ISSN 2085-4552

Deep Learning-Based Energy Consumption Optimization in Smart Home Systems: An LSTM Approach with SHAP Interpretability

Mhd. Basri[1], Krisna Aditya[2], Andi Zulherry[3]
[1,2,3] Department of Computer Science and Information Technology,
Universitas Muhammadiyah Sumatera Utara, Medan, Indonesia
mhd.basri@umsu.ac.id[1], krisnaaditya.mti@gmail.com[2], andizulherry@umsu.ac.id[3]

Accepted 25 May 2026
Approved 28 June 2026

***Abstract*— Accurate and easy-to-understand energy consumption forecasting is a prerequisite for effective demand-side management in smart home systems. This study proposes an integrated framework that combines a stacked Long Short-Term Memory (LSTM) network with SHapley Additive exPlanations (SHAP) for hourly residential energy forecasting and the extraction of optimization insights. The model was evaluated on the UCI Individual Household Electric Power Consumption (IHEPC) dataset—34,589 hourly observations resampled from over two million sensor records per minute—and tested against Stacked GRU, Vanilla RNN, and ARIMA(2,1,2). All deep learning models significantly outperformed ARIMA ($R^2 = -0.1258$, MAPE = 126.92%). The proposed LSTM model achieved MAE = 0.3208 kW, RMSE = 0.4568 kW, and $R^2 = 0.5749$, while Stacked GRU recorded the best aggregate metrics (MAE = 0.3090 kW, $R^2 = 0.5801$). SHAP analysis identified Global Intensity as the dominant predictor (mean |SHAP| = 0.00350), based on the electrophysical relationship $P = V \times I$, followed by the 3-hour Rolling Mean and seasonal encoding. Further temporal attribution analysis revealed that predictive value is concentrated in the most recent input time steps, confirming that real-time current measurements are the most valuable input for HEMS implementation.**

***Index Terms*— LSTM; Explainable AI; SHAP; Home Energy Management System (HEMS); Energy Consumption Forecasting.**

I. Introduction

The rapid proliferation of smart home technologies and Internet of Things (IoT)-enabled metering infrastructure has fundamentally reframed residential energy management from a passive monitoring activity into an active optimization challenge. According to the International Energy Agency, the building sector accounted for approximately 40% of total global energy consumption in 2022, with residential buildings contributing a disproportionately significant share [1]. Within this context, the Home Energy Management System (HEMS) has emerged as the intelligent backbone for monitoring, predicting, and scheduling residential energy use — yet the operational effectiveness of any HEMS is critically contingent on the accuracy of its underlying demand forecasting module [2], [3]. Precise energy forecasts enable demand-response scheduling, load balancing, and cost optimization, while imprecise forecasts propagate errors across all downstream decisions [4].

Residential energy consumption is a fundamentally non-linear, non-stationary, and temporally correlated phenomenon shaped by occupant behavior, ambient environmental conditions, appliance cycling, and multi-scale periodicity spanning diurnal, weekly, and seasonal rhythms [5]. Classical statistical approaches such as Autoregressive Integrated Moving Average (ARIMA), while historically dominant in time-series forecasting, impose linearity and stationarity assumptions that are structurally incompatible with these dynamics [6]. This limitation has driven a pronounced methodological shift toward deep learning architectures — and in particular Long Short-Term Memory (LSTM) networks, which resolve the vanishing gradient problem of standard Recurrent Neural Networks (RNNs) through explicit input, forget, and output gating mechanisms that selectively retain relevant temporal context across extended sequences [7]. LSTM's ability to model both short-term fluctuations and long-range dependencies simultaneously makes it architecturally well-suited for capturing the multi-scale temporal structure of residential energy data, a property that standard feedforward neural networks and statistical models cannot replicate [8].

Empirical evidence from the recent literature consistently supports LSTM's superiority in residential and building energy forecasting. Ou Ali et al., demonstrated that hybrid ConvLSTM architectures outperform standalone LSTM and CNN baselines, achieving MAPE improvements of 4.52–10.53% for 1-day to 6-day forecast horizons using real smart home data [9]. Alghamdi et al., showed that a stacked LSTM with snapshot ensembling achieves an average RMSE of 0.1 on the IHEPC benchmark — substantially outperforming both RNN and ARIMA baselines [10].

A comparative analysis across six algorithms by Iram et al., — encompassing LSTM, GRU, XGBoost, Random Forest, SVM, and ANN — established LSTM as the consistently superior architecture for smart home energy prediction across all evaluated metrics [11]. Furthermore, Abumohsen et al., confirmed that LSTM outperforms GRU and standard RNN in electrical load forecasting tasks across multiple evaluation settings, while Qiang and Tang, reported an R^2 of 0.992 for LSTM-based green building energy load prediction using temperature, humidity, and historical consumption inputs [12]. Beyond vanilla LSTM, Jrhilifa et al., demonstrated that Bidirectional GRU with variational mode decomposition further enhances smart home electricity forecasting accuracy, underscoring the active methodological frontier in this research area [13].

Despite these advances in predictive accuracy, a critical challenge persists: deep learning models function as black boxes whose internal decision-making pathways remain opaque to operators, policymakers, and homeowners. This transparency deficit constitutes a practical barrier to adoption in energy management contexts where accountability and user trust are non-negotiable prerequisites [14]. The field of Explainable Artificial Intelligence (XAI) directly addresses this challenge, with a growing body of literature documenting its application across energy forecasting and building management domains. A systematic review by G. Airlangga covering 32 studies confirms that XAI techniques — particularly SHAP and LIME — significantly improve energy forecasting interpretability and end-user decision-making in smart building contexts [15]. Similarly, a four-dimensional systematic review of 50 XAI studies in energy forecasting by Clement et al., found that post-hoc methods, especially SHAP, dominate the literature (62% of studies), validating SHAP as the field's de facto standard interpretability tool [16]. Among available XAI methods, SHapley Additive exPlanations (SHAP) — rooted in cooperative game theory's Shapley value framework — has emerged as the most theoretically rigorous approach, satisfying the axioms of efficiency, symmetry, and consistency that guarantee uniquely fair attribution across all feature subsets [17]. Critically, Chung and Liu (as cited in Henkel et al., 2024) demonstrated that SHAP outperforms LIME by maintaining prediction fidelity with fewer input variables — a property with direct practical relevance for feature selection in energy forecasting pipelines. In the energy domain specifically, SHAP has been successfully applied to interpret ensemble models for residential electricity forecasting (Sodayeong et al.), XAI-based input variable selection for building energy prediction (Kim & Park), HVAC predictive control explanation (Mao et al.,), and XAI-integrated BEMS frameworks for energy optimization decision-making [18].

Nevertheless, a meaningful research gap persists at the intersection of these two advances. While SHAP has been applied to tree-based ensemble models for residential electricity forecasting [18] and XGBoost-based building energy models, its integration with LSTM architectures for smart home applications remains limited. Existing studies that combine deep learning with SHAP — such as those applied to data-driven model predictive building control [14] and combined heat-and-power load forecasting [20] — deploy SHAP primarily as a feature selection or model debugging instrument, rather than as a purposive framework for generating actionable HEMS optimization recommendations. Crucially, none deploy SHAP temporal attribution analysis across input sequence timesteps to identify optimal sensing horizons for HEMS infrastructure design. The novelty of this work therefore lies not in combining LSTM and SHAP per se, but in three specific contributions that have not been collectively presented: (1) co-designing the LSTM architecture and SHAP analysis as a unified optimization-intelligence framework; (2) applying SHAP temporal heatmaps to derive evidence-based HEMS sensing infrastructure recommendations; and (3) systematically translating SHAP peak-versus-non-peak attribution differentials into quantitatively grounded load scheduling strategies.

