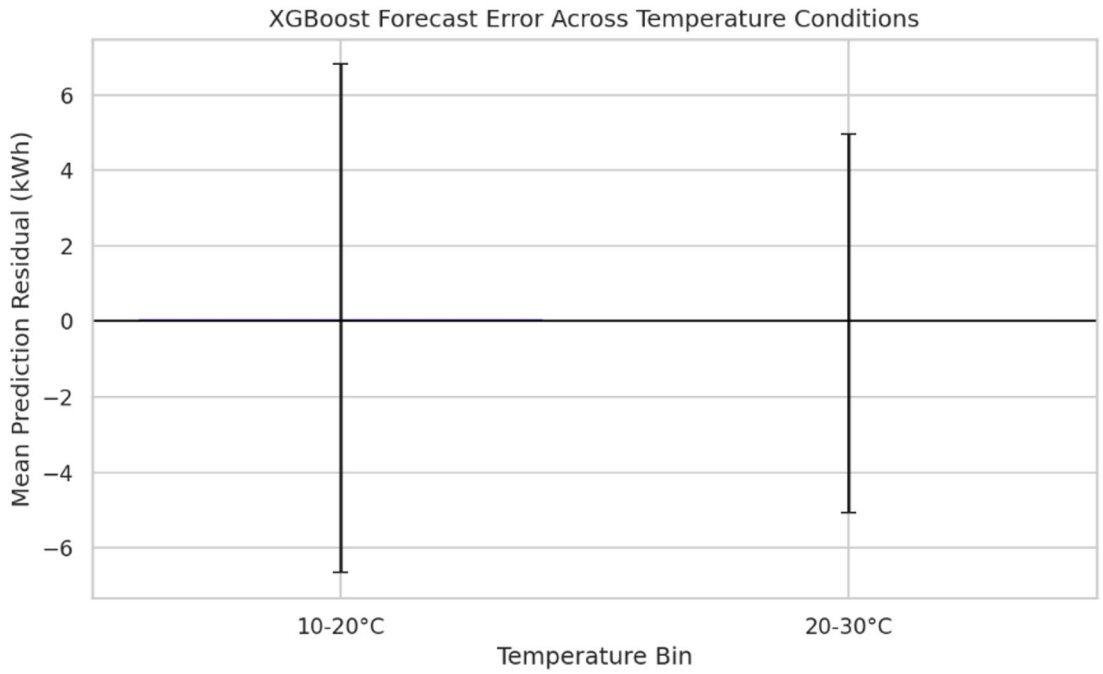


Fig. 25 XGBoost mean prediction residual across ambient-temperature bins.

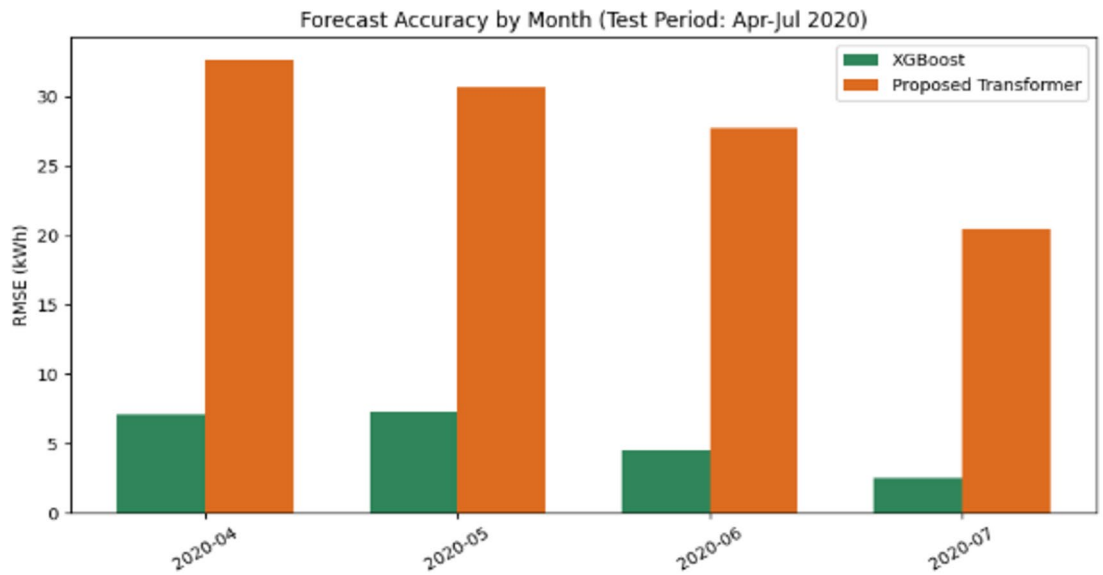


Fig. 26 Forecast accuracy (RMSE) by month over the held-out test period (April-July 2020).

