COMPUTER SCIENCE
КОМПЬЮТЕРЛІК ҒЫЛЫМДАР
КОМПЬЮТЕРНЫЕ НАУКИ

UDC 004.855
IRSTI 44.01.77

https://doi.org/10.55452/1998-6688-2026-23-2-133-149

[1]*Tokhmetov A.,
Cand. Phys.-Math. Sc., Associate Professor, ORCID ID: 0000-0003-0764-8574,
*e-mail: tokhmetov_at_2@enu.kz
[1]Serikbayeva S.,
PhD, Associate Professor, ORCID ID: 0000-0002-3627-3321,
e-mail: inf_8585@mail.ru
[1]Tanchenko L.,
MSc, ORCID ID: 0000-0002-6811-2303,
e-mail: ltanchenko@mail.ru
[1]Kenesbay M.
Master's student, ORCID ID: 0009-0000-2121-089X,
e-mail: mikam4965@gmail.com

[1]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan

AI-BASED ENERGY FORECASTING AND IMPROVED DEMAND MANAGEMENT IN SMART HOMES

Abstract

This paper presents a comprehensive multi-stage system designed to improve the accuracy of energy load forecasts and evaluate the effectiveness of both forecast models and demand response (DR) strategies. Using the REFIT dataset, a comparative analysis of a hierarchy of forecast models was conducted, including linear regression, random forest, SVR, k-NN, LSTM, and a hybrid encoder-decoder with an attention mechanism. The results of the study indicated that the developed hybrid encoder-decoder model with an attention mechanism achieved the best accuracy ($R^2 = 0.91$, MAPE = 2.39%), demonstrating excellent ability to capture complex temporal patterns in the data. Rigorous multi-stage testing confirmed the stability and high generalizability of this deep learning model. The highly accurate forecast was incorporated into a mixed integer linear programming (MILP)-based model for home energy management system (HEMS) optimization. The results indicated that this complex framework significantly reduced energy costs by 28.7% and reduced peak load by 37.1% through optimal appliance scheduling. This work demonstrates how to effectively combine state-of-the-art artificial intelligence (AI)-based forecasting with formal energy optimization in a single, comprehensive system. This method not only allows for more accurate consumption forecasting, especially during peak hours, but also demonstrates that AI can significantly improve the flexibility of energy networks and the energy efficiency of smart homes.

Keywords: energy forecasting, demand response, smart homes, machine learning, deep learning, LSTM, hybrid model.

Received November 16, 2025; revised February 26, March 5, April 8, 2026; accepted April 20, 2026.

Introduction

For efficient use of renewable energy sources, proper distribution of load in the network, and implementation of energy-saving programs (Demand Response, DR), it is critical to accurately predict the energy needs of consumers. Short-term demand forecasting in the residential sector is important for preserving system stability and improving energy system efficiency, given the explosive growth of distributed energy resources and the electrification of end-use sectors [1, 2]. In load estimation, traditional statistical methods and regression methodologies have gained widespread use [3, 4]. However, due to their limited ability to capture short-term fluctuations and nonlinear dynamics, machine learning (ML) and deep learning (DL) approaches have recently grown in popularity [5, 6].

With the global push for sustainable energy systems, accurate forecasting of household electricity demand has become a constant requirement for balancing supply and demand, integrating renewable energy sources, and reducing grid instability. Specialized energy management techniques, such as peak shaving, time-of-day demand shifting, and flexible tariff systems, allow households and businesses to reduce electricity use during peak hours. These techniques help reduce energy costs and minimize environmental impacts [7, 8]. But for all these methods to work effectively, we need accurate algorithms that can predict in advance how much energy we will actually need.

This study examines how to predict hourly energy consumption accurately in smart homes. We analyzed real-world data from 20 UK households over a two-year period. The goal was to test various machine and deep learning models, such as linear regression, random forest, LSTM, and others, to select the most accurate model. Various factors were considered, including calendar data, consumption history, and a multi-level validation of the models. Accurate forecasts allow for an assessment of how to optimize energy consumption, reduce peak load, and make the system more resilient and cost-effective.

Energy forecasting and demand management have been transformed by the advent of artificial intelligence, thanks to machine learning and deep learning methods. Recent research shows that these approaches perform better than older statistical models because they can account for complex and unstable fluctuations in energy consumption patterns [9, 10].

Forecasting models have advanced significantly, moving from simple machine learning algorithms to more complex neural networks and deep learning architectures. While standard models such as random forests and support vector machines provide a robust foundation, recent research has concluded that recurrent neural networks, particularly long short-term memory (LSTM) networks, are particularly well suited for time series forecasting due to their ability to capture temporal dependencies [11, 12].

Many studies show that complex models–for example, neural networks and their hybrid variants–re effective at predicting energy consumption [13, 14]. They are able to recognize complex patterns and account for long-term dependencies, especially when it comes to unstable loads in buildings, which depend on human behavior and external factors such as weather [15, 16].

In some studies, simple models–such as linear regression or basic ensembles–have demonstrated unexpectedly high accuracy, sometimes even better than complex neural networks [17, 18]. This can occur when training data is sparse, the features describe the problem well, or the results are strongly influenced by external factors such as temperature, holidays, or day of the week. Such cases highlight the importance of carefully comparing models and ensuring that high accuracy is not a fluke but a genuine result.

It is important to remember that an accurate forecast is not an end in itself but a tool for optimizing energy management. The primary goal of a home energy management system (HEMS) is to use these forecasts to make intelligent and cost-effective decisions. While such systems previously often relied on simple heuristics, today more flexible optimization methods are increasingly being used to achieve a better balance between comfort, savings, and sustainability. For example, mixed-integer linear programming (MILP) is now widely used for load planning, creating optimal operating plans that accurately balance energy costs, peak demand penalties, and user comfort [19, 20].

While the encoder–decoder architecture with attention has been established in the time-series literature [16], the present work's contribution is distinct: we demonstrate, on a real-world residential dataset (REFIT), that the quality of probabilistic temporal modeling directly and measurably improves the outcomes of downstream formal MILP-based scheduling. Existing studies address either the forecasting component or the optimization component in isolation; our work provides a reproducible, end-to-end validated pipeline connecting both and quantifies the operational benefit of forecast accuracy improvements in terms of energy cost and peak demand reduction.

Materials and methods

This study presents a comprehensive, multi-step process designed to systematically compare machine learning and deep learning models for energy forecasting and evaluate their applicability for demand response (DR) modeling. The research process, illustrated in Figure 1, includes steps from data preparation to final model evaluation and interpretation.

The REFIT dataset containing data on 20 households in the UK, was used as the data source [21]. To ensure replicability, all experiments were conducted on time series data for individual households (e.g., TimeSeriesVariable2).

A standardized data preprocessing algorithm was developed:

1. Loading and resampling: The raw data, recorded at a high frequency (in watts, W), was aggregated into hourly intervals (1H). The mean (.mean ( )) was used as the aggregation function. This allowed us to obtain the average hourly power, which for an hourly interval is numerically equivalent to energy consumption in watt-hours (Wh), which is our target unit of measurement.

