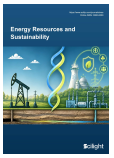
Scilight

*Energy Resources and Sustainability*
https://www.sciltp.com/journals/eras



*Article*

# Modelling and Forecasting of Solar Energy Generation and Consumption in Residential Buildings Using Machine Learning

Behnoosh Shahidpoorfalah, Shuo Li and Wei Peng *

Industrial Systems Engineering, Faculty of Engineering and Applied Science, University of Regina, Regina, SK S4S 0A2, Canada
* Correspondence: Wei.Peng@uregina.ca

**How To Cite:** Shahidpoorfalah, B.; Li, S.; Peng, W. Modelling and Forecasting of Solar Energy Generation and Consumption in Residential Buildings Using Machine Learning. *Energy Resources and Sustainability* **2026**, *1*(1), 4.

Received: 11 August 2026
Revised: 7 September 2026
Accepted: 7 September 2026
Published: 6 November 2026

**Abstract:** Residential buildings account for a substantial share of global electricity use, and their reliance on solar energy is constrained by the intermittency of irradiance and the variability of heating and cooling demand, particularly in cold climates such as Regina, Saskatchewan. This paper presents a modular, three-stage, fully data-driven machine learning framework for intelligent solar-assisted energy control in residential buildings. In Stage 1, an ensemble of XGBoost and LightGBM gradient-boosting models forecasts Global Horizontal Irradiance (GHI) from historical weather data and engineered solar-geometry features, achieving a coefficient of determination ($R^2$) of 0.9975. In Stage 2, a hybrid deep learning architecture that combines a one-dimensional convolutional neural network (CNN), two stacked Long Short-Term Memory (LSTM) layers, and trainable building-identity embeddings predicts hourly heating and cooling energy consumption for twelve residential buildings, reaching $R^2 = 0.9730$ for heating and $R^2 = 0.9140$ for cooling. In Stage 3, a Random Forest classifier trained on rule-derived labels assigns each hour to one of three operational modes (SOLAR, GRID, or OFF) with classification accuracy exceeding 99% across heating and cooling scenarios. The framework is evaluated using observational 2018 weather and GHI data from Regina, Saskatchewan, together with hourly residential building-energy data from twelve buildings in the NREL ResStock AMY2018 dataset located in North Dakota, a climatically similar region, and its performance is benchmarked against multiple peer-reviewed forecasting and control studies. The results show that combining interpretable ensemble learning, sequence-aware deep learning, and rule-informed classification yields a scalable and transparent decision-support pipeline for solar energy management that can be progressively extended toward real-time smart-building deployment.

**Keywords:** solar-assisted energy systems; machine learning; GHI forecasting; CNN-LSTM; residential building energy management; ensemble learning

## 1. Introduction

Residential buildings are responsible for nearly 40% of global electricity consumption and a correspondingly large share of greenhouse gas emissions [1]. In cold climates such as Regina, Saskatchewan, this burden is amplified by high seasonal heating demand and by weather-driven fluctuations in available energy supply. During the warmer months, cooling loads further increase consumption, raising the stakes for effective thermal load management. Solar-assisted energy systems have emerged as a leading strategy for reducing dependence on fossil



**Copyright:** © 2026 by the authors. This is an open access article under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).
**Publisher's Note**: Scilight stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

fuels and supporting low-carbon residential infrastructure, yet their effective use is limited by the intrinsic variability and discontinuity of solar irradiance.

Accurate, short-horizon forecasting of solar irradiance is therefore foundational to any solar-aware control strategy, especially when real-time irradiance sensing is unavailable and hourly radiation must be inferred from routinely measured meteorological variables. Ensemble gradient-boosting algorithms such as XGBoost and LightGBM have been shown to outperform classical physical and statistical models for this task because they capture nonlinear dependencies in meteorological data while remaining robust to noise and multicollinearity [2]. Beyond irradiance forecasting, accurately modeling how a building actually consumes energy is equally important for operational planning, since occupancy, thermal characteristics, and weather interact nonlinearly to shape demand. Hybrid deep learning architectures that combine convolutional and recurrent layers have proven particularly effective at capturing both the short-term, localized fluctuations and the longer temporal dependencies present in energy consumption time series [3].

