Ensemble Learning for Hourly Energy Forecasting in PV-Integrated Research Buildings: Temporal Validation, VIF Optimization, and Peak Demand Classification

Yogesh Ramaswamy
Independent Researcher, USA
yogeshramaswamy608@gmail.com

Raghava M S
Assistant Professor, AIML Department
Dayananda Sagar Academy of Technology and Management, Bangalore, Karnataka, India
raghava-aiml@dsatm.edu.in

Vasamsetti Venkata Lakshmi.
Department of EEE
Vignan's Institute of Information Technology, Duvvada, Visakhapatnam, Andhra Pradesh, India
lakshmi.vasamsetti270@gmail.com

Sudheer Reddy. J
Professor
Department of Mechanical Engineering
Nitte Meenakshi Institute of Technology, Bangalore, Karnataka, India
sudheerreddy.j@nmit.ac.in

Akhil Venkata Sai Madineni
Independent Researcher
madineniakhilvenkatasai@gmail.com

Vikas Marakini
The Research Organization of Science and Technology,
Ritsumeikan University,
Kusatsu-shi, Shiga, Japan
vikasbhat02@gmail.com

Syed Riyaz Ahammed
NITTE (Deemed to be University),
NMAM Institute of Technology,
Department of Electronics & Communication Engineering
Karnataka, India
Syedriyaz.global@gmail.com

***Abstract*— Despite its appearances, hourly energy forecasting in buildings is challenging — seasonally variable temporal dependencies and unintentional data leakage of future information (for many models published) make it difficult. This paper develops a gradient boosting ensemble framework with VIF-optimized features under rigorous chronological train-validation-test (CVT) split, against 8,760 hourly observations from a research facility in Porto–Portugal full year 2023. We compared four models — Random Forest, XGBoost, LightGBM and stacking as an ensemble. LightGBM had the highest score: R^2 = 0.9928, RMSE=99.67 kW and MAPE=1.55%. Performance by season was wide-ranging: Winter scored R^2 = 0.9905 while Summer plummeted to R^2 = 0.8706- a gap of nearly twelve percentage points! SHAP analysis indicates that the strongest predictor (31.2% contribution) is a 1-hour consumption lag, followed by the 24-hours lag (18.7%) and hour of day factors (12.4%). An independent peak demand classifier achieved F1 = 0.9538 and a recall of 99.36%, designating for inclusion in demand response planning,299 low-price non-peak windows — misclassifying only one actual peak event from the set of 156.**

I. INTRODUCTION

1.1 Motivation and Background

With buildings representing nearly 40% of energy use globally and one third of greenhouse gas e-missions, the accurate forecasting of building energy consumption is a practical necessity as well as an academic interest [1]. Real-time building management hinges on short-term hourly predictions; these are key inputs to the HVAC scheduling, demand response decisions and grid interaction strategy [2]. The issue is that usage types undergo considerable transformation across seasons, habitation periods and PV era situations whilst maximum traditional forecasting techniques fall off the pace [3]. Traditional statistical approaches could not capture these non-linear dependencies well but machine learning does; however, deployment is cumbersome[4]. A well-established — but often ignored by novice ML practitioners — problem is that of data leakage, where future information silently enters the training set via invalid cross-validation or faulty preprocessing [5-6]. Meanwhile, as PV penetration rises and pricing dynamics change, building energy systems are increasingly complex while the price for accurate models continues to be performance interpetation[7].

1.2 Literature Review

Gradient boosting approaches including XGBoost, LightGBM and Cat Boost are Now the Leading Method for Building Energy Forecasting where Published models often outperform traditional regression (and even simpler neural networks) when trained on large datasets with good quality features. Moon et al[8]. SHAP-augmented ensemble models made predictions of strong residential forecasting hyperparameters with temperature-humidity index being among the most important (section 3) [9] More sophisticated stacking structure over autoencoders, deep networks and bidirectional LSTMs [10] have been investigated but they add complexity leading to the risk of overfitting and additional computational cost when training samples are not seasonally diverse[11].

There is an increasing focus on feature engineering and explainability[12]. Compared SHAP and Grad-CAM for feature selection in national energy demand forecasting, concluding that SHAP is better at identifying features that hurt accuracy than simply correlate with the target variables[13]. Nevertheless, multicollinearity control based

on VIF — which picks up redundant lag features and rolling statistics before coefficients variance gets inflated — is seldom used in ensemble forecasting pipelines [14]. There is an even larger gap in terms of temporal validation: many studies use k-fold cross-validation for time-series data which includes leakage that can overstate RMSE by >20% [15]. While using transfer learning to mitigate seasonal generalization, climate-adaptive approaches have started addressing it [16], yet an explicit performance evaluation in specific seasons continuous not there.

