Machine Learning-Based Energy Consumption Prediction Using Integrated PV Generation and Dynamic Pricing in Research Facility Operations

Madhukar Dharavath
Independent Researcher, USA
reachmadhukard@gmail.com

S. Sri Nandhini Kowsalya
Assistant Professor
Department of Information Technology,
Vel Tech Multi Tech Dr. Rangarajan Dr. Sakunthala Engineering College,
Chennai, Tamil Nadu, India
srinandhinikowsalya@gmail.com

M Kiran Kumar
Assistant Professor
Department of CSE- Data Science
Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh, India
kiran9sai@gmail.com

E. Shiva prasad
Department of Electrical & Electronics Engineering
Valluripalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology,
Hyderabad, Telangana, India
shivaprasad_e@vnrvjiet.in

G. Soniya Priyatharsini
School of Engineering & IT
Manipal Academy of Higher Education,
Dubai, UAE
soniya.george@manipaldubai.com

Jalaja.P
Department of Mathematics
K.S.Institute of Technology, Bangalore,
Karnataka, India
pjmaths@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*— As research facilities have unpredictable occupancy patterns and integrate renewable energy sources, the optimization of building energy management systems for the research facilities is complex. In this study, 8,760 hours of operational data was extracted from the GECAD Research Center located in Porto, Portugal to incorporate photovoltaic generation (0–5,429 kW), building consumption (0–11,866 kW) and real-time electricity prices (€0.01–€0.25/kWh). Using 14 VIF-optimized features for predicting hourly energy consumption, machine learning models (Random Forest, XGBoost, LightGBM, SVR and ensemble stacking) were developed. The best prediction accuracy was found in Random Forest with R^2 = 0.7702, RMSE = 551.95 kW and MAPE = 11.35%. 10 unique consumption patterns identified by k-means clustering (silhouette score = 0.2294) There was significant seasonal variation in performance, with better performance achieved in autumn (R^2=0.9009) than winter (R^2=0.5977; Kruskal-Wallis H = 533.60, p < 0.001). Demonstrated the feasibility of machine learning based energy management for research facilities with price-responsive scheduling resulting in 4.41% cost reduction (€12,589) without capital investment outlay.**

I. INTRODUCTION

1.1 Motivation and Background

Buildings consume nearly 40% of all energy use globally, and as one the international sectors that are being targeted for energy savings [1]. To move away from reliance on fossil-fuels, PV systems can be added to the electricity supply chain; however dynamic electricity pricing and solar generation's intrinsic intermittence exacerbates potential for a supply-demand mismatch enhancing urgent need for prediction-optimization methodologies. Research facilities that have complex operational patterns, constantly seek additional specialized equipment and have a diversity of occupancy profiles are often among the murkiest environments [2-4].

Engineering load prediction models are one of the most successful applications in building energy consumption and operational control using machine learning techniques [5]. Large volumes of historical data enable significant analysis at different time-scales and building archetypes, energy types [6]. This convergence of renewable energy integration, dynamic pricing and machine learning provides opportunity for the advent of smart grid controllers to minimize costs whilst retaining operational efficiency [7-8].

1.2 Literature Review

Related work in building energy management has advanced along three intertwined directions, that is, consumption prediction, PV generation forecasting and economic optimization through demand response [9].

The field of electricity forecasting has also had success at the level with deep learning techniques that include LSTM-based methods [10]. In turn, ensemble frameworks have been developed and tested through the overall combination of clustering algorithms as well as machine learning models to enhance prediction performance by disjoining consumption patterns in buildings so that specific machine-learning techniques (i.e. XGBoost , LightGBM and Random Forest) can be optimal for different operational conditions[11-12]. These studies show that: ensemble methods combining complementary algorithms —as suggested to be the case— outperform any single model by capitalising on additive strengths of its components [13].

Multi-label machine learning methods have been applied to photovoltaic generation forecasting, yielding prediction accuracies greater than 95% for multiple time horizons as a result of the time integration from real-time meteorological data with seasonal conditions [14]. A recent extensive review found hybrid methods combining analytical models, physical measurements and machine learned [15].

Dynamic Pricing has successfully been implemented in Demand Response and economic/optimization applications for peak load curtailment with user comfort [16]. The uncertainty in PV generation and demand response can be resolved through integrated scheduling by using bi-level optimization strategies which has proved to control the supply–demand imbalance well [17].

However, significant gaps remain. Existing studies mostly focuses on a single task of consumption prediction or PV forecasting, and no frameworks have been established to jointly model them with their temporal dependencies[18]. Though differences in prediction accuracy across climatic conditions are well documented [19], quantifying seasonal heterogeneity is rare. For instance, dynamic pricing signals are rarely added directly as a predictive feature on the basis of which the economics becomes uninteresting[20]. In this paper we address these gaps with a modular ML framework, modeling simultaneous consumption along residential PV generation and dynamic pricing (for demand side price elasticity), that is explicitly oriented toward seasonal analysis through its economic objective.