Translating these two predictive layers into an actionable control decision requires a third component: a classifier capable of deciding, on an hourly basis, whether a building should draw on solar generation, fall back to the grid, or remain idle. Purely rule-based systems derived from solar-engineering thresholds [4] are interpretable but inflexible at the margins of uncertain or borderline conditions, whereas tree-based ensemble classifiers such as Random Forest combine the interpretability of threshold logic with the adaptability of data-driven learning, and have demonstrated strong performance in smart microgrid control settings [5]. The growing literature on machine-learning-enabled building energy management systems (BEMS) further confirms the value of distributed, learning-based infrastructures for load forecasting and intelligent control [6], while recent work on system resilience emphasizes that such systems must remain robust to sensor failure, abrupt demand changes, and incomplete data [7].

Despite this progress, two structural gaps remain in the literature. First, many studies treat irradiance forecasting and load control as independent problems, addressed by separate model families with no shared pipeline connecting forecasting to decision-making [8,9]. Second, machine learning models trained for one building stock or climate often transfer poorly to other regions because of differences in construction, sensor architecture, and occupant behavior [10]. This paper addresses both gaps by proposing an integrated, three-stage modular framework that couples ensemble-based irradiance forecasting, deep-learning-based energy consumption modeling, and rule-informed classification into a single coherent pipeline, and by validating the framework using real weather data from Regina together with energy-consumption data from a climatically similar building stock in North Dakota.

The three stages are intentionally linked as a sequential prediction-to-decision architecture rather than treated as independent modeling tasks. Stage 1 estimates the solar resource available at a given hour from meteorological and solar-geometry information. The resulting GHI forecast is transferred to Stage 2 as an input, where it is combined with weather, temporal, and building-identity information to estimate heating or cooling demand. Stage 3 then uses the forecasted solar availability and forecasted building demand to classify the appropriate operational mode as SOLAR, GRID, or OFF. Thus, the framework follows a logical progression from resource forecasting, to demand forecasting, to operational decision support.

The contribution of this work is not simply the combination of three established machine-learning algorithms, but their integration into a sequential forecast-to-decision framework in which forecasted solar availability directly informs building-demand forecasting and both upstream forecasts are subsequently propagated into operational-mode classification. The framework further incorporates trainable building-identity embeddings to represent latent differences among multiple residential buildings within a shared energy-forecasting architecture, while separate heating and cooling models account for distinct seasonal load dynamics. Finally, the operational classifier is evaluated using both measured and forecasted upstream inputs, enabling explicit assessment of how forecasting errors influence downstream operational decisions. These characteristics distinguish the proposed framework from studies that evaluate irradiance forecasting, building-load forecasting, or control classification independently.

The remainder of this paper is organized as follows. Section 2 reviews related work in machine-learning-based solar forecasting, building energy modeling, and learning-based control. Section 3 describes the proposed three-stage methodology, including data sources, feature engineering, and model architectures. Section 4 presents the experimental results for each stage and discusses the findings in relation to comparable studies. Section 5 concludes the paper, contribution, limitation, and outlines directions for future work.

## 2. Related Work

### *2.1. Machine Learning in Solar Energy Systems*

Machine learning has fundamentally reshaped how building energy systems are modeled, forecast, and controlled. Aguilar et al. [1] emphasize the importance of learning-based building energy management systems, particularly when integrated with sensor networks and Internet-of-Things platforms that enable distributed, real-time adaptability. Nabavi et al. [6] show that deep learning models such as LSTM can identify complex spatio-temporal patterns in smart energy systems, while Kurukuru et al. [11] categorize artificial intelligence applications in solar photovoltaic systems into forecasting, fault detection, dispatch optimization, and performance monitoring. Tina et al. [12] support hybrid deep learning methods that incorporate time-based encodings and solar-geometry features, and Pang et al. [13] show that recurrent architectures are more effective than simple artificial neural networks at capturing short-term irradiance variability, observations that motivate the time-aware feature engineering adopted in this study.