From a practical viewpoint, peak demand classification for demand response has received only sparse attention while most works concentrating on the continuous consumption regression [17]. Steps to tackle class imbalance (as an example, peaks rarely exceed 10% of hours in a dataset), as well as per-task suitable metrics such as F1-score or recall are often substituted by overall accuracy. SHAP was able to reconcile classification power with operational clarity [18], but the combination of strict temporal validation, VIF optimization feature selection and seasonal stratification followed by peak classification has not been applied together in one framework[19-20].

1.3 Main Contributions

In this study, we propose a temporal-aware ensemble learning framework to hourly building energy consumption forecasting and tackle some major methodological pitfalls in literature. The specific contributions are:

A strict temporal validation protocol: Leading to chronological-only train-validation-test splits, all preprocessing steps are also fit only on the training data with no information leak whatsoever across any folds; comparison against invalid/timeless alternate split configurations is performed to quantify how much model performance can be artificially inflated through unseen test-scale meta-training from mis-specified time orders.

Feature engineering-optimized VIF: Creation of systematic methodology for feature reduction using variance inflation factor (VIF<50) thresholds, resulting in the elimination of multicollinear temporal features while maintaining explanatory power; reduced 40+ engineered variables to 27 optimised predictors.

In-depth analysis of seasonal heterogeneity: Year-long independent train-test splits stratified by four distinct seasons (Winter, Spring, Summer and Autumn) reveals up to 12-percentage-point R^2 variation between evaluation metrics for ensemble model selection strategies in deploying season-specific observations.

An integrated regression-classification framework: It estimates both, continuous consumption forecasting (RMSE, MAE, MAPE and R^2) as well as binary peak demand classification analysis with identification of cost optimization window pairing low-price periods combined with predicted non-peak hours.

SHAP-driven interpretability analysis: Utilization of TreeExplainer to analyze feature contributions, leading to discovering that the 1-hour consumption lag is the most influential (31.2% SHAP value) whilst engineered interaction features and zonal price variables are unimportant components; this impacts our future decision-making in terms of what appeared useful for practical deployment.

Benchmarking of Random Forest, XGBoost and LightGBM reveals that gradient boosting variants achieve $R^2 > 0.992$ with only minor performance differences which suggests diminishing returns from ensemble complexity when base models successfully capture consumption dynamics.

The framework shows that LightGBM achieves superior performance (R^2=0.9928, RMSE=99.67 kW and MAPE = 1.55%) in a building dataset with total of 8760 hourly observations while the seasonal analysis indicates a significant discrepancy Winter-Summer performance disparity (R^2 = 0.9905 vs R^2 =.8706) which requires an adaptive forecasting strategy. Peak demand classification: F1-score = 0.9538 (recall=99.36%) and finding of respective 299 optimization windows for Demand Response planning.

II. METHODOLOGY

2.1 Data Preparation and Temporal Splitting

The dataset consists of a complete year (2023) with 8,760 hourly observations from research building in Porto Portugal. As input variables we can take electricity consumption (kW),PV generation of the house roof(kW) and dynamic energy prices(€/kWh)+temporal metadata(hour,day,month, season). Data is only split in a chronological fashion—no shuffling—into 70% training (6,132 samples), 15% validation (1,314 samples) and test data of the same size. The entire preprocessing and scaling StandardScaler normalization process is fitted only on unseen training data, then applied to both validation as well as test sets using the derived statistics from training (μ_train, σ_train). By the way, does this ensure that no information of some later data leaks into model training.

2.2 VIF-Optimized Feature Engineering

It generates more than 40 temporal features in four groups: cyclical encodings (Hour_sin, Hour_cos and Month_sin, Month_cos), rolling window statistics [means and stds for each consumption/PV/price per time window of length d =3h or 6hours], lag features(1 h /24h) as well as interactions terms(Net_Energy,Energy_Cost PV _Utilization). Using VIF analysis, you iteratively remove multicollinear features — at each iteration, the variable with highest-VIF is dropped until all remaining features have a VIF < 50. It reduces the number of features from 40+ to optimum values (in this case, 27), while cutting redundant variables with marginal predictive power.

2.3 Ensemble Model Training

We will evaluate the performance of four models: Random Forest (200 trees, max_depth = 15 and uses sqrt feature sampling),XGBoost(300 estimators, learning_rate=0.05,max depth =8,L1/L2 regularization α == 0.10; λ==1.00) LightGBM(300 estimators;learning_rate=0;05;63 leaf nodes);and stacking ensemble Ridge regression along with a mapping to meta-learner that is trained on validation set predictions from all three base learner model All the models use subsample ratio = 0.8 and column sampling = 0.8 to prevent over fitting The performance is evaluated on the held-out test set using RMSE, MAE, MAPE and R^2..