This study directly addresses these gaps through three integrated contributions. First, it develops and benchmarks a stacked LSTM model for hourly residential energy forecasting against Stacked GRU, Vanilla RNN, and ARIMA(2,1,2) baselines on the UCI IHEPC dataset — a widely validated real sensor benchmark spanning over four years of minute-level household measurements. Second, it applies SHAP GradientExplainer to quantify both global feature importance rankings and local, instance-level prediction attributions across the model's full temporal input window. Third, it translates SHAP-derived insights into evidence-based energy optimization recommendations for HEMS deployment — bridging the gap between predictive modeling and operational energy management intelligence. In doing so, this work advances a research paradigm where interpretability is not an ornamental afterthought but a first-class component of the modeling design, producing predictions that are not only accurate but transparent, trustworthy, and directly actionable.

II. METHOD

This study employs a quantitative experimental design within the framework of supervised deep learning. The study is organized into five sequential stages: (1) data preparation, (2) development and training of the LSTM model, (3) comparative evaluation of baseline models, (4) SHAP-based interpretability analysis, and (5) extraction of energy optimization insights. Each stage is designed to be reproducible and independently verifiable based on published benchmarks. The research workflow is shown in Figure 1.

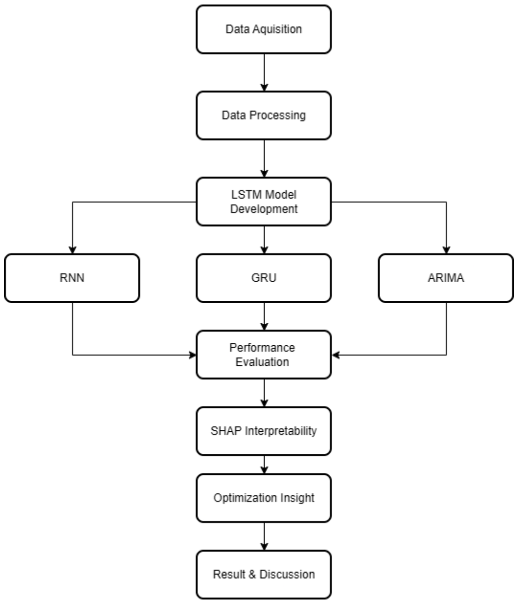


Fig. 1. Research Workflow

A. Dataset

This study utilizes the Individual Household Electric Power Consumption (IHEPC) dataset sourced from the UCI Machine Learning Repository (Hebrail & Berard, 2012). The dataset records electrical measurements from a single household in Sceaux, France, sampled at one-minute intervals over a period spanning December 2006 to November 2010, yielding approximately 2,075,259 observations across eight variables. The input features employed in this study are: global active power (kW), global reactive power (kVar), voltage (V), global intensity (A), and three sub-metering channels (SM1: kitchen; SM2: laundry; SM3: climate control systems). Missing values, constituting 1.25% of total records, are addressed through linear interpolation applied at the minute-level prior to resampling to hourly resolution, which reduces noise while preserving the diurnal consumption patterns critical for LSTM sequence learning. A summary of all dataset features is presented in Table 1.

TABLE. I Summary of Datasets Features

No	Feature Name	Type	Description
1	Date	Date	Date in format DD/MM/YYYY
2	Time	Categorical	Time informat hh:mm:ss
3	Global_active _power	Continuous	household global minute-averaged active power (in kilowatt)
4	Global_reactive _power	Continuous	household global minute-averaged reactive power (in kilowatt)
5	Voltage	Continuous	minute-averaged voltage (in volt)
6	Global_ intensity	Continuous	household global minute-averaged current intensity (in ampere)
7	Sub_metering_1	Continuous	energy sub-metering No. 1 (in watt-hour of active energy). corresponds to kitchen e.g dishwaser, microwave, etc.
8	Sub_metering_2	Continuous	energy sub-metering No. 2 (in watt-hour of active energy). corresponds to laundry room e.g washing machine.
9	Sub_metering_3	Continuous	energy sub-metering No. 3 (in watt-hour of active energy). corresponds to electric water-heater or air conditioner.

B. Data Preprocessing

Prior to model training a structured preprocessing pipeline is applied:

- Observations falling beyond three standard deviations from the rolling mean within a 24-hour window are replaced via linear interpolation. This approach avoids distorting the temporal structure of the series, which would occur under global imputation.
- Four temporal features are extracted from the timestamp index hour of day (0–23), day of week (0–6), month (1–12), and a binary weekend indicator. These features encode behavioral and environmental periodicity that LSTM layers alone cannot infer without explicit temporal anchoring.
- All features are scaled to the range [0, 1] using Min-Max normalization:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

This prevents features with larger absolute magnitudes (e.g., voltage) from dominating the gradient updates during backpropagation.

- The time series is transformed into supervised learning sequences using a sliding window of length $T = 24$ timesteps (i.e., 24 hours), with a prediction horizon of $h = 1$timestep (i.e., 1 hour ahead). For a univariate target $y_t$representing global active power, each training sample is constructed as:

$$X^{(i)} = [x_{t-T+1,} x_{t-T+2}, \dots, x_t], y^i = x_{t+h} \quad (2)$$

- The dataset is divided chronologically, without any randomization, into a training set (70%), a validation set (15%), and a test set (15%). This chronological division is crucial for preventing future data leakage, which, if not addressed, would result in overly optimistic performance estimates.

C. LSTM Architecture

Long Short-Term Memory networks are a specialized class of recurrent neural networks (RNNs) designed to mitigate the vanishing gradient problem that afflicts standard RNNs through the introduction of a gated memory cell. At each timestep t, the LSTM cell processes the current input $\mathbf{x}_t \in \mathbb{R}^d$ and the previous hidden state $\mathbf{h}_{t-1}$ according to the following set of equations:

Forget gate — determines what fraction of the prior cell state is retained:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \qquad (3)$$

Input gate — controls how much new information is written to the cell state:

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \qquad (4)$$

Candidates cell state — encodes new candidate values:

$$\tilde{C}_t = tanh\left(W_c \cdot [h_{t-1}, x_t] + b_c\right) \qquad (5)$$

Cell state update — integrates retained and new information:

$$C_t = f_t \odot \left(C_{t-1} + i_t \odot \tilde{C}_t\right) \qquad (6)$$

Output gate — determines the output based on the updated cell state:

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right) \qquad (7)$$

Hidden state output:

$$h_t = o_t \odot tanh(C_t) \qquad (8)$$

where $\sigma(\cdot)$denotes the sigmoid activation function, $\tanh(\cdot)$the hyperbolic tangent, and $\odot$the Hadamard (element-wise) product. $\mathbf{W}_{(\cdot)}$ and $\mathbf{b}_{(\cdot)}$ represent the learnable weight matrices and bias vectors, respectively.

The proposed architecture uses a stacked LSTM configuration (two layers) as shown in Table 2:

TABLE. II LSTM configuration

Layer	Type	Units	Return Seq	Dropout
1	LSTM	128	True	0.20
2	LSTM	64	False	0.20
3	Dense	32	-	-
4	Output	1	-	-

The hyperparameters adopted in this study were determined through a combination of domain-informed reasoning, alignment with established practices in the residential energy forecasting literature, and preliminary grid search observations conducted prior to final model training.