### *2.2. Solar Irradiance Forecasting*

Physically based irradiance models generally struggle with the discontinuous, weather-driven behavior of solar radiation, motivating a shift toward data-driven alternatives. Kumari and Toshniwal [2] compare LSTM, GRU, and CNN-LSTM architectures across multiple climates and forecast horizons and find recurrent architectures effective for short-term forecasting, while Rajagukguk et al. [14] demonstrate that domain features such as solar elevation and time-of-day encoding materially improve deep learning accuracy and generalizability. Houchati et al. [8] propose a rooftop solar forecasting framework based on historical temporal similarity, and Wang et al. [15] show in a broad survey that ensemble models such as XGBoost and LightGBM often outperform deep architectures for short-term forecasting when paired with log-transformed targets and appropriate feature scaling, an observation directly mirrored by the ensemble design adopted in Stage 1 of this work.

### *2.3. Deep Learning for Building Energy Consumption Forecasting*

Hybrid CNN-LSTM architectures have demonstrated strong performance for residential load forecasting. Kim and Cho [16] propose a CNN-LSTM model for short-term household consumption forecasting that achieves an $R^2$ of 0.918 using weather and occupancy variables, but their architecture is tailored to a single residence and lacks a mechanism for generalizing across buildings. Zhou et al. [3] apply an LSTM-based framework to real-time electric-vehicle charging scheduling, conceptually paralleling the temporally aware forecasting required for residential heating and cooling demand. Le et al. [17] and Sajjad et al. [18] further refine hybrid CNN-recurrent designs using bidirectional LSTM and GRU units, respectively, while Drir et al. [19] and Salama and Mahmoud [20] investigate CNN- and ANN-based architectures for home energy scheduling and load prioritization. None of these studies, however, incorporate a learned building-identity representation capable of capturing inter-building heterogeneity from a shared model, a gap that Stage 2 of the proposed framework directly addresses through trainable building embeddings, following the consumption-pattern profiling methodology of Chou and Tran [21].

### *2.4. Hybrid and Learning-Based Control Architectures*

Several studies combine forecasting and control into unified, often reinforcement-learning-based, architectures. Paul and Padhy [22] integrate short-term machine-learning forecasts with real-time bilevel decision-making for smart apartment buildings, while Zhang et al. [23] use deep reinforcement learning to optimize energy distribution in hybrid grids without explicit physical modeling. Wei et al. [24] develop a proximal-policy-optimization-based deep reinforcement learning controller that coordinates smart-home devices, storage, and photovoltaic input through a combined discrete-continuous action space, and Xu et al. [25] apply model-based reinforcement learning to optimize photovoltaic-battery operation. Salari et al. [5] combine fuzzy logic with Q-learning in a cooperative multi-agent home-microgrid controller, and Xu et al. [26], Lingmin et al. [27], and Lee and Choi [28] respectively propose multi-agent rule-based coordination, Q-learning for combined cooling-heating-power systems, and hierarchical deep reinforcement learning for smart-home appliance and storage control. These architectures achieve high adaptability but typically require extensive training data, reward-function design, and computational resources that limit practical deployment. The Random Forest classification approach adopted in Stage 3 of this study is conceptually inspired by these control frameworks while avoiding the implementation overhead of reward-based learning cycles.

### *2.5. Uncertainty, Resilience, and Transferability*

Robustness to uncertainty and data scarcity is a recurring theme in the literature. Mansour et al. [9] use Wasserstein generative adversarial networks to model uncertainty in solar generation, and Zeynalia et al. [29] apply a two-stage stochastic ANN-based framework to manage uncertainty from electric-vehicle loads and price fluctuations. Wang et al. [10] identify generalizability as a key limitation of machine learning models trained on a single building stock, and Mehrjerdi et al. [7] argue that resilient building energy management systems must perform acceptably under sensor failure or data-exchange delay by relying on historical trends or proxy data sources. Li et al. [30] similarly emphasize that transferring photovoltaic-battery configurations across regions requires careful climate-aware calibration, supporting the use of proxy datasets from climatologically similar regions, such as the North Dakota building data used in this study to model Regina's residential energy behavior.