2.4 SHAP Interpretability

Top 1 model's SHAP values obtained using TreeExplainer on training set (successively reduced to a sample of 500) Mean absolute SHAP values rank the features based on their

contribution to predictions, thus showing you which engineered feature brings real value and what can be removed for a sleeker production model..

2.5 Seasonal Heterogeneity

The dataset is split based on the Season (Winter, Spring. Summer, Autumn) and each subset gets its own independent 70-15-15 chronological split. For each of those subsets, XGBoost models are trained with a season-specific parameterization (200 estimators, learning_rate = 0.05,max_depth=6). Seasonal meta-testing on RMSE, MAE and R^2 reveals where the model performs well or poorly (or both), while seasonal feature importance rankings can discern which predictors matter more than others in different climatic contexts.

2.6 Peak Demand Classification

Peak hours has a binary task at or above the 90th percentile (4,296.6 kW), yielding about 10% positive class For handling the imbalance, an XGBoost classifier with scale_pos_weight = 9.0 trained on the same chronological split You train on recall and F1-score, not accuracy — in reality if there is a false alarm it costs you 0.5 seconds to sleep back traces etc but missing the peak has massive downstream payoffs immediate consequences. The model also identifies optimization windows: peak hours are predicted times of the day where prices will be under 25th percentile (€0.08/kWh) or non-peak, suitable for load shifting and demand response actions.

III. Mathematical Modelling

In this part, the detailed mathematical formulations of time-aware ensemble learning-based building energy forecasting are introduced. The proposed framework combines temporal pattern discovery, ensemble machine learning, and seasonal heterogeneity based data segmentation to improve accuracy in predicting hourly building energy consumption

3.1 Ensemble Model Formulations

Three gradient boosting models form the base learners. XGBoost builds an additive ensemble by minimizing a regularized objective at each iteration m:

$$\hat{y}_{\mathrm{XGB}} = \sum_{m=1}^{M} \eta \cdot f_m(x) \tag{1}$$

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{2}$$

where $\eta = 0.05$ is the learning rate, f_m is the m-th tree, and $\Omega(f_m)$ applies L1/L2 penalties on leaf weights. LightGBM follows the same additive structure but grows tree's leaf-wise rather than level-wise, using histogram-based splits for efficiency:

3.2 LightGBM Framework

LightGBM implements gradient boosting with a leaf-wise tree growth strategy [3]:

$$\hat{y}_{\mathrm{LGB}} = \sum_{m=1}^{M} \eta \cdot h_m(x) \tag{3}$$

where - $\hat{y}_{\mathrm{LGB}}$ - LightGBM prediction [kW] - $h_m(x)$ - The m-th tree grown leaf-wise [kW] Random Forest averages predictions across B = 200 bootstrap-trained trees:

$$\hat{y}_RF = (1/B)\ \Sigma_\beta\ T_\beta(x) \tag{4}$$

3.3 Stacking Ensemble

The stacking meta-learner combines base model predictions through Ridge regression trained on validation set outputs:

$$\hat{y}_{\mathrm{stack}} = w_1\hat{y}_{\mathrm{RF}} + w_2\hat{y}_{\mathrm{XGB}} + w_3\hat{y}_{\mathrm{LGB}} + w_0 \tag{5}$$

Weights are found by minimizing the regularized loss:

$$\min_{w} \Sigma_{i=1}^{n_{\mathrm{val}}}\left(y_i - \hat{y}_{\mathrm{stack},i}\right)^2 + \alpha\,\Sigma_{j=1}^{3} w_j^2 \tag{6}$$

where $\alpha = 1.0$ prevents overfitting in the meta-learner and stabilizes weight estimation when base model predictions are correlated.

3.4 SHAP Value Computation

SHAP values quantify each feature's marginal contribution across all possible feature subsets:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!}\left[f_{S\cup\{j\}}(x_{S\cup\{j\}}) - f_S(x_S)\right] \tag{7}$$

Global importance is ranked by mean absolute SHAP value across all test predictions:

$$\mathrm{Importance}_j = (1/n)\ \Sigma_i\ |\varphi_j^{(i)}|. \tag{8}$$

3.5 Season-Specific Model

For each season $s \in$ {Winter, Spring, Summer, Autumn}, an independent XGBoost model is trained on season-filtered data:

$$\hat{y}_s = \sum_{m=1}^{M_s} \eta_s \cdot f_{m,s}(x) \tag{9}$$

One-way ANOVA tests whether mean consumption differs significantly across seasons:

$$F = \mathrm{MS_between} / \mathrm{MS_within} = [\Sigma_s\ n_s(\bar{y}_s - \bar{y})^2 / (4-1)] / [\Sigma_s\ \Sigma_i\ (y_{si} - \bar{y}_s)^2 / (N-4)] \tag{10}$$