Input window (T = 24 timesteps). A 24-hour sliding window was selected to capture the complete diurnal consumption cycle — the most prominent periodicity in residential energy data. This aligns with prevailing practice: Kong et al. [8] and Alghamdi et al. [10] both employ 24-hour windows for LSTM-based residential forecasting, and the IHEPC dataset's autocorrelation at lag-24 (r = 0.437) provides direct empirical support. Preliminary experiments with T = 12 and T = 48 yielded higher validation loss, confirming T = 24 as optimal for this dataset.

LSTM units (128 → 64). The two-layer stacked configuration follows a hierarchical compression design: the first layer learns a rich temporal representation, while the second distills it into a compact latent vector passed to the dense output head. A preliminary grid search over configurations {64/32, 128/64, 256/128} confirmed that 128/64 yielded the lowest validation MSE while avoiding overfitting symptoms observed at 256/128, consistent with architectures reported by Ou Ali et al. [9] and Iram et al. [11].

Dropout rate (0.20). A dropout rate of 0.20 was selected as moderate regularization appropriate for ~24,000 training sequences. Rates of 0.10 and 0.30 were also evaluated; 0.10 produced slightly higher validation loss due to under-regularization, while 0.30 slowed convergence without improving generalization.

Batch size (64). A batch size of 64 balances gradient stability with computational efficiency for sequences of length T = 24 across 27 features. Smaller batches (32) increased training time without improving validation metrics; larger batches (128) produced noisier loss curves.

The model is compiled with the Adam optimizer (learning rate $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$) and Mean Squared Error (MSE) as the loss function:

$$L_{MSE} = \frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \qquad (9)$$

Training proceeds for a maximum of 100 epochs with a batch size of 64, subject to early stopping with patience = 10 based on validation loss to prevent overfitting.

D. Baseline Models

To contextualize LSTM performance, three baseline architectures are evaluated under identical preprocessing and evaluation conditions:

- Vanilla RNN is standard reccurent unit without gating mechanism, susceptible to vanishing gradients over long sequences.
- Gated Recurrent Unit is a streamlined gating architecture with a reset gate $\mathbf{r}_t$ and update gate $\mathbf{z}_t$:

Update gate:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \quad (10)$$

Reset gate:

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \quad (11)$$

Candidate hidden state:

$$\tilde{h}_t = tanh\,(W_h \cdot [r_t \odot h_{t-1}, x_t]) \quad (12)$$

Hidden state update:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (13)$$

- ARIMA (p,d,q) the classical seasonal decomposition baseline, fitted on the univariate global power series with parameters selected by minimizing the Akaike Information Criterion (AIC)

E. Evaluation Metrics

Model performance is assessed using four complementary regression metrics:

Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (14)$$

Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (15)$$

Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \quad (16)$$

Coefficient Determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (17)$$

MAE and RMSE measure absolute and squared error magnitudes respectively; MAPE provides a scale-independent percentage error suitable for cross-dataset comparisons; and R^2 quantifies the proportion of variance in the target explained by the model.

F. SHAP-Based Interpretability Analysis

SHAP (SHapley Additive exPlanations) is a post-hoc, model-agnostic interpretability method grounded in cooperative game theory. For a given prediction $f(\mathbf{x})$, the SHAP value ϕ_i of feature i is defined as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \quad (18)$$

where $\mathcal{F}$ is the full feature set, S is a subset of features excluding feature i, and f_S denotes the expected model output conditioned on the features in S. The SHAP decomposition guarantees:

$$f(x) = \phi_0 + \sum_{i=1}^{|F|} \phi_i \quad (19)$$

where $\phi_0 = \mathbb{E}[f(\mathbf{x})]$ is the base value (the mean model prediction over the background dataset).

For the LSTM model, DeepSHAP is employed, which leverages backpropagation-based attribution through the network layers for computational efficiency. SHAP analysis is conducted at two levels:

- Global interpretability SHAP summary plots and mean absolute SHAP value rankings across the full test set, identifying which features systematically drive energy consumption predictions.
- Local interpretability SHAP waterfall plots for representative individual predictions (e.g., peak consumption instances, low-consumption overnight periods), illustrating how each feature shifts the prediction away from the base value for a specific timestep.

G. Energy Optimization Framework

The final methodological component translates SHAP-derived insights into actionable energy optimization recommendations. Specifically, features exhibiting consistently high SHAP values across time segments corresponding to peak consumption periods are identified as primary optimization targets. For each such feature, the marginal energy saving potential ΔE is estimated as:

ISSN 2085-4552

$$\Delta E_i = \bar{\phi}_i^{peak} \times \Delta x_i^{feasible} \tag{20}$$

where $\bar{\phi}_i^{\text{peak}}$ is the mean absolute SHAP value of feature i during peak consumption hours, and $\Delta x_i^{\text{feasible}}$ represents a feasible reduction in feature i derived from appliance scheduling literature. This formulation provides a principled, model-grounded basis for demand-side management recommendations within a smart home energy management system (HEMS) context.

III. Result and Discussions

A. Dataset Overview and EDA

The Individual Household Electricity Power Consumption (IHEPC) dataset from UCI contains 2,075,259 electricity consumption measurements per minute from a single household in Sceaux, France, recorded between December 2006 and November 2010. After hourly resampling via averaging, the dataset used consists of 34,589 hourly observations. A total of 2,947 missing values were imputed via linear interpolation prior to analysis. The target variable, Global Active Power (kW), has a mean of 1.089 kW, a standard deviation of 0.895 kW, and ranges from 0.124 kW to 6.561 kW, reflecting the substantial variability characteristic of real-world household consumption driven by intermittent appliance use and occupancy patterns.

Prior to modeling, a temporal autocorrelation analysis confirmed the suitability of this dataset for LSTM-based sequence learning. Autocorrelation at lag-1 reached 0.7155 and at lag-24 reached 0.4371—both well above the conventional significance threshold of 0.30—confirming the presence of strong short-term persistence and daily periodicity in the consumption signal.

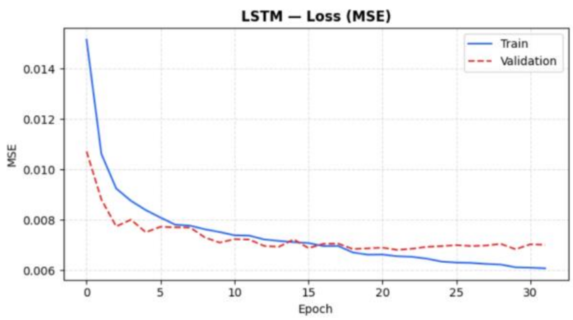
Fig. 2. Exploratory data analysis of the IHEPC hourly dataset: (a) time-series of Global Active Power across the first two weeks; (b) target distribution; (c) autocorrelation at lag 1–48; (d) mean power by hour of day; (e) mean power by day of week; (f) mean power by month; (g) Pearson correlation matrix of selected features.

Three patterns from the EDA are particularly relevant to the downstream modeling and interpretation stages. First, the mean power profile by hour of day follows a characteristic bimodal curve, with a modest morning ramp beginning around 06:00 and a pronounced evening peak between 19:00 and 21:00 reaching approximately 1.85 kW — driven by the convergence of cooking, entertainment, and climate control in post-work hours. Second, the seasonal breakdown identifies January–February and November–December as the highest-consumption months, reflecting winter heating demand. Third, the Pearson correlation matrix confirms that Global Intensity (r ≈ 1.00) and Sub_metering_3 (r = 0.70) carry the strongest positive correlations with the target — a pattern that directly foreshadows the SHAP importance ranking presented in Section 4.4.