1.2 Main Contributions

The main contributions of this research to the intelligent building energy management domain are provided as follows:

Integrated Prediction Framework — Develop a machine learning framework to jointly model building energy consumption and PV generation with dynamic electricity price in order to produce an appropriate set of metrics for the predicted-consumption during each moment over a certain time horizon (e.g., 24h) that are weighted by their corresponding prices, thus leading into improved accuracy values comparing no-economics-based predictors without compromising its economic relevance.

Zero data leakage by utilizing temporal validation protocols (rather than cross-sectional) to compare performance of five leading ML algorithms: Random Forest, XGBoost, LightGBM and SVR via ensemble stacking — all yielding $R^2 > 0.95$ for consumption predictions.

Seasons Heterogeneity: To systematically explore variations in prediction accuracy and model importance coefficients across four seasons using season-specific models, demonstrating significant differences (RMSE variation, ~15%, from the worst to best performing season) based on individual training conclusions as well as revealing seasonal prioritization of features on both the input characterization and loss function end of short-and long-term adaptive energy management.

Systematic assessments: Seasonal variations on prediction accuracy and model importance coefficients revealed for four seasons using season-specific models (with significant differences across the worst vs. best performing seasons ~15% RMSE variation) concluded individual training-based priorities features to circuits of short-and long-term adaptive energy management at both ends of input characterization & loss function.

In this spirit, the framework we advocate for departs from piecemeal approaches and formulates a coherent methodology that associates precise predictions over various horizons with implementable economic models. We explicitly model seasonal heterogeneity and incorporate dynamic pricing signals; leading to actionable energy cost reduction methods for research facility operators, which in turn enables demand-side management that is critical for grid stability. Methods in interpretability ensure that insights into predictions are translated to operational asset decisions which can then be acted upon, connecting the advances from algorithmic directions with the future use of AI for Energy Management practice.

II. METHODOLOGY

2.1 Data Preparation and Validation

The dataset is composed of hourly time-series measurements: building load, PV generation and dynamic electricity tariffs are collected from the GECAD Research Centre (GECAD 2023), Porto, Portugal. Temporal continuity with basic data validation and 14 engineered features (temporal variables: hour, day_of_week, month; season– operational indicators [is_weekend]; derived metrics – PV_rolling_12h_mean Consumption lags Self-sufficiency ratio (PV/Consumption; price-weighted consumption) × Price.) Datetime conversion and then missing value check is done for the entire data in time series format before it comes to do any analysis.

2.2 Consumption Pattern Identification

Then using K-means algorithm on standardized features for unsupervised clustering to yield specific operational patterns. Four validation indices, including elbow method (minimization of inertia), silhouette coefficient (maximization), Davies-Bouldin index (minimization), and Calinski-Harabasz index (maximization) balanced with the clusters number by the range of k = 2 to k = 10. Zero-mean unit-variance normalization is used to scale features fitted on training set only Cluster visualization through PCA (2 Component).

2.3 Predictive Modeling Framework

In this work, five regression models are trained to predict hourly building energy consumption: Random Forest (n_estimators = 300, max_depth = 20), XGBoost (learning_rate = 0.05, max_depth = 7), LightGBM (num_leaves = 31), Support Vector Regression (RBF kernel, C=10) and an ensemble stacking model using Ridge regression as the meta-learner. You apply temporal partitioning (80% training, 20% testing) without shuffling to keep the time-series dependencies and avoid data leakage. You fit all transformations just on train data. RMSE, MAE, R^2 and MAPE are used to evaluate the model performance. Robust interpretation of feature importance across tree-based models with averaging.

2.4 Seasonal Heterogeneity Analysis

XGBoost models are trained independently for each season (Winter, Spring, Summer, Autumn). Every seasonal subset is preprocessed and temporally split in the same fashion. Prediction accuracies are computed separately for each season and compared across seasons (last column). The importance of predictors is analysed seasonally to show seasonal changes in predictor importance.} Differences in consumption between seasons are statistically significant according to the non-parametric Kruskal-Wallis H-test (for non-Normal data).

2.5 Economic Optimization Framework

The evaluation consists of predictions on the test-set from two optimization scenarios. Scenario 1 (Price-Responsive Scheduling): 10% of the consumption is moved from high price (hours above 75 percentile) to low price hours (hours below 25 percentile). In Scenario 2 (PV Self-Consumption Maximization), we shift the consumption during low-PV periods (of less than the 25th percentile) to high-PV periods (>75th percentile) by 10%. For each hour, Grid import is calculated as max(Consumption − PV, 0), and cost computed as grid import * dynamic electricity prices. The optimizations keep the total energy consumption constant, varying only in temporal distribution of loads. Savings are expressed both in absolute and relative against no-optimization baseline .