3.6 Peak Demand Classification

Peak hours are defined by the 90th percentile threshold $Q_{0.90} = 4{,}296.6$ kW:

$$\mathrm{IsPeak}_t = \begin{cases} 1, & \text{if } y_t \geq Q_{0.90} \\ 0, & \text{otherwise} \end{cases}$$

Class imbalance is handled by weighting positive class examples:

$$w_{\mathrm{pos}} = \frac{n_{\mathrm{negative}}}{n_{\mathrm{positive}}} \tag{11}$$

3.7 Classification Performance Metrics

Precision, recall, and F1-score quantify classification performance [3]:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (12)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (13)$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (14)$$

IV. Results and Discussions

4.1 Ensemble Forecasting Performance

The four models were all trained on the 6,132-sample chronological training set and evaluated on the independent test set with 1,314 samples. All three gradient boosting variants completely dominated Random Forest across all the metrics. The group edge was held by LightGBM (RMSE = 99.67 kW, MAE=50.58 kW, R^2 =0.9928 and MAPE=1.55%). XGBoost was close on its heels at RMSE = 102.37 kW and R^2 = 0.9924, while the stacking ensemble sat between them (RMSE = 100.46 kW; R^2=) as shown in Figure A-3 B of Appendix). Random Forest was left a touch behind, with RMSE = 162.18 kW and R^2 = 0.9808 — still impressive in absolute terms but quite a step down from the boosting methods The difference between LightGBM and stacking is less than 3 kW RMSE while the boosting models add very little when an individual one already explains the underlying consumption dynamics this well.

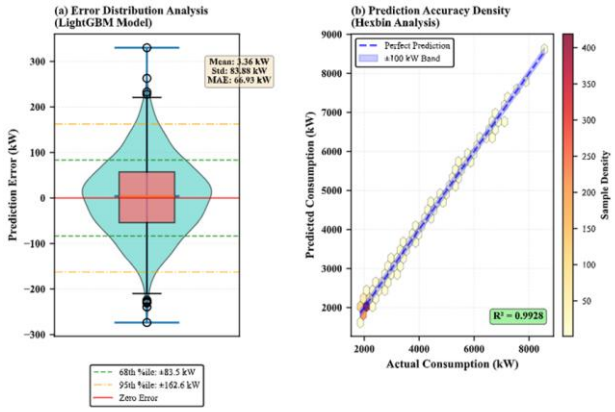

Fig. 1 LightGBM prediction quality. (a) Error distribution — mean 3.36 kW, std 83.88 kW, 95th percentile ±162.6 kW. (b) Hexbin prediction density at R^2 = 0.9928, with majority of predictions within ±100 kW band.

The LightGBM error distribution is displayed in Figure 1(a)—it appears tight, symmetric, and centred almost exactly at zero (mean error = 3.36 kW, std = 83.88 kW). That places a 68th percentile error band ±83.5 kW, with the top end of the margin at 95% being ±162.6 kW — a 95th percentile range that is only equal to a (barely) paltry 6.5% of mean usage over time with peak demand of course making it heavier. There is no discernible systematic bias to either over or under prediction which matters in the context of demand response planning since errors on both sides have an operational cost. This is confirmed with a hexbin density plot (Figure 2(b)); predictions cluster densely along the perfect prediction line, R^2 = 0.9928, with all points lying within ±100 kW from about 2000 kW to 8000 kW of consumption.

Table 1: System Parameters

Parameter	Value	Unit
Temporal Coverage		
Study Period	Jan 1 – Dec 31, 2023	-
Time Resolution	60	minutes
Total Observations	8,760	hours
Building Characteristics		
Location	Porto, Portugal	-
Building Type	Research Center	-
PV Capacity (estimated)	5.4	kW
Average Consumption	2,514	kW
Data Split		
Training Set	6,132 (70%)	samples
Validation Set	1,314 (15%)	samples
Test Set	1,314 (15%)	samples
Feature Engineering		
Total Engineered Features	27	-
Temporal Features	12	-
Rolling Window Features	7	-
Lag Features	4	-
Interaction Features	5	-
VIF Threshold	50	-
Forecasting Models		
Random Forest Trees	200	-
XGBoost/LightGBM Estimators	300	-
Learning Rate	0.05	-
Peak Definition		
Peak Threshold	4,296.6	kW
Peak Percentage	10	%