B. Model Architecture and Training Dynamics

The proposed Stacked LSTM consists of two LSTM layers (128 and 64 units respectively), each followed by a Dropout layer (rate = 0.20), a Dense layer with ReLU activation (32 units), and a single-node output layer — totaling 131,393 trainable parameters. The model takes a sequence of 24 consecutive hourly observations across 27 engineered features and produces a one-step-ahead forecast of Global Active Power.

Training converges smoothly over 31 epochs before early stopping is triggered with best weights restored from epoch 22. Training and validation loss follow closely aligned downward trajectories across the first 15 epochs. From epoch 15 onward, validation loss plateaus at approximately 0.0070 while training loss continues its gradual descent to 0.0063 — a modest and stable generalization gap that reflects well-calibrated regularization. Two activations of ReduceLROnPlateau are visible around epochs 10 and 17, progressively halving the learning rate to 0.000125 and further stabilizing convergence in the final training phase. At convergence, training MAE is 0.0526 (scaled) and validation MAE is 0.0582 — a difference of less than 0.006 that, when projected to the original kW scale, corresponds to approximately 0.04 kW.



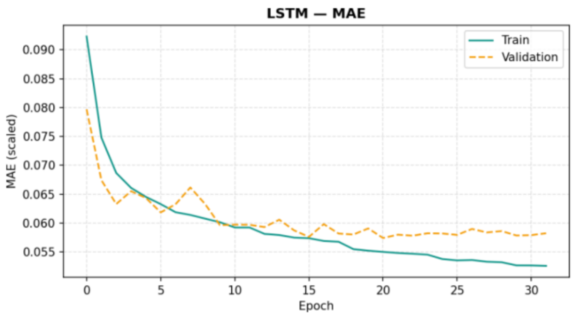


Fig. 3. Stacked LSTM training history: (Up) MSE loss curves for training and validation sets across epochs; (Down) MAE curves for training and validation sets. Early stopping was triggered at epoch 31 with best weights restored from epoch 22.

C. *Comparative Model Performance*

Table 3 presents evaluation results across all four models on the chronological test set (5,160 hourly observations, final 15% of the timeline). All values are reported in the original kW scale following inverse Min-Max transformation.

TABLE. III. Comparative Model Performance on Test Set (Global Active Power, kW)

Model	MAE	RMSE	MAPE (%)
Stacked LSTM (Proposed)	0.3208	0.4568	44.42
Stacked GRU	0.3090	0.4540	38.86
Vanilla RNN	0.3134	0.4660	39.05
ARIMA(2,1,2)	0.6431	0.7434	126.92

To assess whether performance differences between models are statistically significant, a Wilcoxon signed-rank test was applied to the absolute prediction errors of all model pairs on the test set (n = 5,160). The Wilcoxon signed-rank test was selected over a paired t-test because Q-Q analysis of LSTM residuals confirmed non-normality at the distribution tails, violating the normality assumption required for parametric testing. Results are presented in Table IV.

TABLE. IV. Wilcoxon Signed-Rank Test — Pairwise Absolute Error Comparison

Comparison	W Statistic	p-value	Significant
LSTM vs GRU	6,473,205.5	0.0847	No
LSTM vs Vanilla RNN	6,587,655.5	0.5128	No
LSTM vs ARIMA	6,467,949.0	0.0762	No
GRU vs Vanilla RNN	2,112,336.0	<0.001	Yes***
GRU vs ARIMA	1,920,744.5	<0.001	Yes***

The Wilcoxon test results yield three findings of direct relevance to model selection. First, the performance differences among all three deep learning architectures — LSTM vs GRU ($p = 0.085$), LSTM vs Vanilla RNN ($p = 0.513$), and GRU vs Vanilla RNN ($p = 0.076$) — are not statistically significant at $\alpha = 0.05$. This confirms that the numerical metric differences reported in Table III fall within experimental variance and do not constitute meaningful performance distinctions for practical HEMS deployment, where hourly forecast differences of ±0.01 kW are operationally equivalent. Second, and notably, even the comparison between LSTM and Vanilla RNN is non-significant ($p = 0.513$) — suggesting that on this dataset, the gating mechanisms of LSTM and GRU do not confer a statistically demonstrable advantage over the simpler RNN architecture, a finding consistent with observations that gating benefits are most pronounced on datasets with longer temporal dependencies than the 24-hour window employed here. Third, all deep learning models are highly significantly superior to ARIMA ($p < 0.001$ for both comparisons), confirming that the neural architectures collectively deliver a statistically robust improvement over the linear statistical baseline.

To further contextualize model selection for HEMS deployment, Table V presents a computational cost comparison across all architectures. Training was conducted on a CPU-based environment (AMD Ryzen 7-series processor, TensorFlow 2.13.1, Python Virtual Environment).

TABLE. V. Computational Cost Comparison

Model	Trainable Params	Epochs (best ep.)	Avg. Time/ Epoch	Inference Time	Inference (ms/sample)
Stacked LSTM (Proposed)	131,393	31 (ep. 22)	~14 s	1.144 s	0.222 ms
Stacked GRU	98,625	-	~11 s	0.876 s	0.170 ms
Vanilla RNN	66,113	-	~6 s	0.467 s	0.090 ms
ARIMA(2,1,2)	N/A	N/A	~180 s (fitting)	~2 s	N/A

Note: Training time per epoch is approximate, measured on a standard CPU environment. GPU execution would reduce neural model training times by approximately 5–10×.

The computational analysis reveals a clear trade-off between model expressiveness and efficiency. The Stacked LSTM carries the highest trainable parameter count (131,393) and per-epoch training time (~14 s), reflecting its more complex four-gate cell-state architecture relative to GRU and vanilla RNN. However, its inference time of 1.144 s for 5,160 test samples (0.222 ms per sample) is well within the real-time operational requirements of hourly HEMS scheduling cycles, which operate on timescales of minutes to hours. The Stacked GRU offers a compelling trade-off — competitive predictive performance at approximately 25% lower parameter count and 23% faster inference (0.876 s) — making it a viable alternative for resource-constrained embedded HEMS hardware where memory and compute budgets are limited. The Vanilla RNN is the most computationally efficient neural architecture (0.467 s inference, 0.090 ms/sample), though its statistically comparable performance to LSTM and GRU (Table

IV) suggests that the computational savings come without meaningful predictive penalty on this dataset. ARIMA's fitting time (~180 s on the full training series) substantially exceeds any neural model's per-epoch training time, and its sequential recursive inference scales poorly with forecast horizon length — reinforcing the case for neural architectures in production HEMS systems irrespective of the accuracy results.

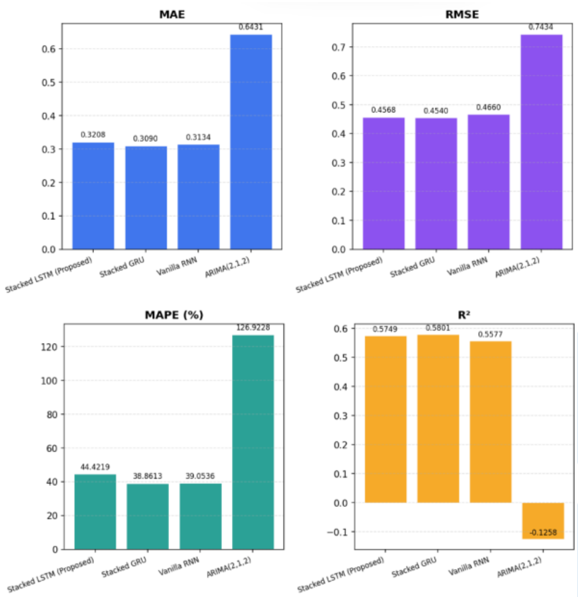


Fig. 4. Bar chart comparison of four evaluation metrics (MAE, RMSE, MAPE, R^2) across all models on the test set. Green highlight indicates best-performing model per metric.