## 3. Proposed Methodology

### *3.1. Overview of the Three-Stage Framework*

The proposed framework follows a fully data-driven approach: rather than relying on physics-based simulation or detailed hardware specifications, each stage learns from observational weather and irradiance data together with hourly residential energy-consumption data from the NREL ResStock AMY2018 dataset. The pipeline is organized into three sequential, modular stages, illustrated in Figure 1. Stage 1 forecasts Global Horizontal Irradiance (GHI) from historical weather variables and engineered solar-geometry features. Stage 2 uses the Stage 1 irradiance forecast, together with weather inputs, to predict hourly heating and cooling energy consumption for individual residential buildings. Stage 3 combines the predicted irradiance and predicted energy consumption with rule-derived thresholds to classify each hour into one of three operational modes (SOLAR, GRID, or OFF) using a Random Forest classifier.

Modularity was an explicit design goal: each stage can be retrained, replaced, or extended independently, and each relies on algorithms (XGBoost, LightGBM, and Random Forest) that natively expose feature-importance scores. This provides a degree of interpretability that allows domain experts to verify that model behavior aligns with known engineering relationships, rather than treating the pipeline as an opaque black box.

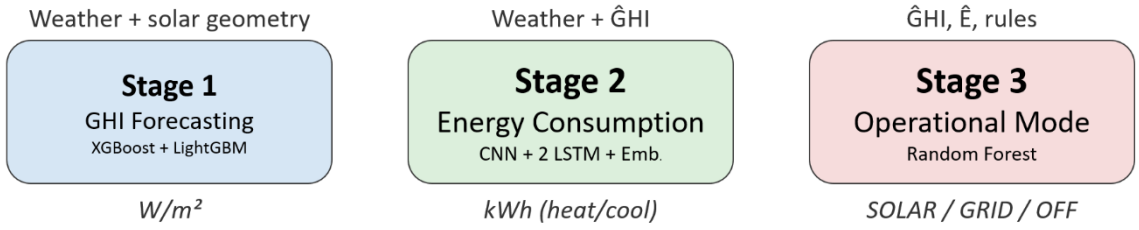



**Figure 1.** Overview of the proposed three-stage machine learning framework for solar-assisted residential energy control.

### *3.2. Stage 1: GHI Forecasting Using Ensemble XGBoost and LightGBM*

Global Horizontal Irradiance is the total shortwave solar radiation (direct and diffuse) received on a horizontal surface, and it is the primary indicator of solar energy potential at a given location [4]. GHI forecasting was framed as a supervised regression task using sixteen engineered meteorological and temporal features: air temperature, dew point, relative humidity, wind speed, station pressure, visibility, solar elevation angle, sine/cosine encodings of hour-of-day, a one-hour lag of GHI, a three-hour rolling mean of GHI, and a clearness index computed as the ratio of the t-1 GHI to the maximum GHI observed at the same hour across the year.

Solar elevation was computed using the pvlib Python library from timestamp and geolocation, capturing the sun's apparent position in the sky and helping the model resolve diurnal irradiance fluctuations. To prevent artificial discontinuities at the day boundary (e.g., between 23:00 and 00:00), the hour of day was encoded cyclically using:

$$\sin_{hour} = \sin\left(\frac{2\pi \cdot hour}{24}\right) \tag{1}$$

$$\cos_{hour} = \cos\left(\frac{2\pi \cdot hour}{24}\right) \tag{2}$$

Because GHI is strongly right-skewed, the target was log-transformed prior to training,

$$y = \log(GHI + 1) \tag{3}$$

and forecasting was converted back to physical units (W/m$^2$) after inference using the inverse transform

$$GHI_{pred} = e^{y} - 1 \tag{4}$$

All numerical features were normalized to [0, 1] using a MinMaxScaler. A Pearson correlation analysis (Figure 2) confirmed that solar elevation and the cyclic hour encodings were among the strongest predictors of log-transformed GHI; although moderate multicollinearity was observed between temperature and dew point (r = 0.92), no features were removed, since both XGBoost and LightGBM are inherently robust to redundant inputs through their regularization mechanisms.