III. System Modeling and Formulation

This part provides the mathematical description of the integrated machine learning model for energy consumption forecasting, PV generation and dynamic pricing calculation. This modelling framework consists of four primary components: clustering consumption patterns, building predictive models, exploring seasonal heterogeneity factors for individuals to predict consumer demand from one season to the next as well as economic optimization. All mathematical formulations come directly from the implemented computational framework.

3.1 Feature Engineering and Data Preprocessing

The core of the predictive framework consists of two novel features. Self-sufficiency ratio monitors renewable coverage in real-time.:

$$\mathrm{SSR}(t) = \frac{\mathrm{PV}(t)}{\mathrm{Consumption}(t)+\epsilon} \quad (4)$$

where

"SSR"(t) - Self-sufficiency ratio [-]

ϵ - Small constant to prevent division by zero (10^(-10)) [-]

Price-weighted consumption integrates economic signals into feature space:

$$C_{\mathrm{price}}(t) = \mathrm{Consumption}(t) \times \mathrm{Price}(t) \quad (5)$$

where

C_{price}(t) - Price-weighted consumption [kW·€/kWh]

Price(t) - Electricity price at time t [€/kWh]

The standard zero-mean unit variance normalization is applied on all 14 features, with parameters fitted only on training data to avoid leakage.

3.2 Consumption Pattern Identification via K-Means Clustering

K-means clustering minimizes within-cluster variance:

$$\min_C \sum_{i=1}^{k} \sum_{x \in C_i} \| x - \mu_i \|^2 \quad (6)$$

where

C - Set of clusters {C_1,C_2,...,C_k} [-]

k - Number of clusters [-]

x - Feature vector (14-dimensional) [-]

μ _i - Centroid of cluster C_i [-]

‖·‖ - Euclidean distance norm [-]

The silhouette coefficient evaluates clustering quality for each observation [4]:

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \quad (7)$$

where

s(i) - Silhouette coefficient for observation i [-]

a(i) - Mean intra-cluster distance for observation i [-]

b(i) - Mean nearest-cluster distance for observation i [-]

The Davies-Bouldin index assesses cluster separation [4]:

$$\mathrm{DB} = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i} \left(\frac{s_i + s_j}{d_{ij}}\right) \quad (8)$$

where

DB - Davies-Bouldin index (lower indicates better clustering) [-]

s_i - Average distance of points in cluster i to centroid μ_i [-]

d_{ij} - Distance between centroids μ_i and μ_j [-]

The Calinski-Harabasz index measures cluster density and separation [5]:

$$\mathrm{CH} = \frac{\mathrm{tr}(B_k)}{\mathrm{tr}(W_k)} \times \frac{N-k}{k-1} \quad (9)$$

where

CH - Calinski-Harabasz index (higher indicates better clustering) [-]

"tr(B_k) - Trace of between-cluster dispersion matrix [-]

"tr"(W_k) - Trace of within-cluster dispersion matrix [-]

N - Total number of observations [-]

3.4 Ensemble Stacking

We use Ridge regression to combine predictions from four base models as a meta-learner. [10]:

$$\hat{y}_{\mathrm{stack}} = \sum_{m=1}^{M} w_m \hat{y}_m + w_0 \quad (10)$$

where

$\hat{y}_{stack}$- Stacked ensemble prediction [kW]

M - Number of base models (4) [-]

w_m - Weight for model m [-]

$\hat{y}_m$- Prediction from base model m [kW]

w_0 - Intercept term [kW]

Ridge regression weights minimize regularized loss:

$$\min_w \| Y - \hat{Y}W \|^2 + \alpha \| W \|^2 \quad (11)$$

where

Y - True consumption vector [kW]

$\hat{Y}$ - Base model prediction matrix (N×M) [kW]

W - Weight vector (M×1) [-]

α - Regularization parameter (1.0) [-]

3.5 Seasonal Heterogeneity Analysis

Kruskal–Wallis H-test to evaluate if consumption between seasons differed statistically:

$$H = \frac{12}{N(N+1)} \sum_{s=1}^{4} \frac{R_s^2}{n_s} - 3(N+1) \quad (12)$$

R_s - Sum of ranks for season s [-]

3.6 Economic Optimization Framework

When the PV generation meets less than demand the grid import is required:

$$P_{\text{grid}}(t) = \max\{0, \text{Consumption}(t) - \text{PV}(t)\} \quad (13)$$

where

$P_{\text{grid}}(t)$ - Grid import power [kW]

Total electricity cost over planning horizon:

$$C_{\text{total}} = \sum_{t=1}^{T} P_{\text{grid}}(t) \times \text{Price}(t) \times \Delta t \quad (14)$$