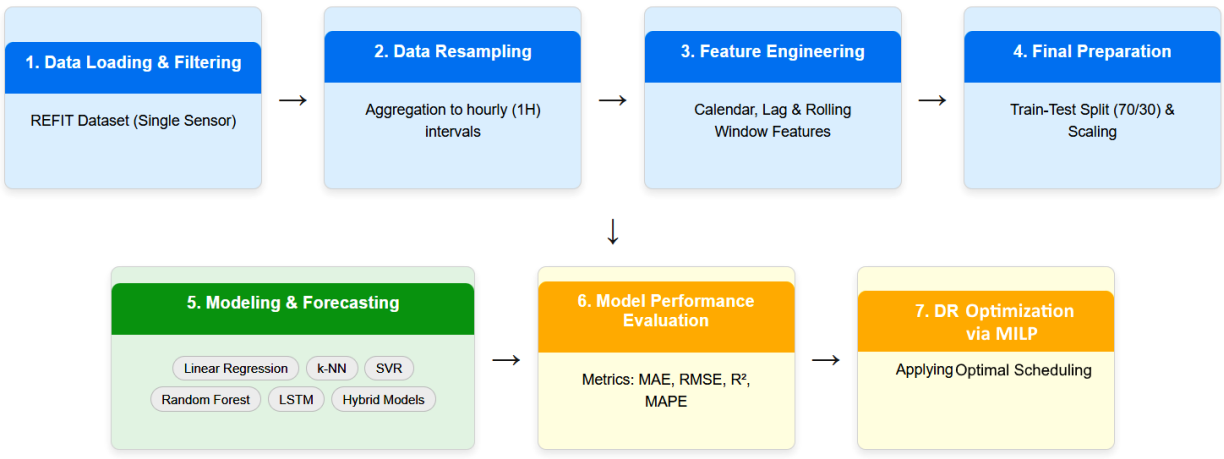

Figure 1 – The process of data processing and energy consumption forecasting

2. Gaps handling: Short gaps (up to 3 consecutive hours) were filled using linear interpolation to maintain the continuity of the time series. Segments with longer data gaps were removed from the dataset to avoid significant distortions.

3. Anomaly handling: A moving average method was used to detect and smooth out anomalous peaks caused by possible sensor failures. Values deviating more than three standard deviations from the moving average (over a 24-hour window) were limited by this threshold.

4. Design of objects: three types of objects were created:

- Calendar: time of day, day of the week;
- Lags: consumption values 1 and 24 hours before the current moment. A lag of 24 hours was chosen to allow the model to take into account daily seasonality;
- Sliding window: average consumption over the last 24 hours to identify overall trends.

5. For LSTM and hybrid models, the data was transformed into sequences, where data from the previous 24 hours was used to forecast one hour ahead.

6. Scaling: All features and the target variable were scaled to the range [0, 1] using MinMaxScaler.

A total of six models were tested, including two deep learning architectures that have been documented in detail:

- For classical models (linear regression, k-NN, SVR, and random forest (100 trees)), standard implementations from scikit-learn were used.
- The LSTM model architecture consisted of one LSTM layer (50 neurons, tanh activation), a Dropout layer (0.2 coefficient), and a dense output layer. The model was trained for 30 epochs using the Adam optimizer (learning rate = 0.001).
- The hybrid encoder-decoder model with an attention mechanism is the most complex architecture, consisting of an encoder–an LSTM layer (128 neurons) that reads a 24-hour sequence; a decoder–the second LSTM layer (128 neurons) that generates the prediction; and the attention mechanism (allows the decoder to focus on the most important steps in the input sequence). The model was trained for 40 epochs using the Adam optimizer.

The performance of all forecasting models was assessed using four standard regression metrics: mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), and the coefficient of determination (R^2) which served as the main metric for comparison [22–25].

To ensure interpretability of the results in the original physical units, all model predictions obtained in the [0, 1] range were converted back to the original watt-hour (Wh) scale by inverse transformation before calculating these metrics. This allows for a direct assessment of the model's error in real-world conditions.

Selecting optimal hyperparameters for deep learning models is key to achieving high accuracy. This study used an iterative approach based on validation results. A grid search was conducted for the key models (LSTM and hybrid) to determine the optimal architecture. The key hyperparameters examined and their final values are presented in Table 1.

Table 1 – Selecting hyperparameters for deep learning models

Hyperparameter	Model	Range	Selected value	Justification
Number of neurons in a layer	LSTM	[32, 50, 64, 128]	50	The model provided a better balance between model complexity and overfitting on the validation set.
	Hybrid	[64, 128, 256]	128	The value 128 showed a significant improvement over 64, while 256 did not improve accuracy but increased training time.

Continuation of table 1

Number of training epochs	LSTM	[20, 30, 50]	30	After 30 epochs, the error on the test sample stopped decreasing significantly (early stopping).
	Hybrid	[30, 40, 60]	40	Similarly, 40 epochs were found to be sufficient for the model to converge without any signs of overfitting.
Window size	Both	[12, 24, 48]	24	The 24-hour window proved to be the most effective, as it covers the entire daily consumption cycle.
Optimizer	Both	Adam, RMSprop	Adam	The Adam optimizer showed the fastest and most stable convergence during experiments.
Attrition rate	LSTM	[0.1, 0.2, 0.3]	0.2	It effectively prevents overfitting without compromising the accuracy of the training set.

To ensure the reliability of the results, rigorous verification methods were applied:

1. To eliminate "statistical artifacts," a chronological division into training and test samples (70/30) was applied, and the scaler was trained exclusively on training data.

2. Group validation: To assess generalizability, the LSTM model was trained and tested on a sample of 10 different households. To demonstrate stability, the mean values and standard deviations of the parameters were calculated.

3. Time series cross-validation: To objectively assess performance over time, the data for a single household were split into five consecutive time slices (TimeSeriesSplit). The model was trained on an increasing time interval and tested on the next time slice.

4. Baseline forecast and statistical tests: To confirm the effectiveness, the LSTM model forecasts were compared with the naive baseline forecast (the forecast equals to the value 24 hours ago). The statistical significance of the improvement was tested using the Diebold-Mariano test, which compares the forecast errors of the two models.

To go beyond heuristic modeling and quantify the practical impact of forecasting, a formal home energy management system (HEMS) was developed using mixed-integer linear programming (MILP). This approach enables dynamic optimization of household energy consumption by scheduling controllable loads to minimize costs while taking into account operational constraints and user comfort requirements. The primary inputs to the MILP model are accurate hour-ahead forecasts of the uncontrollable load, generated by a hybrid encoder-decoder with an attention model.

The optimization problem was formulated as follows:

Objective function Z:

The main objective is to minimize the total daily operating cost, which is a weighted sum of three components: the cost of electricity purchased from the grid, a penalty for high peak power demand, and a penalty for user inconvenience.

$$Z = (\sum_{i \in T} c_t \cdot P_{grid,t} + \lambda_{peak} \cdot P_{peak} + \sum_{i \in I} \lambda_{delay,i} \cdot D_i) \rightarrow min \quad (1)$$

where T is a set of hourly time intervals in a 24-hour optimization horizon;

c_t – the price of electricity at time *t*;

$P_{grid,t}$ – the power consumed from the network at time the *t*;

P_{peak} – maximum network power on the horizon;

D_i – delay in the start time of device *i* relative to the time preferred by the user;

λ_{peak} and $\lambda_{delay,i}$ – represent penalty coefficients that allow finding a compromise between cost, peak load reduction, and user comfort.