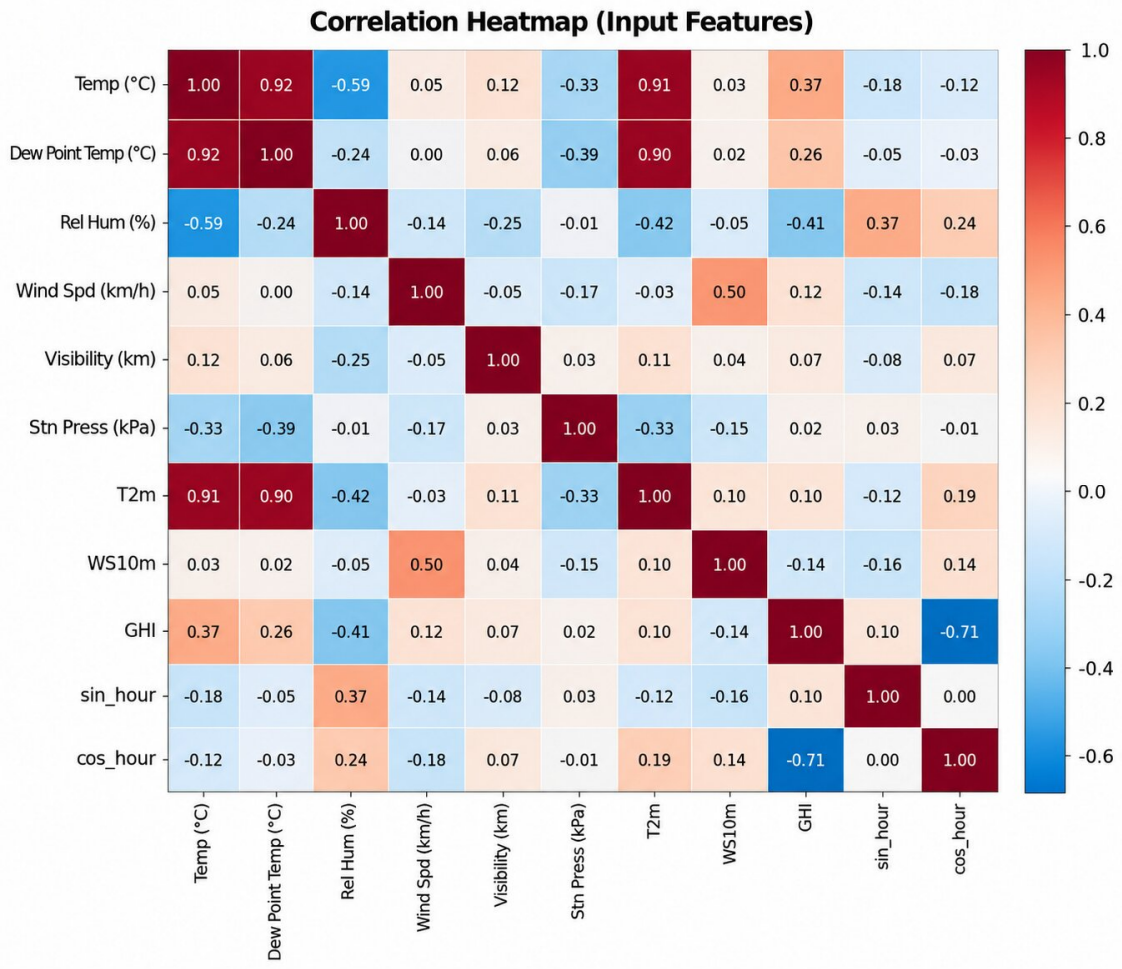



**Figure 2.** Pearson correlation heatmap between meteorological variables, engineered temporal features, and GHI.

XGBoost (Extreme Gradient Boosting) is a scalable, regularized tree-boosting framework that builds decision trees sequentially, with each new tree fit to the residual errors of the previous iteration [31]. It applies both L1 and L2 regularization to promoting feature sparsity through shrinkage. The XGBoost component was trained with 300 estimators, a maximum tree depth of 9, a learning rate of 0.03, subsampling and column-subsampling ratios of 0.9, and a gamma of 0.1 to regularize tree growth.

LightGBM is a histogram-based gradient boosting framework that improves training speed and scalability through Gradient-based One-Side Sampling (GOSS), which retains high-gradient instances while subsampling low-gradient ones, and Exclusive Feature Bundling (EFB), which merges mutually exclusive sparse features to reduce dimensionality [32]. The LightGBM component was trained with the same hyperparameters as XGBoost (300 estimators, maximum depth 9, learning rate 0.03, subsampling ratio 0.9) to ensure comparable model capacity between the two ensemble members.

Preliminary experiments with a univariate LSTM model for GHI forecasting showed poor convergence given the limited temporal history available in hourly irradiance data, motivating the use of a simple unweighted averaging ensemble of the two boosting models,

$$\hat{y}_{emsemble} = 0.5\,\hat{y}_{XGB} + 0.5\,\hat{y}_{LGB} \tag{5}$$

where $\hat{y}_{XGB}$ and $\hat{y}_{LGB}$ denote the individual model forecasts in the log-transformed domain, subsequently inverted using (4).

### *3.3. Stage 2: Energy Consumption Forecasting Using CNN-LSTM with Building Embeddings*

Hourly heating and cooling energy consumption was modeled using a hybrid architecture composed of a one-dimensional convolutional neural network (1D-CNN), two stacked LSTM layers, and a trainable building-identity embedding layer, with separate models trained for the heating and cooling seasons to reflect their distinct load dynamics.