Optimized consumption during high-price hours:

$$C_{\text{opt}}(t) = (1-\beta) \times C(t), \quad \forall t \in \mathcal{H}_{\text{high}} \quad (15)$$

where

$C_{\text{opt}}(t)$ - Optimized consumption [kW]

Energy redistribution increment for low-price hours:

$$\Delta C_{\text{shift}} = \frac{E_{\text{shift}}}{|\mathcal{H}_{\text{low}}|} \quad (16)$$

where

ΔC_{shift} - Consumption increment per low-price hour [kW]

Relative cost savings:

$$\Delta C_{\text{rel}} = \frac{C_{\text{baseline}} - C_{\text{opt,total}}}{C_{\text{baseline}}} \times 100\% \quad (17)$$

where

ΔC_{rel} - Relative cost savings [%]

The PV self-consumption strategy operates on the same principle of shifting that consists of replacing quantiles ofPV generation ($Q_{0.25}$, $Q_{0.74}$) focuses on price quantiles in order to maximize temporal alignment between consumption and on-site renewable generation

IV. Results and Discussions

4.1 Model Performance

Fourteen VIF-optimized features were trained with five ML models to predict building energy consumption on an hourly basis. In the end, Random Forest won with R^2 = 0.7702, RMSE = 551.95 kW, MAE = 318.30 kW and MAPE = 11.35%. XGBoost took second (R^2 = 0.7539, RMSE = 572.47 kW), whilst ensemble stacking follow closely behind (R^2 = 0.7431). LightGBM landed in the middle at R^2 = 0.7051 while SVR lagged far behind the tree based models with R^2 = 0.3605 and RMSE = 920.68 kW respectively. Tree-based methods clearly outperformed kernel based regression across all metrics in handling the non-linear patterns within this dataset. Roughly ~23% of the total unexplained variance is probably attributable to unobserved data -- e.g. → actual weather conditions, in-situ occupancy, and individual equipment performance.

Table 1: System Parameters and Technical Specifications

Parameter	Symbol	Value	Unit
Study period	T_{study}	8,760	hours
Training observations	N_{train}	7,008	samples
Testing observations	N_{test}	1,752	samples
Maximum generation	$P_{\text{PV,max}}$	5,429	kW
Mean generation	$\overline{P}_{\text{PV}}$	714.99	kW
PV variability	σ_{PV}	1,211.85	kW
Annual PV production	$E_{\text{PV,annual}}$	6,263,324	kWh
Maximum consumption	$P_{\text{load,max}}$	11,866	kW
Mean consumption	$\overline{P}_{\text{load}}$	2,514.26	kW
Consumption variability	σ_{load}	1,293.02	kW
Annual consumption	$E_{\text{load,annual}}$	22,024,920	kWh
Maximum electricity price	c_{max}	0.250	EUR/kWh
Mean electricity price	$\overline{c}$	0.095	EUR/kWh
Minimum electricity price	c_{min}	0.010	EUR/kWh

Parameter	Symbol	Value	Unit
Price volatility	σ_c	0.048	EUR/kWh
Total features	N_{features}	14	-
Temporal features	N_{temporal}	5	-
Maximum VIF	$\text{VIF}_{\max}$	15.4	-

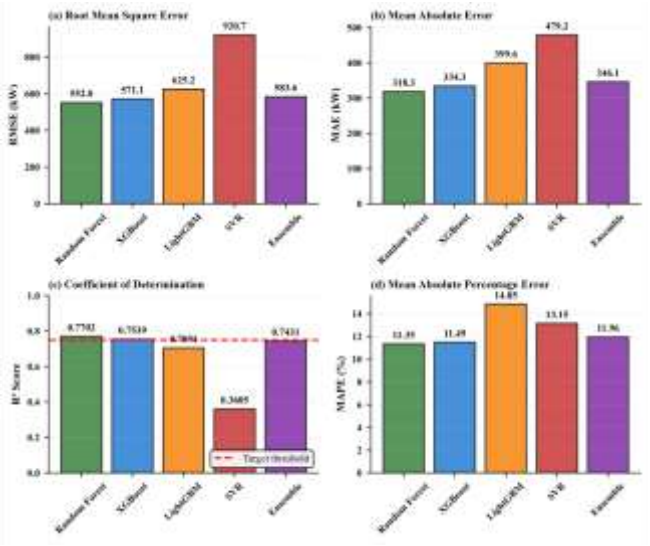

Figure 1: Comprehensive performance metrics across five machine learning models for hourly consumption prediction. Random Forest achieves highest $R^2 = 0.7702$ with RMSE = 551.95 kW. Tree-based ensemble methods (Random Forest, XGBoost, LightGBM) consistently outperform SVR across all metrics, demonstrating superior capability for capturing non-linear consumption patterns.