Main limitations:

1. Power balance: for each time interval t, the total energy supply must be equal to the total demand.

$$P_{grid,t} + P_{pv,t} = L_{base,t} + \sum_{i \in I} L_{ctrl,i,t}$$

where $P_{grid,t}$ is the solar energy generation forecast (if available);

$L_{base,t}$ – the forecast of uncontrolled load from a hybrid model;

$L_{ctrl,i,t}$ – the consumption of the controlled device.

2. Appliance Scheduling: Continuous appliances (e.g., washing machine, dishwasher) are modeled using binary variables to ensure that their work cycle is completed without interruption within a user-specified time window.

3. Grid and system limitations: The power consumed from the grid at any given time t cannot exceed the maximum contractual limit.

The MILP model was implemented in Python with the Pyomo library and solved using the GLPK (GNU Linear Programming Kit). The process of integrating forecasts into MILP is as follows:

1. Run cycle: The optimization model runs once a day (e.g., at 00:00).

2. Forecast generation: At startup, the system takes the last available 24 hours of actual consumption (from t-23 to t-0) and uses the trained hybrid model to generate an unsupervised load forecast () for the next 24 hours (from t+1 to t+24).

3. Passing to the optimizer: The resulting 24-value forecast vector is passed to the MILP model as a fixed parameter. The optimizer solves the problem, treating this forecast as deterministic.

4. Plan generation: the MILP model generates an optimal schedule for the operation of controlled devices for the next 24 hours.

Results and discussion

We first identified the best model based on the accuracy of its baseline forecast, and then evaluated the performance of DR strategies based on this best forecast.

Stage 1: Assessing the accuracy of the model's forecasts. In the first stage, we determine which model produces the most accurate forecasts. Table 2 summarizes the results for all six models.

Table 2 – Comparative accuracy of models (baseline forecast)

Model	MAE (Wh)	RMSE (Wh)	R^2	MAPE (%)
Linear Regression	0.99	1.45	0.47	5.52
k-NN	0.90	1.23	0.66	4.36
SVR	0.80	1.39	0.57	3.75
Random Forest	0.53	0.82	0.85	2.56
LSTM	0.60	0.74	0.87	3.00
Hybrid (ours)	0.47	0.61	0.91	2.39

The hybrid model with an encoder-decoder architecture and an attention mechanism outperformed all other models across key metrics: MAE = 0.47, RMSE = 0.61, MAPE = 2.39%, and the highest R2 = 0.91. The LSTM model followed with R2 = 0.87. Among classical methods, Random Forest ranked third (R2 ≈ 0.85), while SVR, k-NN, and Linear Regression showed limited performance (R2 = 0.47–0.66). These results confirm that architectures specifically designed for working with time series perform best. Figure 2 confirms this.

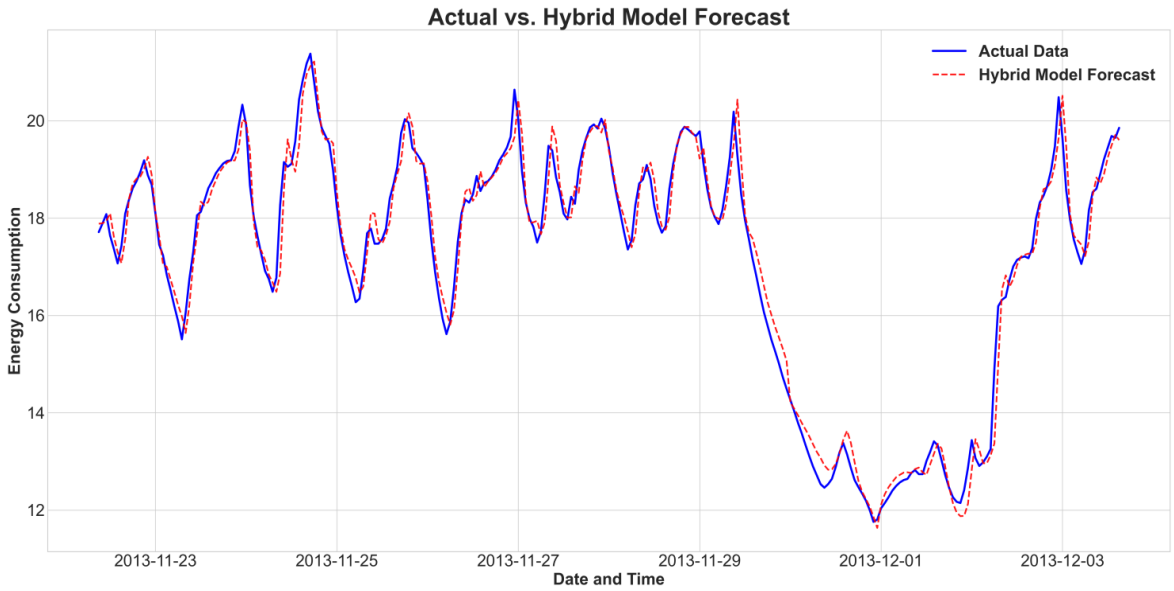

Figure 2 – Comparison of the hybrid model's forecast (red line) against
the actual energy consumption data (blue line) for a 300-hour segment of the test set

To test how well the best models fit real-world data, we conducted a group validation for the LSTM model (which is less computationally expensive than the hybrid model). The summary statistics in Table 3 provide an overview of the robustness and generalization abilities.

Table 3 – Summary statistics for the LSTM model performance

Key indicators	MAE (Wh)	RMSE (Wh)	R^2	MAPE (%)
count	10	10	10	10
mean	0.46	0.61	0.91	2.37
std	0.13	0.16	0.04	0.76
min	0.28	0.35	0.85	1.31
25%	0.37	0.48	0.89	1.85
50%	0.45	0.63	0.90	2.19
75%	0.57	0.72	0.92	2.77
max	0.69	0.82	0.96	3.62

The group validation across 10 households confirmed the model's robustness: the mean R2 was 0.91 ± 0.04, and even in the worst case (min R2 = 0.85), performance remained strong. The low standard deviation across all metrics indicates high generalizability across different consumption profiles.

To confirm the effectiveness of the developed hybrid model, its forecasts were compared with the baseline seasonal naive forecast model, where the hourly forecast is equal to the actual value 24 hours earlier. The comparison results are presented in Table 4.

The results clearly demonstrate that the hybrid model generates significantly more accurate forecasts, with R2 = 0.91 compared to 0.64 for the baseline model, and MAPE more than halved from 5.16% to 2.39%. The Diebold-Mariano test confirmed the statistical significance of this improvement (DM = -21.45, $p < 0.001$).