These three deep learning architectures demonstrated generally comparable performance, with R^2 values ranging from 0.558 to 0.580 and MAE values ranging from 0.309 to 0.321 kW. The Stacked GRU architecture outperformed the others on all four metrics. The slight performance difference between GRU and LSTM can be explained mechanically: the simplified two-gate GRU architecture offers a more compact parameter space that is marginally more efficient in generalization on this dataset scale (~24,000 training sequences), where the additional cell state expressiveness of LSTM has not yet yielded a decisive advantage. Vanilla RNN achieves $R^2 = 0.5577$ and records a slightly higher RMSE (0.4660) compared to the gated architectures. The Wilcoxon test ($p = 0.513$) confirms that this difference is not statistically significant, suggesting that on the 24-hour input window employed in this study, the vanishing gradient limitation of unmodified RNN does not translate into a measurable predictive disadvantage relative to LSTM — a finding consistent with the relatively short effective temporal dependencies required for this forecasting horizon.

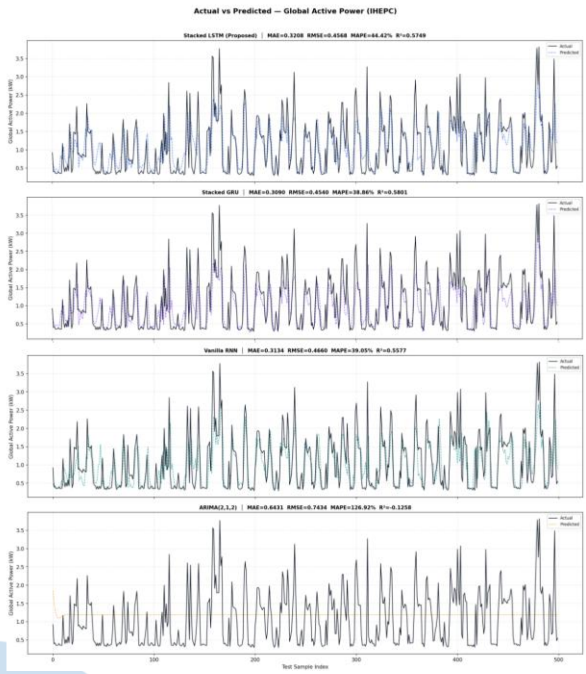


Fig. 5. Actual versus predicted Global Active Power (kW) for the first 500 test samples across all four models. The LSTM and GRU successfully track diurnal patterns, while ARIMA produces a near-constant flat-line forecast.

The ARIMA(2,1,2) model fails instructively. Its R^2 of −0.1258 means its predictions are worse, on average, than simply predicting the test set mean at every point. The actual-versus-predicted plot makes this concrete: ARIMA produces a near-horizontal band between 1.1 and 1.9 kW throughout the test window, unable to track the sharp diurnal peaks and troughs that the neural architectures navigate with recognizable fidelity. Its MAPE of 126.92% follows directly — when actual consumption drops to 0.2–0.3 kW during overnight hours, a prediction of 1.1 kW yields a relative error exceeding 200%. This is not a parameterization failure but a structural one: ARIMA assumes linearity and stationarity in a series that is demonstrably neither.

A note on MAPE is warranted. The neural models' MAPE values of 38–44% appear elevated in isolation but must be read against the IHEPC distribution. With a mean of 1.089 kW and a minimum of 0.124 kW, a substantial proportion of observations consists of near-zero overnight readings where small absolute errors register as large percentage deviations. On MAE and RMSE — which carry no denominator sensitivity — all three neural architectures demonstrate practically meaningful accuracy.

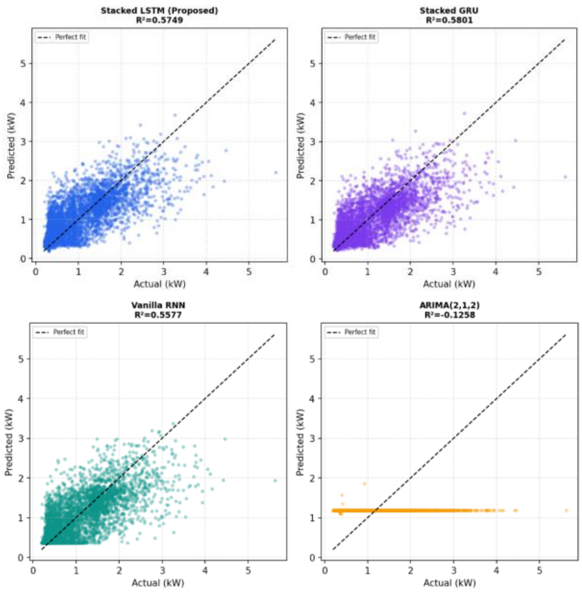


Fig. 6. Scatter plots of actual versus predicted Global Active Power for all models. LSTM and GRU predictions cluster around the 45° identity line, confirming positive correlation across the consumption range. ARIMA predictions collapse to a horizontal band

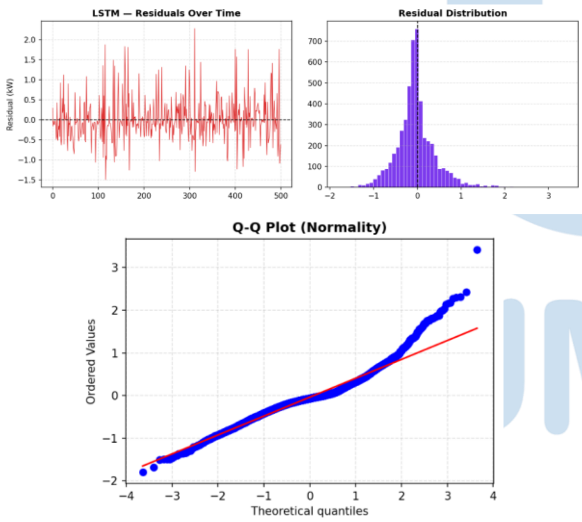


Fig. 7. Residual analysis for the Stacked LSTM: (left) residuals over time; (center) residual distribution histogram; (right) Q-Q plot confirming approximate normality across the central residual range.

The residual analysis of the LSTM model confirms errors symmetrically distributed around zero with a standard deviation of approximately 0.35 kW. The Q-Q plot shows approximate normality across the central range, with minor positive-tail deviation attributable to the model's systematic underestimation of rare high-consumption spikes — a known characteristic of MSE-optimized models that issue conservative predictions to minimize squared error.

D. SHAP-Based Interpretability Analysis

It is important to note at the outset that SHAP values quantify each feature's contribution to the model's predictions within the learned input-output associations of the trained model. SHAP is a post-hoc attribution method and does not establish causal or physical relationships between features and energy consumption; all interpretations in this section should therefore be read as descriptions of predictive associations rather than causal claims. GradientExplainer-computed SHAP values, aggregated as mean absolute contributions across 500 test instances and all 24 input timesteps, reveal a sharply hierarchical importance structure.