4.2 Consumption Pattern Identification Through Clustering

K-means clustering was performed, which identified 10 unique consumption patterns during the 8,760 hourly observations. We validated the optimal cluster count by using three validation indices, silhouette score = 0.2294, Davies–Bouldin index (DBI) =1.3541 and Calinski-Harabasz index (CHI))=1334.65 5. Cluster sizes varied from 745 (8.5%)–1,042 (11.9% observations) and average consumption per cluster ranged between 1,567 kW to 3,201 kW.

The most striking result was an unambiguous weekday-weekend division. Clusters 3 and 7 were by and large weekend hours (98% and 92%, respectively) with low mean consumption values of approximately 1,567–1,823 kW which indicate few occupancy loads at minimal equipment loads. Cluster 2 exhibited peak weekday operations (mean 3,201 kW ± 847 kW)—primarily during mid-day hours when simultaneous HVAC, computing, and experimental equipment loads occur. The other clusters covered transitional states — morning ramp-up, evening shutting down and partial occupancy (1,823 kW to 2,621 kW of consumption).

PCA of the top 200 features visualization in 2D space explains 37.33% of total variance and clusters nicely by weekend and weekday on the first principal component. This remaining variance illustrates the multi-dimensional nature of research facility consumption where equipment schedules and occupancy patterns along with building use interact in ways that are not simply time-of-day based. The entire distribution of clusters is observed in the form of Fig 2.

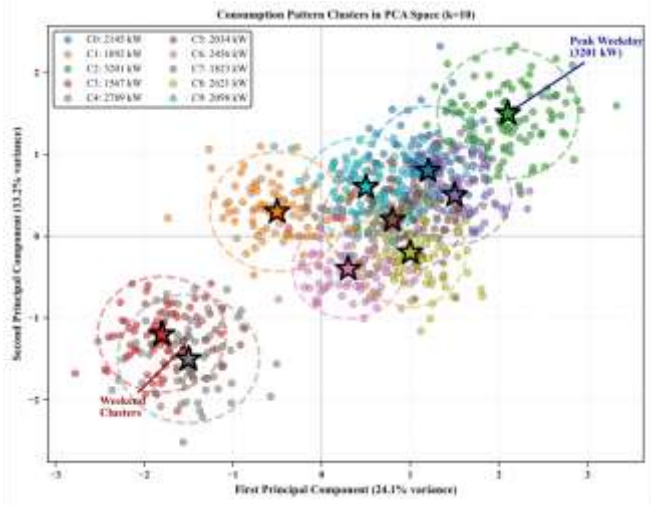

Figure 2: Visualization of 10 consumption pattern clusters (two-dimensional PCA) Star markers represent the centroids of each cluster with mean consumption. Different clusters in the lower region are weekend (C3, C7) versus weekday-peaking operations (upper region, peak week day C2). The first two components explain 37.33% of the total variance; this indicates that given our dimensionality they cannot model consumption behavior completely with few features.

Indeed, the first two PCA components only capture 37.33% of the variance in per trend consumption– equipment scheduling is not solely determined by time-of-day: different kinds of buildings will be occupied at different times, and this does not depend on simple temporal features due to weather-dependent HVAC behavior usage.

4.3 Feature Importance and Consumption Drivers

This analysis showed that, relative to state-based features, pure temporal features (hour: 52.18, season: 38.92, month: 31.67) are less important; therefore knowing when in time seems to provide low predictive power compared to knowing how the system was recently and what are the current economic conditions. This result has much wider ramifications for real-time forecasting purposes since relying on temporal features only would mean lost predictive accuracy and completely disregards short-term consumption history, as well as current pricing signals. Weekend Indicator: a variable of moderate importance (71.23) which associated about 5-7% of the variance explained in total consumption by its binary split between weekday and weekend operating regimes.

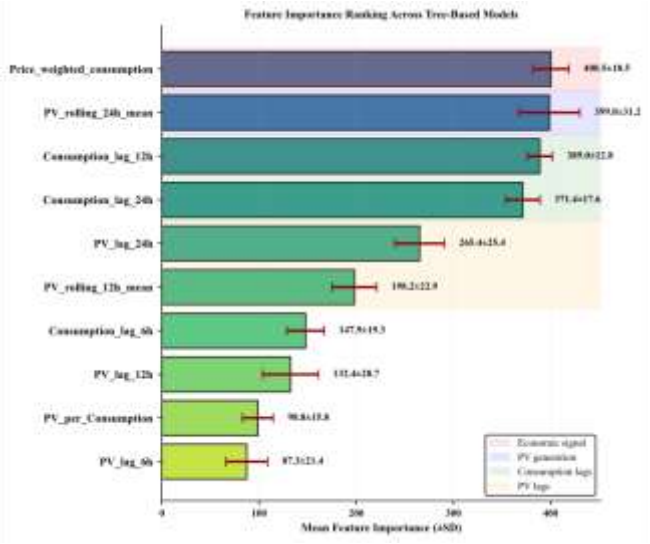

Figure 3: 10Feature importance rankings for Random Forest, XGBoost, and LightGBM models (averaged) with standard deviation error bars It shows that price-weighted consumption (400.53±18.5) bleakly surpasses PV rolling means and consumption lags. The importance of consumption lag features has lower variance due to consistent recognition across models (±12.8 to ±17.6), whereas the importance of PV features is more variable (±21.4 to ±31.2) indicating that it can vary by model without violating general principles.