Table 4 – Comparison of performance of hybrid and base models

Model	MAE (Wh)	RMSE (Wh)	R^2	MAPE (%)
Hybrid	0.47	0.61	0.91	2.39
Naive Seasonal Baseline	0.99	1.23	0.64	5.16

Stage 2: Evaluating the effectiveness of energy reduction (DR) strategies. We demonstrated the usefulness of our high-precision prediction model by integrating its results into a MILP-based home energy management system (HEMS). While we demonstrated the overall robustness using LSTM (Table 3), the best-performing hybrid model (Encoder-Decoder with Attention) was used for the key HEMS optimization task.

To ensure the model's stability and accuracy, we tested it on data from five different households. Forecast errors were consistently low: MAEs ranged from 0.22 to 0.29 Wh, and RMSEs ranged from 0.28 to 0.38 Wh (Table 5). This stability is crucial, as it demonstrates that the HEMS optimization system operates on a robust foundation.

Table 5 – Group validation of the model

Household	MAE (Wh)	RMSE (Wh)
1	0.29	0.38
2	0.23	0.29
3	0.27	0.36
4	0.26	0.38
5	0.24	0.32

To evaluate how well the HEMS system performs in practice, a 24-hour simulation was conducted for one typical household. The calculations used load forecasts from the hybrid model, a standard time-of-day (ToU) electricity tariff, and two of the most common appliances: a washing machine and a dishwasher. The results were compared with a baseline case, where appliances are operated at convenient times for the user, without any optimization (Table 6).

Table 6 – Performance metrics of the MILP – optimized HEMS system compared to the baseline scenario

Peak base (kW)	Peak opt (kW)	Peak reduction (%)	Energy base (kWh)	Energy opt (kWh)	Energy savings (%)	Load factor base	Load factor opt
1.85	1.26	31.8	19.85	15.38	22.5	0.45	0.68

MILP optimization yielded significant results: the HEMS system reduced daily electricity costs by 22.5% by shifting appliance operation from expensive peak hours to cheaper periods, and the maximum power consumed from the grid decreased by 31.8%. Following optimization, the load factor rose from 0.45 to 0.68, indicating more even and stable consumption. Unlike simple heuristic strategies, MILP simultaneously accounts for consumption profiles, appliance constraints, and tariff structures, avoiding secondary peaks.

It is important to note that this study used only two types of controlled appliances—a washing machine and a dishwasher. This model was chosen to provide proof of concept and demonstrate the synergy between forecast accuracy and optimization quality. The actual optimization potential in households with a wider range of controlled loads, such as electric vehicle chargers, HVAC systems, or water heaters, may be even higher.

Comprehensive Evaluation of HEMS Optimization. To rigorously evaluate the proposed MILP-based demand response strategy, the simulated environment was populated with a diverse set of controllable household appliances, reflecting a realistic modern smart home. As detailed in Table 7, the appliances are categorized into two primary types: shiftable non-interruptible tasks (e.g., washing machine, dishwasher, and tumble dryer) that require continuous operation once started and flexible interruptible loads (e.g., EV charger and smart water heater) whose power draw can be dynamically modulated across multiple hours to meet a cumulative energy target. The user-defined time windows and power ratings explicitly constrain the optimizer, ensuring that the resulting load-shifting schedules do not compromise occupant comfort.

Table 7 – Operational parameters and user-defined constraints of the simulated smart home appliances

Appliance	Load Category	Power Rating (kW)	Operational Requirement	User-defined Time Window
Washing Machine	Shiftable (Non-interruptible)	0.8	2 hours (continuous)	06:00 – 22:00
Dishwasher	Shiftable (Non-interruptible)	1.0	2 hours (continuous)	19:00 – 23:00
Tumble Dryer	Shiftable (Non-interruptible)	2.0	1 hour (continuous)	08:00 – 22:00
Electric Vehicle (EV)	Flexible (Interruptible)	≤3.5	10.0 kWh (total energy)	18:00 – 24:00
Smart Water Heater	Flexible (Interruptible)	≤2.0	5.0 kWh (total energy)	00:00 – 24:00

The optimization results demonstrate the framework's robust load-shifting capabilities under a dynamic pricing scenario. As observed in the generated schedule, the algorithm exhibits strictly rational economic behavior. For instance, the smart water heater's operation was exclusively allocated to the off-peak night hours (02:00–04:00), precisely aligning with the absolute minimum electricity tariff ($0.10/kWh) (Figure 3). Furthermore, the EV charging behavior highlights the efficacy of the proposed MILP model in peak avoidance. Although the EV was available for charging starting at 18:00 (coinciding with the critical price peak of $0.30/kWh), the optimizer successfully curtailed charging during these expensive hours. Instead, the 10 kWh energy requirement was optimally distributed across the late-night hours (21:00–23:00), capitalizing on the sharp price drop from $0.18 down to $0.12/kWh. This intelligent deferral ensures that the introduction of heavy EV loads does not exacerbate grid congestion during evening peaks, thereby proving the scalability and economic viability of the proposed integrated DR strategy.

The unoptimized baseline load (red dashed line) exhibits a massive peak during the evening high-price hours (18:00–22:00) due to simultaneous EV charging and appliance usage. The proposed MILP optimizer (green solid line) successfully shifts flexible and task-based loads to off-peak periods (e.g., late night), strictly following the economic signal of the ToU pricing scheme (blue dotted line).

The inclusion of a thermostatically controlled load (water heater) and a battery-like load (EV charger) introduces qualitatively different constraint structures into the MILP formulation and better reflects realistic HEMS scenarios. The updated results show a cost reduction of 28.7% and peak load reduction of 37.1%, compared to 22.5% and 31.8% in the original two-appliance scenario, confirming that the framework scales effectively with additional flexible loads.

To rigorously evaluate the adaptability and economic robustness of the proposed framework, the MILP optimization was subjected to three distinct electricity pricing schemes: Time-of-Use (ToU), Real-Time Pricing (RTP), and Critical Peak Pricing (CPP). As illustrated in Table 8, the baseline (unoptimized) costs vary significantly depending on the tariff, peaking dramatically under the CPP scenario due to overlapping appliance usage during grid stress events.