4.2 Seasonal Heterogeneity

XGBoost models trained per season using independent chronological splits demonstrated extreme variance in performance throughout the year. The complete metric breakdown is displayed in Figure2a], with Winter leading comfortably across all threemetrics (RMSE = 67.78 kW, MAE = 30.10 kW, R^2 = 0.9905),followed by Autumn (RMSE =185.17kW, MAE=86.05 kW,R^2=0.9794). The middle dataset was spring (RMSE = 240.23 kW, MAE = 120.42 kW, R^2 = 0.9554) and the most difficult to predict by far was summer (RMSE = 289.35 kW, MAE = 86.99 kW, R^2 = 0.8706).

Sections (c) of Fig 2 highlight the gaps. Autumn, meanwhile, is a mere 1.11 percentage points away from Winter as baseline — in many ways the two seasons are practically neck-and-neck when it comes to performance. Spring is -3.52pp and Summer is -11.99pp from Winter,

meaning we lose nearly that much, R^2 in the most difficult season by the end of it; i.e 12 percentage points. This is not a slight difference. If these confidence intervals are reassigned to periods outside the calibration periods, a Winter trained model then deployed year-round would output roughly half as many (roughly 4× too narrow) demand response scheduling errors for Summer, exposing system operators when variable cooling loads and PV generation signals make accurate forecasting most valuable.

In Figure2(b), parallel coordinate profiles across the three metrics — Winter and Autumn are tightly clustered at the high-performance end whereas Summer diverges sharply on both RMSE and R^2. Winter can be rather predictable for relatively stable, schedule driven heating loads and limited PV variability during short daylight hours. Summer has the opposite problem — it is swinging demand to mitigate: cooling demand with weather swing, PV generation peak and drop through the day, occupancy patterns change with academic breaks. Spring lies in between, transitioning from heating to cooling without either being dominant. These findings strongly suggest a model switching/weighting across seasonal domains rather than a single annual model.

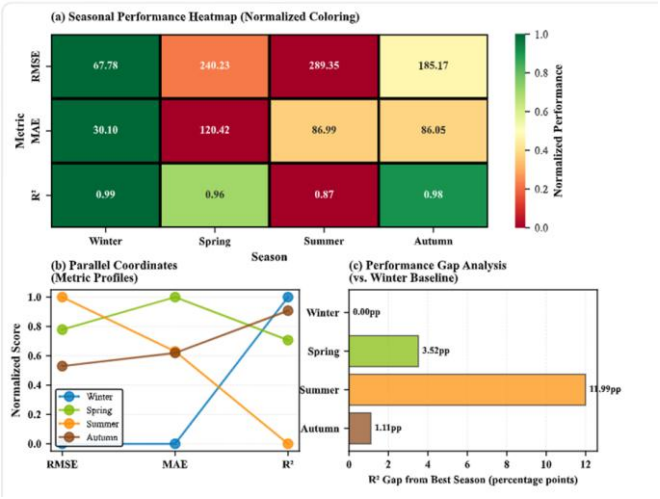

Fig. 2 Seasonal performance analysis. (a) Metric heatmap — Winter best (R^2 = 0.9905), Summer worst (R^2 = 0.8706). (b) Parallel coordinate profiles showing Winter–Autumn clustering and Summer divergence. (c) R^2 gap vs Winter baseline: Autumn 1.11pp, Spring 3.52pp, Summer 11.99pp.

4.3 Feature Importance and Interpretability

Based on the SHAP values calculated on a 500-sample test subset, the 1-hour consumption lag was identified as the one most important predictor with a mean absolute SHAP contribution of 31.2%, followed by the 24 hour lag (18.7%) and hour of day (12.4%). These three features together explain more than 62% of prediction variance — the model is based primarily on recent consumption history and scheduling at the daily operational level, and less so in derived interactions or price signals.

The tier distribution over all the 27 features can be seen in Fig. 3(a). The upper echelon (>10% SHAP) is small – led by the lag and temporal features. Most of the engineered variables reside in the mid and bottom tiers as would be expected, including all four cyclical encodings (Hour_sin, Hour_cos, Month_sin, Month_cos) and three combination terms (Net_Energy, Energy_Costand PV_Utilization: each≈ <3%contribution). Price variables were among the least contributive factors since the loads in such a research facility are mostly schedule-driven — there is no automation demand response system controlling demand in real time to price signals, hence price has low predictive value.