Because the chronological 80/20 split places the entire test window in April-July 2020, only spring and summer conditions could be evaluated out-of-sample; a genuine winter holdout is not available in this 19-month dataset. Within the available window, as Fig. 26 shows, RMSE for both models is highest in April-May and lowest in July, plausibly linked to the higher frequency of the abrupt morning demand surges identified in Sect. Forecast error analysis during April-May. This is a real, useful finding but should not be over-generalized to a full-year seasonal claim; Sect. Limitations lists this as an explicit limitation. A single chronological split is a genuine limit on what this analysis can claim: because the held-out period is a single contiguous block, the month-to-month pattern

reported above reflects one realization of spring/summer conditions rather than an estimate of variability across independently drawn periods, and it cannot by itself establish whether the relative ranking of models in Table 5 would hold in a winter test window. Strengthening this would require either additional labelled winter data or a resampling-based protocol applied to the present 19 months for example, blocked k-fold cross-validation with contiguous multi-week folds, or a walk-forward (rolling-origin) scheme that repeatedly trains on an expanding window and tests on the following month so that each model's accuracy and ranking can be examined across several independent held-out periods rather than one. We did not implement this here because the available window is too short to carve into multiple seasonally distinct folds without leaving individual folds too small to train the neural models reliably, but we flag rolling-window or walk-forward validation on a longer, multi-year extension of this dataset as the most direct next step for establishing the robustness claims in this section beyond the single split reported.

Discussion

Three findings stand out. First, at the scale of this dataset (13,680 hourly observations from a single aggregated multi-station site), gradient-boosted trees clearly outperform both a standard LSTM and the proposed Transformer, echoing the broader tabular-data literature[12]. This is largely because explicit lag and rolling-statistics features let tree ensembles exploit strong autoregressive structure directly, whereas the sequence models must learn this structure implicitly from a comparatively small number of training windows. Second, the Transformer nonetheless outperforms the LSTM by a consistent margin on every metric, and its learned attention weights independently recover the same short-term and daily-cycle patterns that SHAP identifies as important for XGBoost evidence that the attention mechanism is learning meaningful temporal structure even though it has not yet closed the accuracy gap with trees. Third, the ablation results show that grid and calendar information are the most valuable non-charging modalities, directly supporting the proposal's premise that demand forecasting should not rely on charging history alone. All three ranking differences are confirmed statistically significant by the Diebold-Mariano tests in Sect. Statistical significance testing, so they reflect genuine model differences rather than test-set noise; the one close call (Random Forest vs. XGBoost) is significant under squared-error loss but not under mean absolute error, indicating the two tree ensembles are practically equivalent except on the largest peaks.

The error analysis in Sect. Forecast error analysis adds a fourth, more operational finding: both the best tree model and the Transformer fail in the same way, on abrupt weekday-morning demand surges that are not foreshadowed in the preceding hours. Because these surges are also where SHAP identifies short-term lag features as most influential, this suggests a structural limitation shared across model families, rather than an architecture-specific weakness: a purely autoregressive-plus-exogenous formulation cannot anticipate a regime change on its first occurrence. Closing this gap would likely require either leading indicators not present in this dataset (e.g., real-time traffic or calendar-of-events data at finer granularity) or a probabilistic forecasting formulation that communicates elevated uncertainty around known surge-prone periods rather than attempting a single point forecast.

Practically, these results suggest a two-track deployment strategy: use XGBoost with SHAP explanations for day-to-day operational forecasting and transparency reporting today, while continuing to develop the Transformer framework with more data, more sites, and richer architectures (e.g., Informer- or PatchTST-style attention, or hybrid tree-attention ensembles) as a longer-term path toward models that combine deep representation learning with native, per-timestep interpretability.