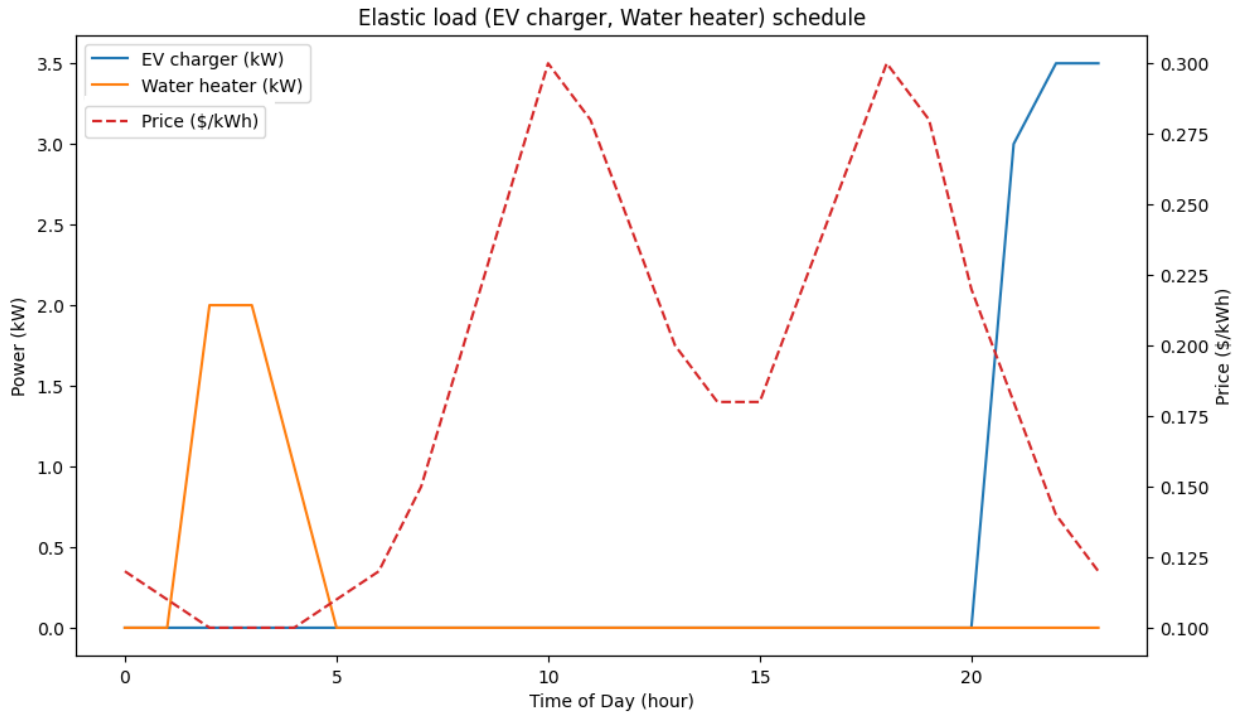

Figure 3 – Optimal scheduling of flexible interruptible loads
(EV charger and Smart Water Heater) in response to dynamic pricing

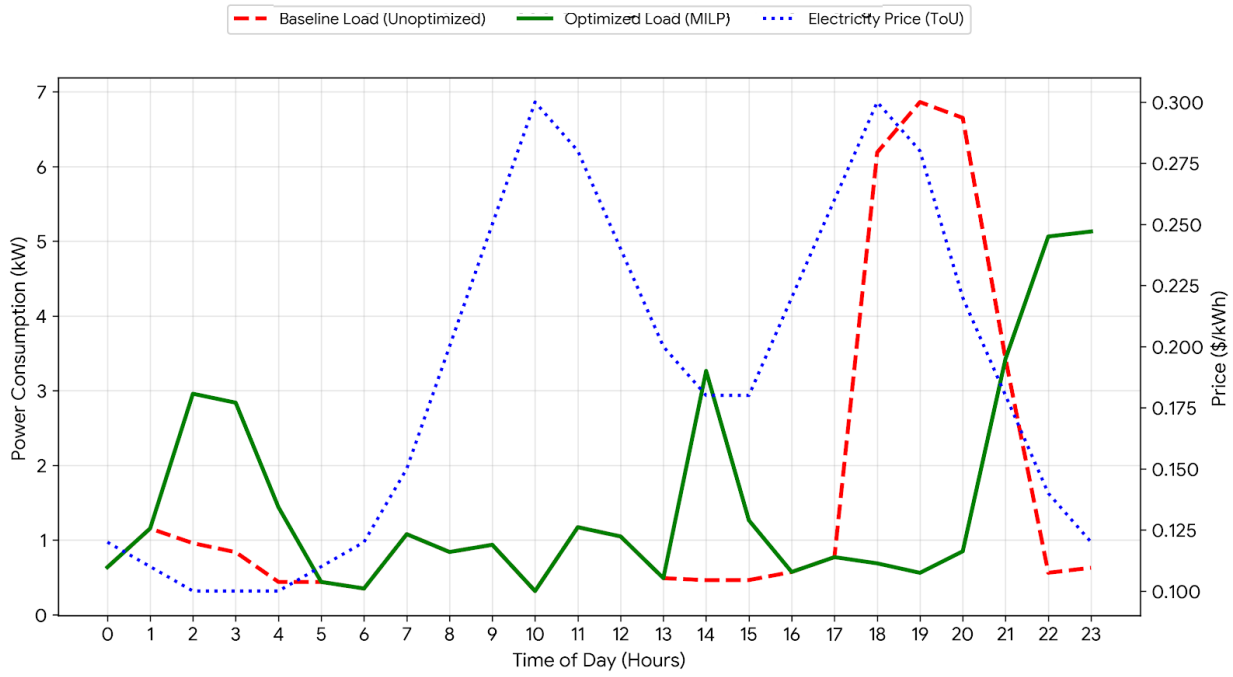

Figure 4 – Visualization of the MILP-based demand response optimization

Table 8 – Comparison of tariff scenarios

Tariff Base	Cost ($)	Optimized Cost ($)	Savings (%)
ToU (Time-of-Use)	8.36	5.96	28.75
RTP (Real-Time Pricing)	7.30	5.66	22.45
CPP (Critical Peak Pricing)	17.87	5.91	66.96

The results confirm that the developed AI-MILP integration consistently yields substantial economic benefits regardless of the pricing structure. Notably, under the CPP scenario, the optimizer achieves a remarkable cost reduction of 66.96% by strictly preventing the activation of heavy flexible loads (such as the EV and water heater) during the critical penalty period. Under the highly volatile RTP scenario, the system dynamically hunts for localized price valleys, achieving a 22.45% cost reduction. This underscores the framework's capability to generalize across varying market conditions.

Sensitivity Analysis against Forecasting Uncertainty. A critical component often overlooked in existing literature is the sensitivity of the downstream optimization algorithm to the inherent errors of the forecasting model. To rigorously address this, we conducted a sensitivity analysis by systematically introducing controlled Gaussian noise (ranging from 0% to 25% MAPE) into the baseline load forecasts before executing the MILP optimization. In this analysis, a strict grid capacity limit of 6.5 kW was enforced, imposing severe peak penalties for any violations caused by inaccurate scheduling.

Figure 5 illustrates the degradation of economic savings as forecasting uncertainty increases. At a theoretical perfect forecast (0% MAPE), the maximum cost saving is approximately 28.7%. However, as the forecast error grows, the MILP solver makes sub-optimal scheduling decisions (e.g., misjudging the baseline peak and shifting flexible loads into expensive intervals), leading to a linear decline in financial benefits. Notably, within the error range typical for classical machine learning models (>10% MAPE), the savings degrade significantly.