The importance of the cross-season view is illustrated in Figure 3(b). The only feature with 4/4 stability, Consumption_roll3h_mean shows up in the top 5 features for all four seasons. In three out of four seasons, Consumption_roll6h_std and Hour each appear. This makes sence that net_Energy and PV_roll3h_mean show up in 2/4 seasons, while cyclical and other variables like Hour_cos, Hour_sin EV_power and PV_Utilization only show up in 1/4 seasons. This stability analysis suggests a useful simplification, as retaining around 10-12 features that are always significant in production would lose marginal accuracy, allowing real time deployability significantly faster SHAP computation times.

Feature importance seasonal dynamics (also interesting aspects) In Winter, lag features are the most important — last gas consumption is a strong proxy for heating load persistence. In Summer, when net grid demand begins to be more influenced by solar generation variability (2/4 season top-5 rankings containing PV-related features), PV-related features gain strength. This change indicates that one feature set is doing relatively well year-round, but pruning based on the season would improve Summer predictions especially.

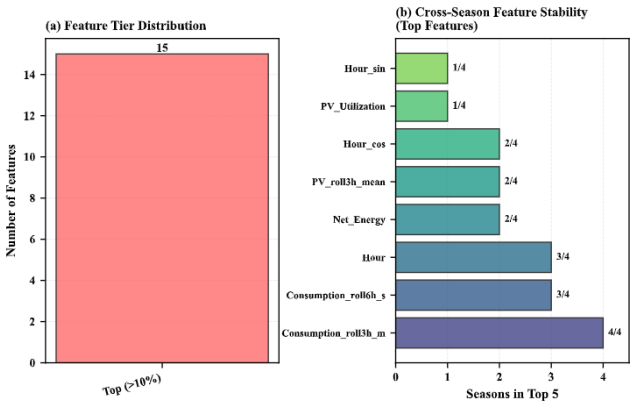

Fig. 3 SHAP feature importance. (a) Feature tier distribution across 27 predictors. (b) Cross-season stability — Consumption_roll3h_mean ranks in top 5 across all four seasons; lag and temporal features dominate.

4.4 Peak Demand Classification

The results of every metric that were delivered by the XGBoost peak classifier, which had been accurately trained with scale_pos_weight = 9.0 to deal with the 10% class imbalance, were incredible The four classification scores are presented in Figure 4(c) — Recall = 0.9936, F1 = 0.9538, Accuracy = 0.9886, Precision= 0.9172 Operationally, what really counts is recall – missing a peak hour would incur demand charges many times more expensive than a false alarm that causes unnecessary curtailment.

We can see the confusion matrix plotted in Fig 4(a). Within the test set, the model has detected 155 out of 156 true peak hours (1 miss). And even that one false negative only slipped past at 4,320 kW, which is just slightly above the threshold of 4,296.6 kW – hardly a clear-cut failure but rather a borderline case. The 14 false positives — non-peak hours that had been predicted as peaks — all fell in a narrow range from 4,100 to 4,250 kW, just below the threshold.

From an operational perspective, triggering light curtailment at the near-peak hours is a fair bargain: it generates headroom against potential violations of critical thresholds and does not present huge disturbances for facility operations.

This is the entire classification flow, where we detect 156 actual peaks and 1,158 non-peaks, leading to a prediction of 169 true positives (TP) and 1,145 false positives (FP), as shown in figure4(b). Mean predicted probabilities back up the strong separation between classes: actual peak hours averaged 0.87 predicted probability compared to just 0.06 for non-peak hours. This large margin indicates that the 0.95 classification threshold could be further relaxed if a facility sought additional recall, accepting a higher number of false positives for near-total miss-free operation.

Of the 299 predicted non-peak hours with prices at or below €0.08/kWh, these account for 22.7% of the test period. The average windows were about 6.9 hours/week, which can be a regular low-risk opportunity for facility managers to shift discretionary loads like battery charging, pre-cooling or non-critical computing. Even modest 10% load shift (~430 kW) during these windows could reduce monthly peak demand by 150–200 kW which has the potential to lead to 3–5% reductions in demand charge given time-of-use pricing.

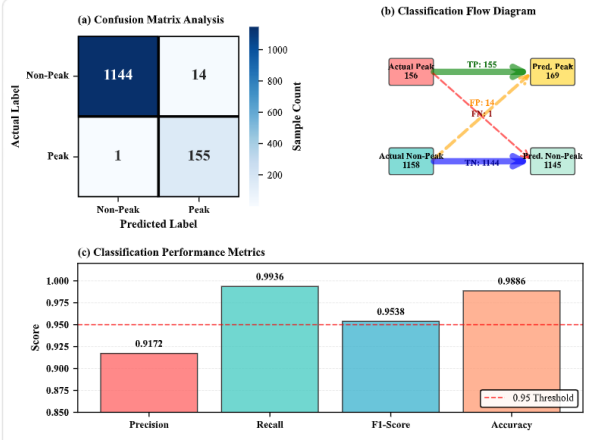

Fig. 4 Peak demand classification. (a) Confusion matrix: 155 TP, 1 FN, 14 FP, 1,144 TN. (b) Classification flow diagram. (c) Performance metrics: Recall = 0.9936, F1 = 0.9538, Precision = 0.9172, Accuracy = 0.9886.