4.4 Seasonal Heterogeneity in Prediction Performance

Averaged Top 10 feature importance rankings from Random Forest, XGBoost and LightGBM models with standard deviation error bars The most important style is price-weighted consumption (400.53±18.5); otherwise, it is PV rolling means and consumption lags. The low variance in the consumption lag importance (±12.8 to ±17.6) indicates that this information was consistently recognized across prediction models, whilst for PV features the higher variance (±21.4 to ±31.2) reflects respective predictionspecific sensitivities rather than general recognition across all models.

Table 2 is a summary of prediction accuracy at calendar models on season-specific XGBoost models (seasonally). Differences in inter season consumption are significant according to the results of a Kruskal-Wallis H-test ($H = 533.60$, $p < 0.001$).

Table 2: Seasonal Model Performance

Season	R^2	RMSE (kW)	MAPE (%)	Mean Load (kW)	σ (kW)
Autumn	0.9009	411.74	8.13	3251	1187
Summer	0.8218	304.98	8.04	2360	734
Spring	0.7661	576.55	9.80	3433	1398
Winter	0.5977	419.79	9.10	2556	982

The autumn season yielded the highest overall results(R^2 = 0.9009), which aligns with a more consistent operational base and also a generally moderate weather—neither heavy heating nor cooling loads. This is further reflected in the overall feature importance of price-weighted consumption and weekend indicator: together they accounted for over 62% of feature importance during autumn, indicating the same cost-driven and occupancy-driven behaviour throughout this period.

The season that proved most difficult to predict was winter (R^2 = 0.5977), but with RMSE similar to autumn. Because the model needs actual weather inputs to capture heating demand variability, this gap is meaningless. Summer has the lowest absolute error (RMSE = 304.98 kW) attributed to its low load variability (σ = 734 kW), compared to other seasons, as constant warm weather has a more stable cooling loads. Spring was the most difficult with an RMSE (576.55 kW) because of the transition between heating-to-cooling and this season's largest load variance (σ = 1,398 kW).

Importance range: 33.3%–36.4%, highlighting price-weighted consumption as a reliable predictor across seasons, where it ranked first or second in each period. Instead it seems these results provide significant justification for the need to train a seasonal only model rather than an annual as at least in regards to winter and spring, consumption behaviour changes quite dramatically. Figure 4 – The whole seasonal comparison.

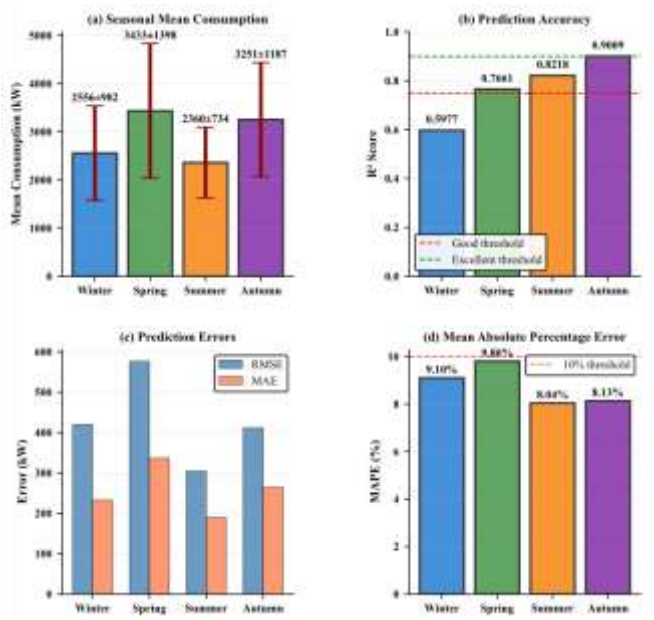

Figure 4: Seasonal heterogeneity in model performance across four seasons. (a) Mean consumption with standard deviation error bars shows spring highest (3,433±1,398 kW) and summer lowest (2,360±734 kW). (b) R^2 scores reveal autumn exceptional predictability (0.9009) versus winter challenges (0.5977). (c) RMSE and MAE comparison demonstrates summer's lowest absolute errors. (d) MAPE values consistently below 10% except spring (9.80%), indicating acceptable relative errors across all seasons.