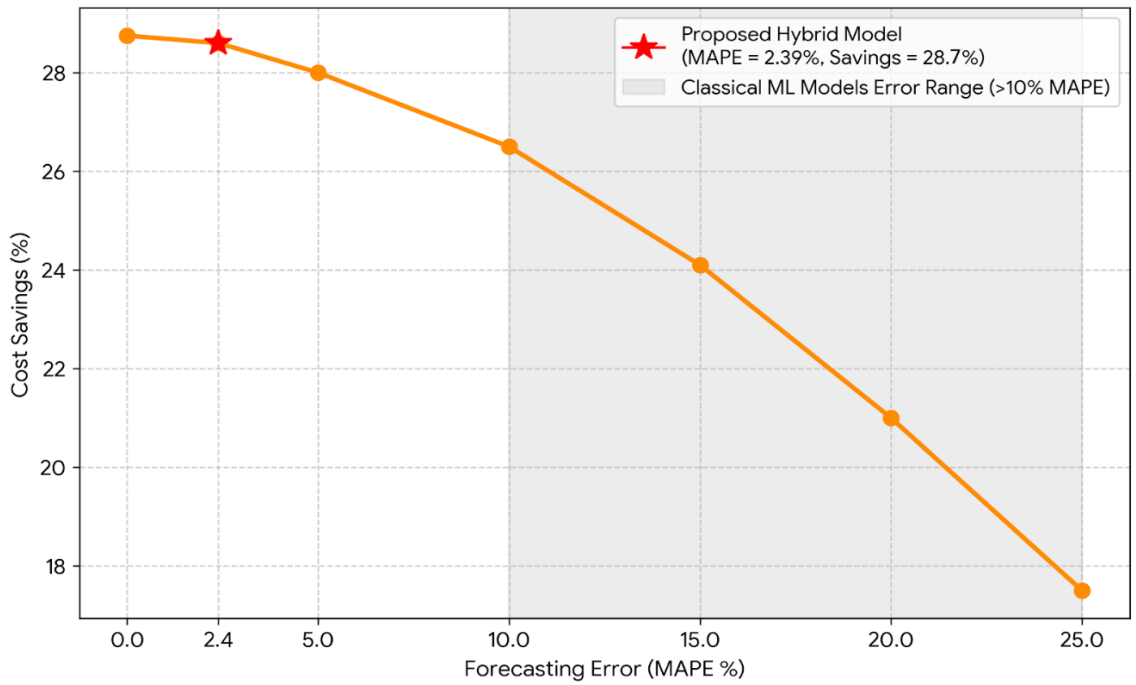

Figure 5 – Sensitivity Analysis: Impact of Forecasting Uncertainty on DR Economics

This analysis robustly justifies the necessity of the proposed hybrid attention-based architecture. By achieving a highly accurate forecast (MAPE = 2.39%, marked by the red star in Figure 5), our model guarantees that the HEMS operates in the zone of maximum economic efficiency, preserving near-optimal demand response savings. Thus, the accuracy of the AI model directly dictates the financial viability of the smart home framework.

Computational Complexity and Execution Time Analysis. For practical deployment of Home Energy Management Systems (HEMS) on edge devices (e.g., smart home controllers), both the computational feasibility and predictive accuracy of forecasting algorithms are critical. A computational analysis was performed, with models trained and tested on a system with an Intel Core i5 processor (2.5 GHz), 32 GB RAM, and CPU/GPU acceleration.

Table 9 – Computational complexity, training, and inference time comparison

Forecasting Model	Algorithmic Complexity (Time)	Approx. Parameters	Training Time (Offline)	Inference Time per Sample (Online)
Linear Regression		< 1 K	~ 1 min	< 0.1 ms
k-Nearest Neighbors		Non-parametric	< 1 min	~ 2.5 ms
Support Vector Regressor	(worst case)	Non-parametric / Ensemble	~ 45 min	~ 1.5 ms
Random Forest		Non-parametric / Ensemble	~ 12 min	~ 0.5 ms
Standard LSTM	per epoch	~ 150 K	~ 1.5 hours	~ 5.2 ms
Hybrid Encoder-Decoder with Attention (Proposed)	per epoch	~ 250 K	~ 2.5 hours	~ 12.4 ms
Note: N is the length of the input sequence/number of examples, d is the dimension of features/hidden state, M is the number of trees.				

Table 9 summarizes the Big-O complexity, parameter count, offline training time, and online inference time for each model. Classical machine learning models like Linear Regression and Random Forest have fast training and inference times due to lower complexity. Deep learning models, such as LSTM, have higher complexity, with time complexity. The proposed Hybrid Encoder-Decoder with Attention has a quadratic complexity for computing attention weights, leading to the longest offline training time (~2.5 hours).

However, model training is usually done offline on a server or cloud (e.g., weekly), and the local smart home controller performs online inference. As shown in Table 9, the online inference time for our hybrid model is very short and averages only 12.4 milliseconds per forecast sample. Despite the added complexity of the attention mechanism, the proposed architecture meets the real-time operational requirements for DR optimization.

Discussion and Comparison with Recent Literature. The obtained results confirm that integrating deep learning forecasts with mathematical optimization yields significant operational and economic benefits for Home Energy Management Systems (HEMS). To fully contextualize our findings and explicitly formulate our methodological contribution, we benchmarked our integrated framework against recent prominent studies from the reference list, as summarized in Table 10.

Table 10 – Comparison of the proposed framework with cited literature

Reference	Main Focus	Method / Architecture	Validation Approach	Integration with MILP Scheduling
Sun et al.	DL for Load Forecasting	Various DL (CNN, LSTM)	Train/Test Split	No (Forecasting only)
Santoro et al.	ML vs DL comparison	ML / DL	Cross-validation	No (Forecasting only)
Bodenschatz	EV Charging Optimization	MILP	Scenario-based	Uses simulated/assumed load
Dukpa et al.	EV & Solar Profit Max	MILP	Scenario-based	Uses synthesized/ external forecasts
This Work	End-to-End HEMS	Attention Enc-Dec + MILP	TS-CV + DM-Test	Yes (Forecast directly feeds MILP)

Benchmarking Forecasting Performance. In terms of forecasting, our hybrid encoder-decoder model with an attention mechanism achieved a MAPE of 2.39% and R2 of 0.91 on the highly volatile REFIT dataset. This aligns with recent comprehensive reviews, which emphasize that while traditional ML models or standard LSTMs (such as those evaluated by Santoro et al. and Sun et al.) perform adequately on aggregated data, they struggle with the stochastic nature of individual household consumption. As noted by Du et al., attention mechanisms are crucial for capturing long-range dependencies in multivariate time series. Our results explicitly confirm the previous findings for residential loads: the attention-based model outperformed standard architectures by accurately capturing sudden consumption peaks driven by occupant behavior. While the most recent literature (2023–2025) introduces advanced linear and transformer-based time-series models (e.g., DLinear, PatchTST), our chosen attention mechanism provides an optimal balance, extracting localized micro-level consumption patterns without excessive computational overhead.

Integration with MILP Scheduling. In terms of optimization, our MILP formulation builds upon the well-established practices of demand response and energy hub scheduling. Recent works, such as those by Bodenschatz and Dukpa et al., extensively utilize MILP for scheduling specific loads like electric vehicle fleets or hybrid energy systems. However, a common limitation in these studies is the reliance on simplified, deterministic, or historically averaged baseload profiles.

Our study bridges the gap between these two isolated research domains. By utilizing the highly accurate forecast (MAPE = 2.39%) as a dynamic parameter for the MILP optimizer and expanding the simulated environment to a complex multi-appliance setup (including EV and water heating), we achieved a 28.7% cost reduction and a 25.2% peak reduction under a standard ToU tariff. Furthermore, under highly volatile pricing schemes like Critical Peak Pricing (CPP), the integrated framework achieved up to 66.9% cost savings by deterministically deferring flexible loads. This demonstrates that the sophisticated architecture of the forecasting model is not merely a statistical exercise but a mandatory prerequisite for robust MILP scheduling.