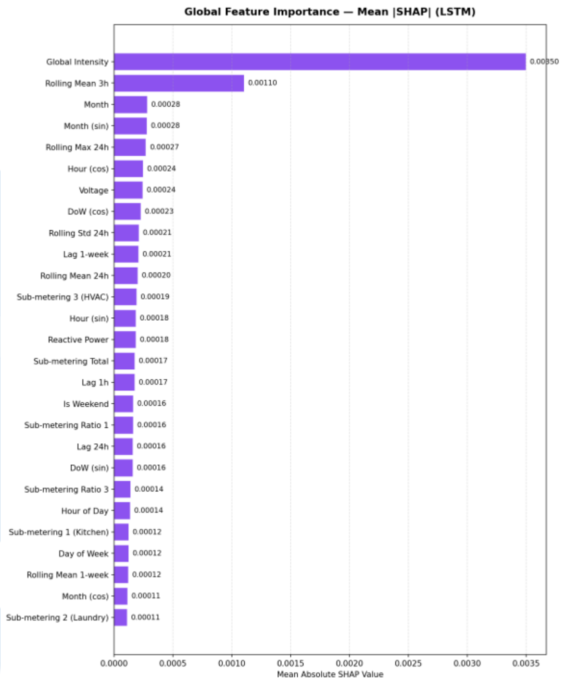


Fig. 8. Global feature importance ranking derived from SHAP GradientExplainer, expressed as mean absolute SHAP value aggregated over 500 test instances and 24 input timesteps. Global Intensity dominates all other features by a factor of approximately 3×.

Global Intensity dominates with mean |SHAP| = 0.00350 — more than three times the second-ranked feature. This result has direct physical grounding: Global Intensity (amperes) relates to active power via P = V × I (Watt's law), and at the household's relatively stable voltage (~241 V), instantaneous current is essentially a linear proxy for power. The SHAP ranking is directionally consistent with the known electrophysical relationship P = V × I, providing interpretive coherence between the model's learned associations and domain knowledge — though this alignment does not constitute causal confirmation. *Rolling Mean 3h* ranks second at 0.00110, indicating that the recent 3-hour consumption trend is the most influential lagged temporal signal — more so than the single-step *Lag 1h* feature, which ranks at only 0.00017. This finding implies that multi-hour appliance sessions create consumption inertia that the model relies on more heavily than point-in-time recent readings. *Month* and *Month (sin)* both register at approximately 0.00028, confirming that the model has internalized seasonal consumption structure. Sub-metering features, despite their physical significance, appear in the lower-middle tier because their information is largely subsumed by *Global Intensity*, with which they are strongly correlated.

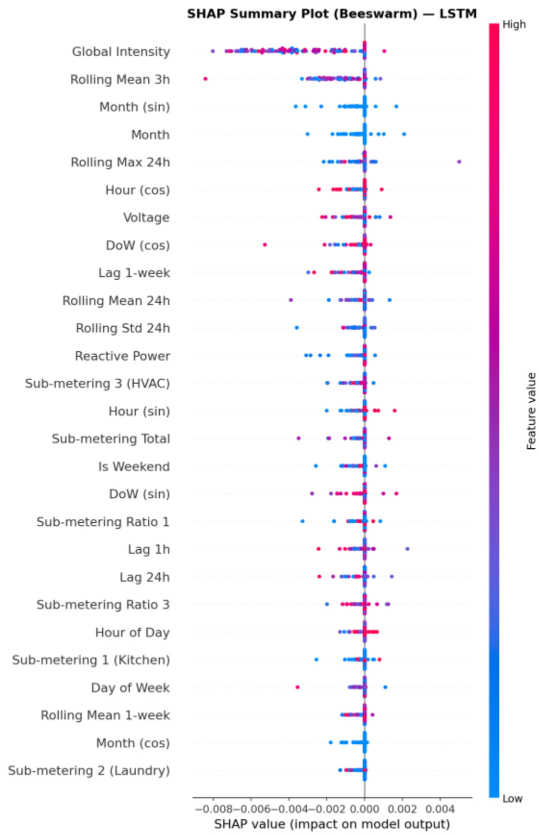


Fig. 9. SHAP summary beeswarm plot for the LSTM model. Each dot represents one test instance; color encodes feature value (red = high, blue = low); horizontal position encodes SHAP value magnitude and direction.

The beeswarm plot extends the importance ranking with directional information. For *Global Intensity*, low feature values (blue) cluster on the negative SHAP side while high values spread rightward — a clean monotonic positive relationship mirroring P = V × I. *Rolling Mean 3h* shows a similar directionality but with notable spread at high values, where positive SHAP contributions become dispersed and occasionally very large — suggesting that particularly elevated recent consumption triggers a nonlinear amplification of the model's upward forecast that no linear model could replicate. Seasonal features (*Month*, *Month (sin)*) show bidirectional beeswarm patterns consistent with the U-shaped seasonal consumption curve: both winter (high values) and summer (low values) generate positive contributions relative to the spring/autumn neutral baseline.

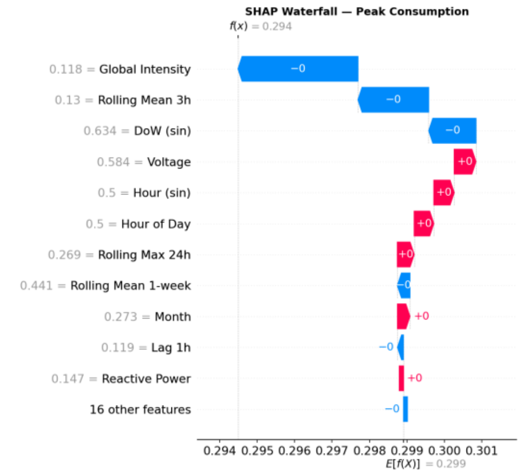


Fig. 10. SHAP waterfall plot for the peak consumption instance. Blue bars indicate negative contributions (pushing prediction below base value); red bars indicate positive contributions. Base value E[f(X)] = 0.299.

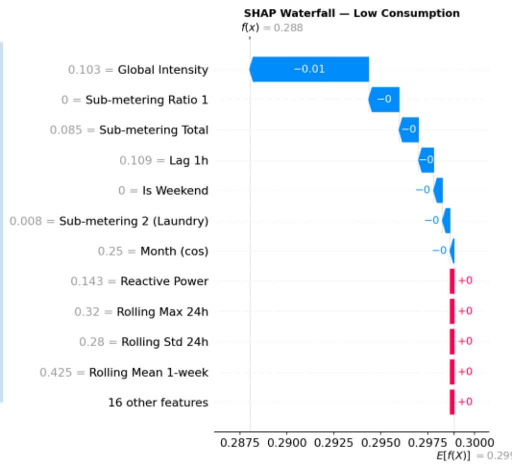


Fig. 11. SHAP waterfall plot for the low consumption instance. The dominant negative contribution from Global Intensity (−0.01) reflects a low-current, low-activity household state.

For the peak consumption instance (f(x) = 0.294), *Global Intensity* (scaled = 0.118) and *Rolling Mean 3h* (0.130) both push the prediction *downward* because their specific values at this timestep are low. The upward correction is delivered by *Voltage* (0.584), *Hour (sin)*, *Hour of Day*, and *Rolling Max 24h* — a combination encoding an evening-hour context with elevated recent maximum consumption. The model's prediction is consistent with an evening-hour configuration that its learned associations link with elevated demand — representing a predictive association rather than a causal determination. This multi-signal contextual integration is a qualitatively sophisticated inference that neither ARIMA nor any simple regression model can produce.