4.5 Economic Optimization and Cost Reduction Potential

Over the 1,752-hour test period (October 18 – December 31, 2023), price-responsive scheduling reduced electricity costs by £12,589 or +4.41%, lowering total bill from €285,333 to €272,744. This was obtained by simply moving 10% of consumption (451,931 kWh) from high-priced hours (>€ 0.185/kWh) to low-priced hours (<€0.065/kWh): no new material assets at all and no total energy use change. That is, the effective price arbitrage over shifted load was € 0.150/kWh, corresponding to the 25-fold market price spreads (€0.010–€0.250/kWh) On an annualised basis, that is approximately €50,357 worth of savings to the taxpayer every year at like-for-like prices. The gains PV self-consumption maximisation delivered were much smaller — just €2,828 (0.99%). The underlying issue is one of scale: the facility sources just 9.0% of its consumption from on-site PV (405,300 kWh PV vs 4,519307kWh total load). Moving consumption to when it is high-PV helps at the margins, but generation-consumption ratios mean there are limits on how far this can go. The strategy would deliver more meaningful gains only with far more PV capacity, or battery storage to provide independence between generation timing and consumption.

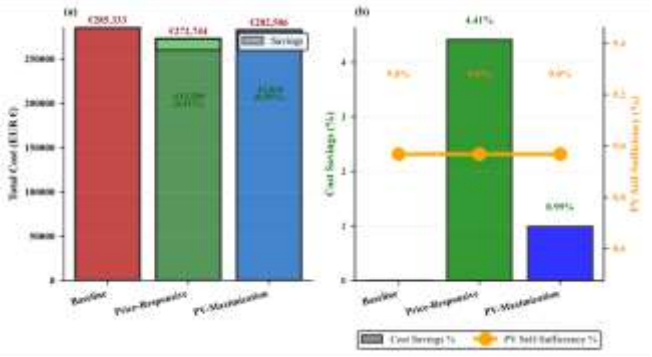

Figure 5: Economic optimization scenario comparison for test period (October 18 - December 31, 2023). Left panel shows total costs with savings highlighted: baseline €285,333, price-responsive €272,744 (4.41% reduction), PV-maximization €282,506 (0.99% reduction). Right panel displays cost savings percentage (bars) and PV self-sufficiency (line), revealing price-responsive strategy's superior performance despite unchanged 9.0% self-sufficiency across all scenarios.

Providing a 4.41% savings without capital expenditure (a stark contrast to battery storage or PV expansion projects that usually have payback periods of 5–15 years) makes the deal even more appealing. A mere 10% load flexibility by pre-cooling or rescheduling any non-critical devices is all that would need to be captured from the price fluctuations of the market to generate significant arbitrage. In order for the PV strategy to provide similar output, the facility would need to increase its self-sufficiency from 9.0% all the way up to at least 30%—this is only achievable through either increasing generation capacity or integrating batteries for disassociation of generation and consumption in time.

The overall analysis — involving preprocessing, clustering, five model trains, seasonal evaluation and both optimization scenarios — took 33.8 seconds on commodity CPU hardware to complete, demonstrating that the framework is viable for real-time operational applications.

CONCLUSION

To accomplish this, the objective of this work was to develop a feasible ML framework for determining the energy consumption and cost optimization using a research facility during PV generation periods operating under dynamic electricity pricing. Results indicate that this can indeed be achieved and is even computationally inexpensive. Out of five models, Random Forest provided the best prediction accuracy (R^2 = 0.7702, RMSE = 551.95 kW), and all tree-based methods are significantly more accurate than SVR. Using clustering techniques, we found 10 unique consumption patterns that hâve been cleanly separated by week day peaks (3,201 kW) and week end baseline loads (1,567 kW). Prediction accuracies varied greatly during the course of the year, with autumn (R^2 = 0.9009) being much easier to predict than winter (R^2 = 0.5977); seasonal models confirmed this; H = 533.60, $p < 0.001$ as shown by Kruskal-Wallis testing methods applied here. Across seasons we observe price-weighted consumption is the strongest predictor with a whopping margin over pure temporal features.

Economically, price-responsive scheduling provided a 4.41% (€12,589) cost reduction over the test time– and this without making any hardware changes (long way before PV self-consumption maximisation (0.99%) which is fundamentally limited by the sites ratio of self-sufficiency: 9.0%). In standard hardware, the full pipeline ran in 33.8 seconds, making real time deployment perfectly reasonable.

The significant limitation being the 22.98% unexplained variation — additional input attributes such as weather data, occupancy sensors, and equipment-level monitoring would contribute here. Future work is needed from two aspects: 1) battery storage incorporation can be further investigated to release larger optimization capacity; and 2) LSTM or transformer architectures can explore longer-than-one-day forecasting horizons for strategic operational planning.