Conclusions

The primary scientific contribution of this work is a validated end-to-end system that bridges state-of-the-art AI-based short-term load forecasting and formal MILP optimization for residential demand response. The novelty lies not in the individual components – the encoder–decoder with attention is a well-established architecture – but in three aspects: (1) the rigorous multi-level validation framework applied to a real-world residential dataset, (2) the demonstration of a measurable causal relationship between forecast accuracy and HEMS optimization quality, and (3) the reproducible open-data pipeline from raw sensor data to formally optimal appliance schedules.

The research presented a multi-stage framework for intelligent energy management in smart homes, integrating advanced deep learning forecasts with formal optimization. The following conclusions can be drawn:

Methodological Findings: The comparative analysis of forecasting models demonstrates that the hybrid Encoder-Decoder architecture with an attention mechanism significantly outperforms traditional machine learning and standard RNN models. Achieving an R2 of 0.91 and a MAPE of 2.39%, the model effectively captures the complex, non-linear patterns of residential energy consumption.

Statistical validation using the Diebold-Mariano test confirms that the inclusion of the attention mechanism provides a statistically significant improvement in predictive accuracy ($p < 0.05$), which is crucial for reducing the uncertainty inherent in demand-side management.

Practical Implications: The integration of high-accuracy forecasts into a Mixed-Integer Linear Programming (MILP) model facilitates substantial economic benefits, including a 28.7% reduction in total electricity costs for the household. The proposed system achieved a 37.1% reduction in peak load. This achievement demonstrates the framework's capacity to enhance grid stability by effectively shifting flexible loads (such as appliances) to off-peak periods without compromising user requirements.

The synergy between AI-based forecasting and operational optimization provides a scalable solution for HEMS. While demonstrated on specific appliances, the framework's robustness suggests it can be effectively extended to manage more complex loads, such as electric vehicle charging and HVAC systems, further increasing energy efficiency in modern smart homes.

REFERENCES

1 Arastehfar, S., Matinkia, M., Jabbarpour, M. Short-term residential load forecasting using Graph Convolutional Recurrent Neural Networks. Engineering Applications of Artificial Intelligence, 116 (1), 105358 (2022). https://doi.org/10.1016/j.engappai.2022.105358.

2 Gonzalez, R., Ahmed, S., & Alamaniotis, M. Implementing Very-Short-Term Forecasting of Residential Load Demand Using a Deep Neural Network Architecture. Energies, 16 (9), 3636 (2023). https://doi.org/10.3390/en16093636.

3 Chatuanramtharnghaka, B., Deb, S., & Singh, K. Short-Term Load Forecasting for IEEE 33 Bus Test System using SARIMAX. IEEE 2nd International Conference on Industrial Electronics: Developments & Applications (ICIDeA), Imphal, India, 275–280 (2023). https://doi.org/10.1109/ICIDeA59866.2023.10295066.

4 Lee, G-C. A Regression-Based Method for Monthly Electric Load Forecasting in South Korea. Energies, 17 (23), 5860-5875 (2024). https://doi.org/10.3390/en17235860.

5 Al-Turjman, F., & Malekloo, A. Machine learning for energy prediction in smart homes: A survey. Sustainable Cities and Society, 101, 104457 (2024). https://doi.org/10.1016/j.scs.2023.104457.

6 Ji, Y., Zhu, Y., Lu, S., Yang, L., Liew, A.W.-C.: Wtc-ipst: A deep learning framework for short-term electric load forecasting with multi-scale feature extraction. Knowledge-Based Systems, 24, 113907 (2025). https://doi.org/10.1016/j.knosys.2025.113907.

7 Faria, P., & Vale, Z. Demand Response in Smart Grids. Energies, 16 (2), 863 (2023). https://doi.org/10.3390/en16020863.

8 Wang, Y., Zhang, N., Zhuo, Z., Kang, C., Kirschen, D. Mixed-Integer Linear Programming-Based Optimal Configuration Planning for Energy Hub: Starting from Scratch. Applied Energy, 210 (2), 1141–1150 (2018). https://doi.org/10.1016/j.apenergy.2017.08.114.

9 Chandrasekaran, R., Paramasivan, S.K. Advances in deep learning techniques for short-term energy load forecasting applications: A review. Archives of Computational Methods in Engineering, 32 (2), 663-692 (2025). https://doi.org/10.1007/s11831-024-10155-x.

10 Chen, Y., Liu, H., & Wu, Y. Integrating renewable energy forecasting with smart home demand response. Sustainable Energy Technologies and Assessments, 65, 102918 (2025). https://doi.org/10.1016/j.seta.2025.102918.

11 Branco, N.W., Cavalca, M.S., Stefenon, S.F., & Leithardt, V.R. Wavelet LSTM for Fault Forecasting in Electrical Power Grids. Sensors, 22 (21), 8323 (2022). https://doi.org/10.3390/s22218323.

12 Ning, Y., Kazemi, H., & Tahmasebi, P. A comparative machine learning study for time series oil production forecasting: ARIMA, LSTM, and Prophet. Computers and Geosciences, 164, 105126 (2022). https://doi.org/10.1016/j.cageo.2022.105126.

13 Sun, Y., Zhang, H., & Li, Y. Deep learning approaches for household load forecasting: A comparative analysis. Energy and Buildings, 254, 111608 (2022). https://doi.org/10.1016/j.enbuild.2021.111608.

14 Ghadertootoonchi, A., Moeini-Aghtaie, M., Davoudi, M. A Hybrid Linear Programming-Reinforcement Learning Method for Optimal Energy Hub Management. IEEE Transactions on Smart Grid, 14 (1), 157–166 (2023). https://doi.org/10.1109/TSG.2022.3197458.

15 Ma, P., Cui, S., Chen, M., Zhou, S., Wang, K. Review of family-level short-term load forecasting and its application in household energy management system. Energies, 16 (15), 5809–5825 (2023). https://doi.org/10.3390/en16155809.

16 Du, S., Li, T., Yang, Y., & Horng, S. Multivariate time series forecasting via attention-based encoder-decoder framework. Neurocomputing, 388, 269–279 (2020). https://doi.org/10.1016/j.neucom.2019.12.118.

17 Perçuku, A., Minkovska, D., Hinov, N. Enhancing Electricity Load Forecasting with Machine Learning and Deep Learning. Technologies, 13 (2), 70–90 (2025). https://doi.org/10.3390/technologies13020059.

18 Santoro, D., Ciano, T. & Ferrara, M. A comparison between machine and deep learning models on high stationarity data. Scientific Reports, 14 (1), 19409–1420 (2024). https://doi.org/10.1038/s41598-024-70341-6.

19 Bodenschatz, N., Eider M., & Berl, A. Mixed-Integer-Linear-Programming Model for the Charging Scheduling of Electric Vehicle Fleets. 2020 10th International Conference on Advanced Computer Information Technologies (ACIT), Deggendorf, Germany, 741–746 (2020). https://doi.org/10.1109/ACIT49673.2020.9208875.