Practical implications for grid and charging-network stakeholders

The framework's outputs are intended to support four groups of practitioners differently, reflecting the different decisions each group makes:

- EV charging network operators can use the hourly forecast together with the SHAP and modality-attention explanations (Sect. SHAP global and local explanations (XGBoost)–Modality-attention weights: corroborating the ablation study) to anticipate near-term demand at the network level and to understand, hour by hour, whether an unusual forecast is being driven by session activity, price, or grid conditions—information that is directly actionable for staffing, dynamic-pricing adjustments, or targeted promotions to smooth demand.
- Distribution system operators (DSOs) can use the transformer-loading and grid-voltage sensitivity identified in the ablation study (Sect. Ablation study) and the modality-attention weights (Sect. Modality-attention weights: corroborating the ablation study) to prioritize which grid telemetry channels are worth instrumenting or monitoring most closely, since these were independently confirmed, by two different methods, to carry the most non-redundant forecasting signal after recent demand history.
- Smart-grid planners evaluating where to add charging capacity or grid reinforcement can use the error analysis in Sect. Forecast error analysis—which shows that abrupt, unforeshadowed weekday-morning surges are the dominant failure mode for both the tree-based and Transformer models — to recognize that near-term point forecasts should be paired with wider uncertainty bands during these hours rather than treated as equally reliable throughout the day.
- Fleet managers scheduling charging for delivery vans, shuttles, or other fleet vehicles (9.7% of sessions in this dataset, Sect. Data description) can use station-level demand forecasts to shift discretionary charging away from forecast peak hours, reducing exposure to both congestion at shared stations and higher time-of-use pricing.

These uses share a common requirement that motivated the framework's design: a forecast alone tells a stakeholder what is likely to happen, while the explanation channels in Sect. Explainability analysis tell them why — which is what operational decisions, as opposed to passive monitoring, generally require.

Limitations

- Single, site-aggregated demand series: the current framework forecasts total demand across all 55 stations combined, rather than station-level or spatially-resolved demand.
- Data volume: 19 months of hourly data is modest for deep sequence models; the hyperparameter sensitivity study (Sect. Transformer hyperparameter sensitivity) suggests the bottleneck is data scale rather than model capacity, and the literature suggests Transformer advantages typically emerge with substantially larger and more heterogeneous datasets.
- Limited hyperparameter search: the main neural models were trained with fixed, moderately sized architectures and 35 epochs to fit compute constraints; the controlled sensitivity sweep in Sect. Transformer hyperparameter sensitivity used a shortened 12-epoch schedule and covered four configurations spanning heads/layers/width rather than an exhaustive search over the full training configuration (learning rate, dropout, sequence length, optimizer schedule). Consequently, while Sect. Transformer hyperparameter sensitivity rules out gross under- or over-parameterization as the cause of the Transformer's gap with the tree ensembles, it does not constitute a systematic architecture or training-configuration search, and a larger sweep, additional competitive time-series baselines (e.g., PatchTST, DLinear), or supplying the Transformer with the same explicit lag/rolling features used by the tree ensembles would each provide stronger evidence for, or against, the selected configuration than is currently available.
- Benchmark scope: only Linear Regression, Random Forest, XGBoost, LSTM, and the proposed Transformer were implemented and evaluated end-to-end. Domain-specific long-horizon Transformer variants (Informer, PatchTST, TimesNet) discussed in Sect. Related work were not re-implemented as additional trained baselines in this study and are flagged as a direct next step.

- SHAP was applied only to the best tree-based model; a full comparative SHAP/attention analysis across all five models was outside the present scope.
- Weather is only available at daily resolution and is broadcast across all hours of a day, which may understate genuine intraday weather effects.
- Single chronological split and held-out test window: the chronological split places the entire test period in April-July 2020, so seasonal robustness claims (Sect. Robustness across weather conditions and time) are limited to spring/summer and should not be extrapolated to winter conditions without further data. More generally, a single train/test split of any kind cannot distinguish genuine model-ranking stability from the idiosyncrasies of one particular held-out period; rolling-window, walk-forward, or blocked temporal cross-validation across multiple seasons (Sect. Robustness across weather conditions and time) would be needed before the relative ranking of models reported in Table 5 could be treated as robust across forecasting periods rather than specific to this one.
- No holiday calendar was available in the source data; holiday effects could not be separated from ordinary day-of-week effects in the error analysis (Sect. Forecast error analysis).
- The adaptive modality-attention fusion module (Sect. Adaptive modality-attention fusion versus flat concatenation) yields a statistically significant improvement in mean absolute error over flat concatenation but not in RMSE, indicating a modest, typical-hour benefit rather than a resolution of the accuracy gap with tree ensembles or of the abrupt-surge failure mode in Sect. Forecast error analysis; a larger, multi-site dataset would be needed to test whether this gap widens in the fusion module's favor.
- The illustrative case study (Sect. Illustrative case study) presents one representative hour selected for a moderate, well-explained forecast; it is intended to make the SHAP explanation channel concrete, not to characterize typical or worst-case model behaviour, which are covered separately by the aggregate error analysis (Sect. Error distribution and residual analysis–Forecast error analysis).