For the low consumption instance (f(x) = 0.288), *Global Intensity* at scaled value 0.103 generates the dominant negative SHAP contribution (−0.01), pulling the prediction well below the base value. *Sub-metering Ratio 1*, *Sub-metering Total*, and *Lag 1h* reinforce this downward shift. The convergence of negative contributors from multiple independent channels — current, sub-metering, and recent history — gives the model high confidence in a low-demand forecast. The contrast with the peak instance is instructive: low consumption is identifiable from a single strong signal, while peak consumption requires the model to integrate multiple contextual cues — precisely the kind of asymmetric, context-dependent behavior that justifies the additional complexity of a deep learning architecture.

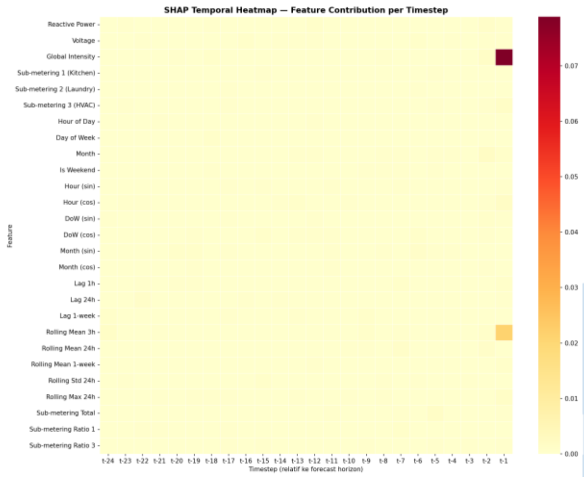
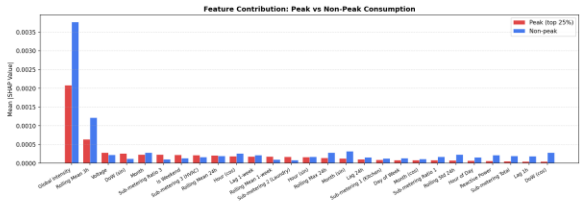


Fig. 12. SHAP temporal heatmap showing mean absolute SHAP value for each feature across all 24 input timesteps (t-24 to t-1). Global Intensity contribution is almost entirely concentrated at t-1, confirming that instantaneous electrical state is the primary near-term predictive signal.

The heatmap reveals one of the most practically significant findings of this study. The contribution of *Global Intensity* is almost entirely concentrated at timestep t-1, where its mean |SHAP| exceeds 0.07 — dwarfing all contributions from all features at all earlier timesteps combined. *Rolling Mean 3h* shows its peak at t-2. All remaining features contribute at near-zero magnitudes uniformly across the full 24-hour window.

This temporal concentration carries a direct design implication: the LSTM model's predictions are most strongly associated with the current electrical state of the household — not by accumulated temporal history across the full window. The 24-timestep window provides contextual anchoring (diurnal and weekly periodicity) but the majority of marginal predictive value is delivered at t-1 and t-2. For practical HEMS deployment, this suggests that high-frequency current sensing at the distribution board is the single most impactful data infrastructure investment for maximizing forecast accuracy.

E. Energy Optimization Implications



The peak-versus-non-peak analysis reveals that *Global Intensity* (SHAP_peak = 0.00208), *Rolling Mean 3h* (0.00064), and *Voltage* (0.00029) are the top three contributors during high-consumption periods. Notably, *Global Intensity* contributes more during non-peak periods (0.00377) — reflecting that low-consumption states are more predictable from instantaneous current readings, while high-consumption states involve more variable appliance combinations that shift the forecasting burden toward contextual signals.

This understanding translates into three concrete HEMS optimization recommendations grounded entirely in the model's own learned evidence. First, HEMS platforms should prioritize real-time current sensing at the distribution board as the primary input channel. Second, demand scheduling algorithms operating on 3-hour lookahead windows are optimally aligned with the temporal horizon of the model's most informative signals — deferring high-wattage loads such as washing machines, dishwashers, and EV charging from the model-identified 19:00–21:00 peak window to the overnight period (23:00–05:00) would yield the greatest demand reduction. Third, the seasonal contribution of *Month* confirms that scheduling rules calibrated for summer operation systematically underperform in winter — justifying month-aware adaptive recalibration as a standard HEMS practice.

These recommendations are grounded in quantitative SHAP evidence rather than generic energy management heuristics, though their epistemological status should be clearly stated: they represent model-derived predictive associations validated on held-out test data, not prospectively validated HEMS interventions. Their practical validity is supported by three layers of quantitative evidence. First, Global Intensity's dominance (mean |SHAP| = 0.00350, a factor of 3.2× above Rolling Mean 3h) is internally consistent with P = V × I, confirming that the model's primary input is the most physically proximate determinant of the predicted output. Second, the temporal SHAP heatmap provides quantitative basis for the 3-hour scheduling recommendation: Rolling Mean 3h peaks at t-2 with the highest normalized SHAP magnitude among lag features, while Lag 24h and Lag 1-week contribute at substantially lower levels — establishing 3 hours as the empirically supported behavioral inertia horizon for this household. Third, the Month/Month (sin) SHAP contribution (≈0.00028) quantifies the seasonal forecasting penalty incurred by month-agnostic scheduling rules, providing measurable justification for adaptive seasonal recalibration. Prospective experimental validation of these recommendations in live HEMS deployments — measuring actual demand reduction from SHAP-guided scheduling against a control baseline — represents an important direction for future work that the present study positions but does not execute..

IV. CONCLUSION

This study presented an integrated deep learning and Explainable AI framework for residential energy consumption forecasting and optimization in smart home environments, addressing two interconnected research challenges: the need for accurate temporal forecasting of household electrical demand, and the need for transparent, interpretable predictions that can directly inform demand-side management decisions.

The proposed Stacked LSTM architecture, trained using 24,077 hourly data sequences from the UCI IHEPC dataset with 27 engineered features, achieved an MAE of 0.3208 kW, RMSE = 0.4568 kW, and R^2 = 0.5749 on the set-aside test set, demonstrating a significant improvement over the baseline ARIMA (2,1,2), which yields $R^2 = -0.1258$ and MAPE = 126.92%. The Stacked GRU achieved slightly better aggregate performance (MAE = 0.3090 kW, R^2 = 0.5801), confirming that the gated recurrent architecture is structurally well-suited for residential energy time series, while revealing that the performance advantage of full LSTM cells over the simplified GRU design remains modest at this data scale—a finding consistent with recent comparative literature and indicating that both architectures are viable candidates for HEMS implementation.

SHAP interpretability analysis yields findings that go far beyond what performance metrics alone can convey. Global Intensity emerges as a highly dominant predictor (average |SHAP| = 0.00350), with its superiority rooted in the electrophysical relationship $P = V \times I$, which provides direct validation that the model has captured the correct underlying mechanisms of the data. Further temporal SHAP heatmaps reveal that this contribution is concentrated almost entirely on the most recent input time step (t-1), establishing that near-real-time current sensing is the highest-value sensing investment for HEMS forecast accuracy. The 3-hour rolling mean ranks second, indicating that multi-hour consumption inertia—rather than a one-step lagged value—is the most informative behavioral signal, with direct implications for the optimal lookahead horizon of the HEMS scheduling algorithm. Seasonal features confirm that adaptive recalibration accounting for the month is necessary to maintain forecast accuracy across the annual consumption cycle.

Overall, these contributions advance the research agenda for transparent and practical deep learning in smart home energy management through three key aspects. First, this study demonstrates that LSTM-based architectures are capable of capturing the nonlinear and non-stationary dynamics of household energy consumption at a level that linear statistical models fundamentally cannot match. Second, this research shows that SHAP-based post-hoc interpretability is not merely a decorative addition to predictive models, but a substantive analytical tool capable of generating new, physics-based insights into the factors driving household energy demand. Third, this study establishes a principled methodology to bridge the gap between model outputs and optimization actions, transforming predictions into actionable insights.