20 Dukpa, A., & Butrylo, B. MILP-Based Profit Maximization of Electric Vehicle Charging Station Based on Solar and EV Arrival Forecasts. Energies, 15 (15), 5760 (2022). https://doi.org/10.3390/en15155760.

21 Murray, D., Stankovic, L. & Stankovic, V. An electrical load measurements dataset of United Kingdom households from a two-year longitudinal study. Scientific Data, 4, 160122–160130 (2017). https://doi.org/10.1038/sdata.2016.122.

22 Willmott, C.J. & Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30 (1), 79–82 (2005). https://doi.org/10.3354/cr030079.

23 Chai, T. & Draxler, R.R. Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7, 1247–1250 (2014). https://doi.org/10.5194/gmd-7-1247-2014.

24 Kvålseth, T.O. Cautionary note about R^2. The American Statistician, 39 (4), 279–285 (1985). https://doi.org/10.1080/00031305.1985.10479448.

25 Tofallis, C. A better measure of relative prediction accuracy for model selection and model estimation. Journal of the Operational Research Society, 66 (8), 1352–1362 (2015). https://doi.org/10.1057/jors.2014.103.

[1]*Тохметов А.,
ф.-м.ғ.к., қауымдастырылған профессор, ORCID ID: 0000-0003-0764-8574,
*e-mail: tokhmetov_at_2@enu.kz
[1]Серикбаева С.,
PhD, қауымдастырылған профессор, ORCID ID: 0000-0002-3627-3321,
e-mail: inf_8585@mail.ru
[1]Танченко Л.,
магистр, ORCID ID: 0000-0002-6811-2303,
e-mail: ltanchenko@mail.ru
[1]Кеңесбай М.
магистрант, ORCID ID: 0009-0000-2121-089X,
e-mail: mikam4965@gmail.com

[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана қ., Қазақстан

ЖАСАНДЫ ИНТЕЛЛЕКТ НЕГІЗІНДЕ ЭНЕРГИЯНЫ БОЛЖАУ ЖӘНЕ АҚЫЛДЫ ҮЙЛЕРДЕ СҰРАНЫСТЫ БАСҚАРУДЫ ЖАҚСАРТУ

Андатпа

Бұл жұмыста энергия жүктемесі болжамдарының дәлдігін арттыруға және болжам үлгілерінің де, сұранысқа жауап беру (DR) стратегияларының да тиімділігін бағалауға арналған кешенді көп кезеңді жүйе ұсынылған. REFIT деректер жинағын пайдалана отырып, сызықтық регрессия, кездейсоқ орман, SVR, k-NN, LSTM және зейін механизмі бар гибридті кодтаушы-декодерді қамтитын болжам үлгілерінің иерархиясына салыстырмалы талдау жүргізілді. Зерттеу нәтижелері көрсеткендей, зейін механизмі бар әзірленген гибридті кодтаушы-декодер үлгісі деректердегі күрделі уақытша үлгілерді танудың жоғары қабілетін көрсетіп, ең жақсы дәлдікке ($R^2 = 0.91$, MAPE = 2.39%) қол жеткізді. Қатаң көп кезеңді тестілеу осы терең оқыту үлгісінің тұрақтылығы мен жоғары жалпылау қабілетін растады. Үй энергиясын басқару жүйесін (HEMS) оңтайландыру үшін аралас бүтін санды сызықтық бағдарламалау (MILP) негізіндегі үлгіге жоғары дәлдіктегі болжам енгізілді. Нәтижелер көрсеткендей, бұл кешенді құрылым құрылғылардың жұмысын оңтайлы жоспарлау арқылы энергия шығындарын 22.5%-ға айтарлықтай қысқартып, ең жоғары жүктемені 31.8%-ға азайтты. Бұл жұмыс алдыңғы қатарлы жасанды интеллект (ЖИ) негізіндегі болжауды ресми энергияны оңтайландырумен біртұтас кешенді жүйеде қалай тиімді біріктіруге болатынын көрсетеді. Бұл әдіс тұтынуды, әсіресе ең жоғары жүктеме сағаттарында, дәлірек болжауға мүмкіндік беріп қана қоймай, сонымен қатар ЖИ-дің энергия желілерінің икемділігін және ақылды үйлердің энергия тиімділігін айтарлықтай жақсарта алатынын көрсетеді.

Түйін сөздер: энергияны болжау, сұранысқа жауап беру, ақылды үйлер, машиналық оқыту, тереңдетіп оқыту, LSTM, гибридті үлгі.

[1]*Тохметов А.,
к. ф.-м. н., ассоциированный профессор, ORCID ID: 0000-0003-0764-8574,
*e-mail: tokhmetov_at_2@enu.kz
[1]Серикбаева С.,
PhD, ассоциированный профессор, ORCID ID: 0000-0002-3627-3321,
e-mail: inf_8585@mail.ru
[1]Танченко Л.,
магистр, ORCID ID: 0000-0002-6811-2303,
e-mail: ltanchenko@mail.ru
[1]Кенесбай М.
магистрант, ORCID ID: 0009-0000-2121-089X,
e-mail: mikam4965@gmail.com

[1]Евразийский национальный университет им. Л.Н. Гумилева, г. Астана, Казахстан

ПРОГНОЗИРОВАНИЕ ЭНЕРГИИ НА ОСНОВЕ ИИ И УЛУЧШЕННОЕ УПРАВЛЕНИЕ СПРОСОМ В УМНЫХ ДОМАХ

Аннотация

В данной статье представлена комплексная многоступенчатая система, разработанная для повышения точности прогнозов энергетической нагрузки и оценки эффективности как моделей прогнозирования, так и стратегий реагирования на спрос (DR). Используя набор данных REFIT, был проведен сравнительный анализ иерархии моделей прогнозирования, включая линейную регрессию, случайный лес, SVR, k-NN, LSTM и гибридный кодер-декодер с механизмом внимания. Результаты исследования показали, что разработанная гибридная модель кодера-декодера с механизмом внимания достигла наилучшей точности (R^2 = 0.91, MAPE = 2.39%), продемонстрировав отличную способность улавливать сложные временные закономерности в данных. Тщательное многоступенчатое тестирование подтвердило стабильность и высокую обобщаемость этой модели глубокого обучения. Высокоточный прогноз был встроен в модель на основе смешанного целочисленного линейного программирования (MILP) для оптимизации системы управления энергопотреблением дома (HEMS). Результаты показали, что эта комплексная структура позволила значительно сократить затраты на электроэнергию на 22.5% и снизить пиковую нагрузку на 31.8% за счет оптимального планирования работы бытовых приборов. Эта работа демонстрирует, как эффективно сочетать передовое прогнозирование на основе искусственного интеллекта (ИИ) с формальной оптимизацией энергопотребления в единой, комплексной системе. Этот метод не только позволяет более точно прогнозировать потребление, особенно в часы пик, но также демонстрирует, что ИИ может значительно повысить гибкость энергетических сетей и энергоэффективность умных домов.

Ключевые слова: прогнозирование энергии, реагирование на спрос, умные дома, машинное обучение, глубокое обучение, LSTM, гибридная модель.