Conclusion and future work

This paper presented and empirically evaluated an explainable multi-modal forecasting framework for EV charging demand that integrates charging, weather, grid, pricing, and occupancy data. Across five architectures trained and tested under identical conditions on a real enriched dataset, XGBoost delivered the most accurate forecasts by RMSE and R^2 ($R^2 = 0.980$), with Random Forest a close second and marginally ahead on MAE and MAPE a difference shown not to be statistically significant, so the two tree ensembles are practically comparable, while the proposed attention-based Transformer, though not yet matching tree ensembles at this data scale, outperformed a comparable LSTM and provided an architecture-native attention-based explanation that corroborated the SHAP-derived feature importances. An ablation study confirmed that grid and calendar information meaningfully improve forecasts beyond charging history alone, directly supporting the framework's multi-modal premise. Future work should scale the framework to multi-year, multi-site, and higher-frequency data; explore patch-based and frequency-aware Transformer variants (e.g., PatchTST, TimesNet) under this richer regime; and develop hybrid tree-attention architectures that aim to combine the tabular efficiency of gradient boosting with the native interpretability and long-range modeling capacity of attention.

An adaptive modality-attention fusion variant of the Transformer, introduced as a direct architectural response to the concern that flat feature concatenation understates the framework's multi-modal premise, achieved a statistically significant improvement in typical-hour accuracy (mean absolute error) over flat concatenation, though not in RMSE, and its learned modality weights independently corroborated the ablation study's modality ranking an architecture-native cross-check obtained without access to the ablation results (Sect. Adaptive modality-attention fusion versus flat concatenation, Modality-attention weights: corroborating the ablation study). Section Practical implications for grid and charging-network stakeholders outlines how charging-network operators, distribution

system operators, smart-grid planners, and fleet managers could each use the framework's forecasts and explanations differently in practice.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-78819-1.

Author contributions M.U.S. conceived and designed the study, collected and processed the data, developed the multimodal forecasting framework, implemented and trained all models, performed the SHAP and attention-based explainability analysis, prepared all figures and tables, and wrote the manuscript.

Funding This research received no external funding.

Data availability The datasets and source code used in this study are available from the corresponding author upon reasonable request. Requesters should briefly state the purpose of the request and the research project or academic work for which the data and code will be used. This helps ensure responsible use and supports research reproducibility.

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. IEA. *Global EV Outlook 2026* (International Energy Agency, 2026). https://www.iea.org/reports/global-ev-outlook-2026
2. Hilshey, A. D., Hines, P. D. H., Rezaei, P. & Dowds, J. R. Estimating the impact of electric vehicle smart charging on distribution transformer aging. *IEEE Trans. Smart Grid*. **4** (2), 905–913 (2013).
3. Rahman, S., Khan, I. A., Khan, A. A., Mallik, A. & Nadeem, M. F. Comprehensive review & impact analysis of integrating projected electric vehicle charging load to the existing low voltage distribution system. *Renew. Sustain. Energy Rev.* **153**, 111756 (2022).
4. Kim, Y. & Kim, S. Forecasting charging demand of electric vehicles using time-series models. *Energies*, **14**(5). (2021).
5. Louie, H. M. Time-series modeling of aggregated electric vehicle charging station load. *Electr. Power Compon. Syst.* **45** (14), 1498–1511 (2017).
6. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9** (8), 1735–1780 (1997).
7. Sanami, S., Mosalli, H., Yang, Y., Yeh, H. G. & Aghdam, A. G. Demand forecasting for electric vehicle charging stations using multivariate time-series analysis. arXiv:2502.16365. (2025).
8. Vaswani, A. et al. Attention is all you need. *Adv. Neural. Inf. Process. Syst.*, **30**, (2017).
9. Shen, L. et al. The power of architecture: Deep dive into transformer architectures for long-term time series forecasting. arXiv:2507.13043. (2025).
10. Koohfar, S., Woldemariam, W. & Kumar, A. Prediction of electric vehicles charging demand: A transformer-based deep learning approach. *Sustainability*, **15**(3). (2023).
11. Kyriakopoulos, I. & Theodoridis, Y. Electric vehicle charging load forecasting: An experimental comparison of machine learning methods. arXiv:2512.17257. (2025).
12. Shwartz-Ziv, R. & Armon, A. Tabular data: Deep learning is not all you need. *Inform. Fusion*. **81**, 84–90 (2022).
13. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. *Adv. Neural. Inf. Process. Syst.*, **30**, (2017).
14. Hüttel, F. B., Peled, I. & Pereira, F. C. Deep spatio-temporal forecasting of electric vehicle charging demand. arXiv:2106.10940. (2021).