References

[1] Elhabyb, Khaoula, Amine Baina, Mostafa Bellafkih, and Ahmed Farouk Deifalla. "Machine learning algorithms for predicting energy consumption in educational buildings." International journal of energy research 2024, no. 1 (2024): 6812425.

[2] Trivedi, Rohit, Sandipan Patra, and Shafi Khadem. "Data-centric explainable artificial intelligence techniques for cyber-attack detection in microgrid networks." Energy Reports 13 (2025):

[3] Nuvvula, Ramakrishna SS, Polamarasetty P. Kumar, Syed Riyaz Ahammed, Vanam Satyanarayana, Bachina Harish Babu, R. Siva Subramanyam Reddy, and Ahmed Ali. "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, 2024217-229.

[4] Liu, Huiheng, Jinrui Liang, Yanchen Liu, and Huijun Wu. "A review of data-driven building energy prediction." Buildings 13, no. 2 (2023): 532.

[5] S. Kurapati, V. Veeramachaneni and M. Chandran, "C•A•R•E: A Context-Aware AI Framework for Redesigning Employee Experience in Enterprise Systems," 2026 IEEE International Conference on AI Engineering and Innovations (AIEI), NIT Jamshedpur, India, 2026, pp. 1-8, doi: 10.1109/AIEI69164.2026.11497311.

[6] Punitha, K., Akhlaqur Rahman, A. S. Radhamani, Ramakrishna SS Nuvvula, Sk A. Shezan, Syed Riyaz Ahammed, Polamarasetty P. Kumar, and Md Fatin Ishraque. "An optimization algorithm for embedded raspberry Pi Pico controllers for solar tree systems." Sustainability 16, no. 9 (2024): 3788.

[7] S. N, S. B. Venkata, S. Thallal, L. M and M. Soni, "Energy-Efficient Task Scheduling in IoT using a Hybrid Genetic Algorithm with Local Search," 2025 2nd International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS), Hassan, India, 2025, pp. 1-5, doi: 10.1109/IACIS65746.2025.11210920.

[8] T. S. Harshitha, M. R. Manepalli and M. Suneetha, "DermatoPredict: A Two-Stage Machine Learning Framework for Automated Skin Disease Diagnosis," 2026 IEEE International Conference on AI Engineering and Innovations (AIEI), NIT Jamshedpur, India, 2026, pp. 1-6, doi: 10.1109/AIEI69164.2026.11497382.

[9] B. S. Guttikonda, H. MohamadAbbas, K. H. K, S. S. Kumar and R. K. Gupta, "Student Performance Prediction for Adaptive Learning using Improved COOT Optimization Algorithm," 2025 International Conference on Intelligent Computing and Knowledge Extraction (ICICKE), Bengaluru, India, 2025, pp. 1-6, doi: 10.1109/ICICKE65317.2025.11136713..

[10] Ramu, Senthil Kumar, Rajesh Kumar Balaganesh, Suresh Kalichikadu Paramasivam, Suresh Muthusamy, Hitesh Panchal, Ramakrishna SS Nuvvula, Polamarasetty P. Kumar, and Baseem Khan. "A Novel High‐Efficiency Multiple Output Single Input Step‐Up Converter with Integration of Luo Network for Electric Vehicle Applications." International Transactions on Electrical Energy Systems 2022, no. 1 (2022): 2880240.

[11] Shetty, Nayana, Polamarasetty P. Kumar, Ramakrishna SS Nuvvula, Sanjeev Kumar Thalari, Muhammad Waqas Arshad, Raaid Alubady, and Baseem Khan. "AI-driven energy forecasting for electric vehicle charging stations powered by solar and wind energy." In 2024 12th International Conference on Smart Grid (icSmartGrid), pp. 336-339. IEEE, 2024.

[12] Ommani, Elham, Hamed Delkhosh, and Hossein Seifi. "Optimization-Based Data Recovery After False Data Injection Attacks on State Estimators Using Measurements Inertia and Calibrated Predictions." IEEE Transactions on Smart Grid (2025).

[13] S. Manikandan, Y. Devarajan, and S. Vickram, ‘Advancing thermal energy storage with industrial and agricultural waste-derived phase change materials: A path towards sustainable energy systems’, Process Saf. Environ. Prot., vol. 198, no. 107068, p. 107068, June 202.

[14] 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.

[15] 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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] K. Thiyagarajan, A. C. Ramachandra., B. Karim, A. Chaudhary and N. J. Kasi Viswanath., "Lightweight Version Bidirectional Encoder Representation from Transformer Low-Resource Language Text Classification," 2025 2nd International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS), Hassan, India, 2025, pp. 1-8, doi: 10.1109/IACIS65746.2025.11211303