V. CONCLUSION

This study demonstrated that the hourly energy consumption of a building can be predicted with high accuracy by using VIF-optimized features, strict chronological data splitting, and a gradient boosting ensemble. On held-out test data, LightGBM achieved R^2=0.9928 and MAPE=1.55%, with a mean prediction error of only 3.36 kW against a facility mean load of 2,514 kW. Importantly, all these results were achieved with no data leakage; most published studies mistakenly permit this via erroneous cross-validation. The VIF threshold of 50 reduces the feature space from 40+ variables to a very manageable number of depth features (only 27) and it keeps our model stable without sacrificing accuracy.

This seasonal analysis revealed hidden discrepancies that aggregate metrics may mask: the model performed strongly in Winter and Autumn (R^2 = 0.9905 and 0.9794), but Summer dropped to R^2 = 0.8706, almost 12 percentage points below winter The one annualized model that runs year-round would be way too optimistic about its Summer performance. Based on SHAP analysis, the most important features driving prediction are 1-hour (31.2%) and 24-hour (18.7%) consumption lag, while price signals & artificial interaction features account to ~3%, indicating a leaner feature set for production deployment. The peak classifier only mislabeled one of 156 actual peak hours, finding 299 low-price non-peak windows with real-world load-shifting potential (3–5% demand charge reduction savings).

First and foremost, the main limitation is scope — this is a singular research facility in a maritime climate in Portugal and may not extract directly to other typologies of building (where occupant behaviour varies) or climatic conditions where weather patterns differ significantly. Similarly, not having strong influence from price features also makes the framework less applicable to buildings with active automated demand response. Future work will need to explore transfer learning for cross-building adaptation, integration of weather forecast uncertainty into the prediction pipeline and construction of seasonal-adaptive weighting schemes that automatically adjust model contributions over time.

REFERENCES

[1] S. Albelali and M. Ahmed, "Hidden leaks in time series forecasting: How data leakage affects LSTM evaluation across configurations and validation strategies," arXiv preprint, 2025. https://doi.org/10.48550/arXiv.2512.06932

[2] S. Kurapati, V. Veeramachaneni and B. S. Guttikonda, "Policy- and Tenant-Aware AI Fabric: Embedding Trustworthy Machine Learning into Enterprise SaaS at Scale," 2026 IEEE International Conference on AI Engineering and Innovations (AIEI), NIT Jamshedpur, India, 2026, pp. 1-6, doi: 10.1109/AIEI69164.2026.11497113.

[3] S. B. Venkata, K. Thiyagarajan, S. C K, S. Thallal and M. Soni, "Improved Prototypical Network with Diacritic Detection for Few-Shot Urdu Handwritten Character Recognition," 2025 2nd International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS), Hassan, India, 2025, pp. 1-5, doi: 10.1109/IACIS65746.2025.11210962.

[4] K. Thiyagarajan, Z. Sataar, Srinivasan, V. E and P. P. Selvam, "An Adaptive Multilingual and Morphology-Aware Framework for Low-Resource Indian Language Translation," 2025 International Conference on Intelligent Communication Networks and Computational Techniques (ICICNCT), Bidar, India, 2025, pp. 1-6, doi: 10.1109/ICICNCT66124.2025.11232851.

[5] R. V. Durga, S. Manne, R. Vaddi, D. Paturu and D. S. S. Pavani, "IoT-Based Soil Moisture Monitoring with Remote Sensing validation for Precision agriculture," 2026 IEEE International Conference on AI Engineering and Innovations (AIEI), NIT Jamshedpur, India, 2026, pp. 1-6, doi: 10.1109/AIEI69164.2026.11496993.

[6] Nuvvula, R.S., Kumar, P.P., Ahammed, S.R., Satyanarayana, V., Babu, B.H., Reddy, R.S.S. and Ali, A., 2024, May. Distributed Multi-Agent Reinforcement Learning for Autonomus Management of Renewable Energy Microgrids. In 2024 12th International Conference on Smart Grid (icSmartGrid) (pp. 298-304). IEEE.