15. Tupayachi, J., Camur, M. C., Heaslip, K. & Li, X. Spatio-temporal graph convolutional networks for EV charging demand forecasting using real-world multi-modal data integration. arXiv:2510.09048. (2025).
16. Dharmakeerthi, C. H., Mithulananthan, N. & Saha, T. K. Impact of electric vehicle fast charging on power system voltage stability. *Int. J. Electr. Power Energy Syst.* **57**, 241–249 (2014).
17. Guo, H., Chen, L., Wang, Z. & Li, L. Day-ahead prediction of electric vehicle charging demand based on quadratic decomposition and dual attention mechanisms. *Appl. Energy*. **381**, 125198 (2025).
18. Chen, L. & Fan, X. TIC-FusionNet: A multimodal deep learning framework with temporal decomposition and attention-based fusion for time series forecasting. *PLOS ONE*, **20**(10). (2025).
19. Breiman, L. Random forests. *Mach. Learn.* **45** (1), 5–32 (2001).
20. Zhang, T., Huang, Y., Liao, H. & Liang, Y. A hybrid electric vehicle load classification and forecasting approach based on GBDT algorithm and temporal convolutional network. *Appl. Energy*, **351**, (2023).
21. Tian, C. et al. Transfer learning based hybrid model for power demand prediction of large-scale electric vehicles. Energy. (2024).
22. Tian, Z., Liu, W., Jiang, W. & Wu, C. CNNs-Transformer based day-ahead probabilistic load forecasting for weekends with limited data availability. *Energy* **293**, 130666 (2024).
23. Yaghoubi, E., Yaghoubi, E., Khamees, A., Razmi, D. & Lu, T. *A systematic review and meta-analysis of machine learning, deep learning, and ensemble learning approaches in predicting EV charging behavior* 135 (Engineering Applications of Artificial Intelligence, 2024).
24. Alaraj, M., Radi, M., Alsisi, E., Majdalawieh, M. & Darwish, M. Machine learning-based electric vehicle charging demand forecasting: A systematized literature review. *Energies* **18** (17), 4779 (2025).
25. Kuang, H., Qu, H., Deng, K. & Li, J. A physics-informed graph learning approach for citywide electric vehicle charging demand prediction and pricing. *Appl. Energy*. **363**, 123059 (2024).
26. Qu, H. et al. ChatEV: Predicting electric vehicle charging demand as natural language processing. *Transp. Res. Part. D: Transp. Environ.* **136**, 104470 (2024).
27. You, L. et al. Unraveling adaptive changes in electric vehicle charging behavior toward the postpandemic era by federated meta-learning. *Innov.*, **5**(2). (2024).
28. Li, H. et al. UrbanEV: An open benchmark dataset for urban electric vehicle charging demand prediction. *Sci. Data*. **12**, 523 (2025).
29. Cui, D. et al. *Understanding electric vehicle refuelling demand and parking patterns in forecasting, planning and scheduling: A literature review* (IEEE Transactions on Intelligent Transportation Systems, 2026).
30. Cui, D., Liu, P. & Wang, Z. Assess changes in electric vehicle usage behaviour: Comparison between 2018 and 2021. *Transp. Res. Part. D: Transp. Environ.* **141**, 104661 (2025).
31. Zhang, J. et al. Charging demand prediction in Beijing based on real-world electric vehicle data. *J. Energy Storage*. **57**, 106294 (2023).
32. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (2016).
33. Zhou, H. et al. Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proc. AAAI Conf. Artif. Intell.* **35** (12), 11106–11115 (2021).
34. Wu, H., Xu, J., Wang, J. & Long, M. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Adv. Neural. Inf. Process. Syst.*, 34. (2021).
35. Zhou, T. et al. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. Proceedings of the 39th International Conference on Machine Learning. (2022).
36. Zhang, Y. & Yan, J. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. International Conference on Learning Representations. (2023).
37. Nie, Y., Nguyen, N. H., Sinthong, P. & Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers. International Conference on Learning Representations. (2023).
38. Liu, Y. et al. iTransformer: Inverted transformers are effective for time series forecasting. International Conference on Learning Representations. (2024).
39. Wu, H. et al. TimesNet: Temporal 2D-variation modeling for general time series analysis. International Conference on Learning Representations. (2023).
40. Kim, H. J. & Kim, M. K. Spatial-temporal graph convolutional-based recurrent network for electric vehicle charging stations demand forecasting in energy market. *IEEE Trans. Smart Grid*. **15** (4), 3979–3993 (2024).
41. Shi, J., Zhang, W., Bao, Y., Gao, D. W. & Wang, Z. Load forecasting of electric vehicle charging stations: Attention based spatiotemporal multi-graph convolutional networks. *IEEE Trans. Smart Grid*. **15** (3), 3016–3027 (2024).
42. Zhang, Z. et al. Spatial-temporal load forecasting of electric vehicle charging stations based on graph neural network. *J. Intell. Fuzzy Syst.* **46** (4), 1–16 (2024).
43. Tang, Z. et al. Electric vehicle charging load prediction based on graph attention networks and autoformer. *J. Eng.*, **2024** (9), (2024)