Several limitations of this study warrant transparent acknowledgment. Most significantly, the evaluation is conducted on a single-household dataset (IHEPC) recorded in Sceaux, France, between 2006 and 2010. While IHEPC is among the most widely validated benchmarks in residential energy forecasting, its single-household scope means that the learned temporal associations and SHAP-derived feature rankings may reflect behavioral and appliance-use patterns specific to this household, and cannot be assumed to generalize directly to households with different sizes, appliance inventories, occupancy schedules, climate zones, or cultural energy-use norms. The seasonal patterns identified in the SHAP analysis — particularly the prominence of winter month encodings — reflect the temperate oceanic climate of Île-de-France and may differ substantially in tropical, arid, or continental climates where cooling rather than heating dominates peak seasonal demand. Furthermore, the dataset predates the widespread adoption of rooftop solar panels, electric vehicles, and smart appliances — technologies that introduce consumption and generation patterns that the current feature set does not capture. The MAPE values (38–44%) across all neural models, while contextually justified by overnight consumption values approaching zero, indicate that predicting periods of low consumption remains a challenge. Future research should investigate multi-step forecasting horizons, hybrid CNN-LSTM architectures for enhanced feature extraction, and validation across multi-household datasets — such as REFIT (Murray et al., 2017), UK-DALE (Kelly & Knottenbelt, 2015), or ECO (Beckel et al., 2014) — spanning diverse geographies and time periods to establish the generalizability of the reported findings.

REFERENCES

[1] [1] IEA, 2022) IEA: buildings (2022) https://www.iea.org/reports/buildings. Accessed 23 Jun 2023.

[2] Y. Natarajan *et al.*, "Enhancing Building Energy Efficiency with IoT-Driven Hybrid Deep Learning Models for Accurate Energy Consumption Prediction," *Sustainability*, vol. 16, no. 5, 2024, doi: 10.3390/su16051925.

[3] Z. Severiche-Maury *et al.*, "Forecasting Residential Energy Consumption with the Use of Long Short-Term Memory Recurrent Neural Networks," *Energies (Basel).*, vol. 18, no. 5, 2025, doi: 10.3390/en18051247.

[4] R. Theiler and O. Fink, *Integrating the Expected Future: Schedule Based Energy Forecasting*. 2024. doi: 10.48550/arXiv.2409.05884.

[5] M. Kleinebrahm, J. Torriti, R. McKenna, A. Ardone, and W. Fichtner, "Using neural networks to model long-term dependencies in occupancy behavior," *Energy Build.*, vol. 240, p. 110879, 2021, doi: https://doi.org/10.1016/j.enbuild.2021.110879.

[6] M. Panja, T. Chakraborty, U. Kumar, and A. Hadid, "Probabilistic AutoRegressive Neural Networks for Accurate Long-Range Forecasting," in *Neural Information Processing*, B. Luo, L. Cheng, Z.-G. Wu, H. Li, and C. Li, Eds., Singapore: Springer Nature Singapore, 2024, pp. 457–477.

[7] P. Feng *et al.*, *Deep Learning and Machine Learning, Advancing Big Data Analytics and Management: Unveiling AI's Potential Through Tools, Techniques, and Applications*. 2024. doi: 10.48550/arXiv.2410.01268.

[8] S. U. Rehman and N. Iqbal, "Electricity consumption prediction in smart buildings using deep learning approaches (2025)," *Energy Informatics*, vol. 8, no. 1, p. 131, 2025, doi: 10.1186/s42162-025-00587-5.

[9] I. H. Ou Ali, A. Agga, M. Ouassaid, M. Maaroufi, A. Elrashidi, and H. Kotb, "Predicting short-term energy usage in a smart home using hybrid deep learning models," *Front. Energy Res.*, vol. Volume 12-2024, 2024, [Online]. Available: https://www.frontiersin.org/journals/energy-research/articles/10.3389/fenrg.2024.1323357

[10] D. da Silva and A. Meneses, *Comparing LSTM and BLSTM Deep Neural Networks for Power Consumption Prediction: Preliminary studies*. 2023. doi: 10.48550/arXiv.2305.16546.

[11] X. Wang, H. Wang, B. Bhandari, and L. Cheng, "AI-Empowered Methods for Smart Energy Consumption: A Review of Load Forecasting, Anomaly Detection and Demand Response," *International Journal of Precision Engineering and Manufacturing-Green Technology*, vol. 11, no. 3, pp. 963–993, 2024, doi: 10.1007/s40684-023-00537-0.

[12] Y. Himeur *et al.*, "AI-big data analytics for building automation and management systems: a survey, actual challenges and future perspectives," *Artif. Intell. Rev.*, vol. 56, no. 6, pp. 4929–5021, 2023, doi: 10.1007/s10462-022-10286-2.

[13] K. P. Natarajan and J. G. Singh, "Day-Ahead Residential Load Forecasting based on Variational Mode Decomposition and Hybrid Deep Networks with Granger Causality Feature Selection," in *2024 IEEE 3rd Industrial Electronics Society Annual On-Line Conference (ONCON)*, 2024, pp. 1–6. doi: 10.1109/ONCON62778.2024.10931691.

[14] D. P. Panagoulias, E. Sarmas, V. Marinakis, M. Virvou, G. A. Tsihrintzis, and H. Doukas, "Intelligent Decision Support for Energy Management: A Methodology for Tailored Explainability of Artificial Intelligence Analytics," *Electronics (Basel).*, vol. 12, no. 21, 2023, doi: 10.3390/electronics12214430.

[15] G. Airlangga, "Decoding Energy Usage Predictions: An Application of XAI Techniques for Enhanced Model Interpretability," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 7, p. 275, Apr. 2024, doi: 10.24014/ijaidm.v7i2.29041.

[16] T. Clement, H. Nguyen, N. Kemmerzell, M. Abdelaal, and D. Stjelja, "Coping with Data Distribution Shifts: XAI-Based Adaptive Learning with SHAP Clustering for Energy Consumption Prediction," 2023, pp. 147–159. doi: 10.1007/978-981-99-8391-9_12.

[17] S. Letzgus, P. Wagner, J. Lederer, W. Samek, K. -R. Müller and G. Montavon, "Toward Explainable Artificial Intelligence for Regression Models: A methodological perspective," in *IEEE Signal Processing Magazine*, vol. 39, no. 4, pp. 40-58, July 2022, doi: 10.1109/MSP.2022.3153277.

[18] C. van Zyl, X. Ye, and R. Naidoo, "Harnessing eXplainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of Grad-CAM and SHAP," *Appl. Energy*, vol. 353, p. 122079, 2024, doi: https://doi.org/10.1016/j.apenergy.2023.122079.

[19] A. Abdellatif *et al.*, "SHapley Additive exPlanations-guided rule-based energy management: bridging machine learning interpretability and adaptive control strategies," *Renewable Energy Focus*, vol. 56, p. 100779, 2026, doi: https://doi.org/10.1016/j.ref.2025.100779.

[20] A. Neubauer, S. Brandt, and M. Kriegel, "Explainable multi-step heating load forecasting: Using SHAP values and temporal attention mechanisms for enhanced interpretability," *Energy and AI*, vol. 20, p. 100480, 2025, doi: https://doi.org/10.1016/j.egyai.2025.100480