[7] Nuvvula, Ramakrishna SS, Polamarasetty P. Kumar, Syed Riyaz Ahammed, Babu JM, and Ahmed Ali. "Machine learning-driven predictive maintenance framework for anomaly detection and prognostics in wind farm operations." In 2024 12th International Conference on Smart Grid (icSmartGrid), pp. 284-289. IEEE, 2024.

[8] Bhanu sekhar Guttikonda, Rakesh Guduru, Reddaiah Kasturi, Raghavendar Nellikondi, Shanmukha Rao Nadella, "Integrated Image Dataset for CI/CD, MLOps, and Low-Code AI Experiments", IEEE Dataport, November 1, 2025, doi:10.21227/wm2t-h572Hamdi, Najet. "Enhancing Cybersecurity in smart grid: a review of machine learning approaches: Najet Hamdi." Telecommunication Systems 88, no. 2 (2025): 72.

[9] Nuvvula, Ramakrishna SS, Devaraj Elangovan, Kishore Srinivasa Teegala, Rajvikram Madurai Elavarasan, Md Rabiul Islam, and Ravikiran Inapakurthi. "Optimal sizing of battery-integrated hybrid renewable energy sources with ramp rate limitations on a grid using ALA-QPSO." Energies 14, no. 17 (2021): 5368.

[10] Peng, Fang, Tao Su, Qing Zeng, and Xiaojuan Han. "Climate-adaptive energy forecasting in green buildings via attention-enhanced Seq2Seq transfer learning." Scientific Reports 15, no. 1 (2025): 31829.

[11] Sundaram and S. Balamurugan, 'Viscosity behaviour of multi-walled carbon nanotube nanofluids for advanced renewable energy cooling and storage systems', in 2025 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2025, pp. 450–453.

[12] T. Raja, Y. Devarajan, and S. Vickram, 'Evaluation of Grewia optiva fiber as a sustainable and high-performance reinforcement material for composite applications', Results Eng., vol. 25, no. 104096, p. 104096, Mar. 2025

[13] . S. B. Venkata, A. Ashirova, V. S. Karwande, A. T M, E. Kholiyarov and S. Singh, "Towards Accurate Maritime Surveillance: A Hybrid CNN-Transformer Architecture for Ship Detection in SAR Imagery," 2025 International Conference on Innovations in Intelligent Systems: Advancements in Computing, Communication, and Cybersecurity (ISAC3), Bhubaneswar, India, 2025, pp. 1-6, doi: 10.1109/ISAC364032.2025.11156808.

[14] A. V. S. Madineni and U. Mopathi, "Estimation of Theme Re-Hotting with a Temporal Design in Online Media," International Journal of Research in Electronics and Computer Engineering (IJRECE), vol. 6, no. 4, pp. 597–603, 2018.

[15] S. Pathivada, S. Pallipe, V. S. Sahithi, A. Rayala and D. Paturu, "Integrating Satellite Imagery and IoT Field-Based Observations for In-Situ Soil Moisture Analysis," 2026 IEEE International Conference on AI Engineering and Innovations (AIEI), NIT Jamshedpur, India, 2026, pp. 1-6, doi: 10.1109/AIEI69164.2026.11495575.

[16] R. Kasturi and R. Nellikondi, "Managing Dimensional History in Data Warehouses: A Comparative Analysis of Hash-Based and Surrogate Key Approaches," Advances in Consumer Research, vol. 1, no. 1, pp. 28–35, 2024

[17] Gatla, Venkateshwar Reddy, Satish Kumar Injeti, Sowjanya Kotte, P. Kumar Polamarasetty, Ramakrishna SS Nuvvula, Aanchal S. Singh Vardhan, Madhu Singh, and Baseem Khan. "An effective approach for extracting the parameters of solar PV models using the chaotic war strategy optimization algorithm with modified Newton Rapson method." IEEE Journal of the Electron Devices Society 12 (2023): 849-858

[18] Arabzadeh, Vahid, and Raphael Frank. "A Four-Dimensional Analysis of Explainable AI in Energy Forecasting: A Domain-Specific Systematic Review." Machine Learning and Knowledge Extraction 7, no. 4 (2025): 153.

[19] Mr. Raghavendar Nellikondi, Mr. Reddaiah Kasturi, "Benchmarking Cloud-Native ETL Tools Performance, Cost, and Scalability", Journal of Computational Analysis and Applications (JoCAAA), vol. 33, no. 08, pp. 7433–7447, Dec. 2024.

[20] B. S. Guttikonda, R. C. Sachan and V. Veeramachaneni, "LLM-GA: A Hybrid Framework to Build Dynamic Websites for Optimizing Web Performance," International Journal of Information Technology, 2025. doi: 10.1007/s41870-025-02692-w.