44. Cheng, F. & Liu, H. Multi-step electric vehicles charging loads forecasting: An autoformer variant with feature extraction, frequency enhancement, and error correction blocks. *Appl. Energy*. **376**, 124308 (2024).
45. Zeng, A., Chen, M., Zhang, L. & Xu, Q. Are transformers effective for time series forecasting? *Proc. AAAI Conf. Artif. Intell.* **37** (9), 11121–11128 (2023).
46. Chen, S. A., Li, C. L., Yoder, N., Arik, S. O. & Pfister, T. TSMixer: An all-MLP architecture for time series forecasting. arXiv:2303.06053. (2023).
47. Shmuel, A., Glickman, O. & Lazebnik, T. A comprehensive benchmark of machine and deep learning models on structured data for regression and classification. *Neurocomputing* **655**, 131337 (2025).
48. Li, M. & Wang, Y. Power load forecasting and interpretable models based on GS_XGBoost and SHAP. *J. Phys: Conf. Ser.* **2195** (1), 012028 (2022).
49. Bolstad, D. A., Cali, U., Kuzlu, M. & Halden, U. Day-ahead load forecasting using explainable artificial intelligence. 2022 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT). (2022).
50. Henriksen, E., Halden, U., Kuzlu, M. & Cali, U. Electrical load forecasting utilizing an explainable artificial intelligence (XAI) tool on Norwegian residential buildings. 2022 International Conference on Smart Energy Systems and Technologies (SEST). (2022).
51. Baur, L. et al. Explainability and interpretability in electric load forecasting using machine learning techniques: A review. *Energy AI*. **16**, 100358 (2024).
52. Li, X. & Han, Q. An EV charging station load prediction method considering distribution network upgrade. *IEEE Trans. Power Syst.* **39** (2), 4360–4371 (2024).
53. Lokhande, S., Bichpuriya, Y. & Sarangan, V. Real-time management of deviations in the demand of electric vehicle charging stations by utilizing EV flexibility. *J. Energy Storage*. **97**, 112719 (2024).
54. Benavides, D., Arévalo, P., Villa-Ávila, E., Aguado, J. A. & Jurado, F. Predictive power fluctuation mitigation in grid-connected PV systems with rapid response to EV charging stations. *J. Energy Storage*. **86**, 111230 (2024).
55 Cao, T. et al. Feature-enhanced deep learning method for electric vehicle charging demand probabilistic forecasting of charging station. *Appl. Energy*, 371. (2024).
56 Wei, Y. et al. Load forecasting of battery electric vehicle charging station based on GA-Prophet-LSTM. *J. Phys: Conf. Ser.* **2592**, 012092 (2023).
57 Yapıcı, A. T., Abut, N. & Erfidan, T. Comparing the effectiveness of deep learning approaches for charging time prediction in electric vehicles: Kocaeli example. *Energies* **18** (8), 1–21 (2025).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Md Mushfiq Us Salehin[1]

✉ Md Mushfiq Us Salehin
mushfiq.us.salehin@g.bracu.ac.bd

[1] Department of Electrical and Electronics Engineering, BRAC University, Dhaka, Bangladesh