The 1D-CNN layer extracts short-term local temporal features from the multivariate input sequence: temperature, humidity, predicted GHI, and related variables, acting as a feature preprocessor before the recurrent layers. The CNN output is passed through two stacked LSTM layers, which mitigate the vanishing-gradient limitations of standard recurrent networks through gated memory cells [33]. Each cell maintains a memory state updated through forget, input, and output gates:

$$f_t = \sigma\big(W_f \cdot [h_{t-1}, x_t] + b_f\big), \qquad i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{6}$$

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)\,, \qquad c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t \tag{7}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \qquad h_t = o_t \cdot \tanh(c_t) \tag{8}$$

where $x_t$ is the input vector at time $t$ (normalized weather, GHI, and time features), $h_{(t-1)}$ is the previous hidden state, and $c_t$ is the memory cell. The first LSTM layer returns the full output sequence to preserve temporal structure, while the second layer encodes higher-level abstracted patterns. This stacked configuration has been shown to improve forecasting accuracy on non-stationary, time-dependent energy datasets [34].

To capture inter-building heterogeneity without relying on external metadata, each of the twelve buildings was mapped to a trainable embedding vector of dimension four, following the consumption-pattern profiling approach of Chou and Tran [21]. This dimension was chosen as a parsimonious trade-off: large enough to capture latent differences in baseline demand, temperature sensitivity, and diurnal consumption patterns, yet small enough to avoid overfitting given a one-year dataset. The embeddings were concatenated with the CNN-LSTM output and passed through a dense layer prior to the final regression output. All numerical inputs were normalized to [0, 1] using MinMaxScaler, and a 24-h sliding window was used to construct each training example, so that each forcast is conditioned on the preceding 24-hourly observations, enabling the model to learn full diurnal consumption cycles. The model was trained for 120 epochs with the Adam optimizer and a mean-squared-error loss, using an 80:20 train-test split, the train-test partition was chronological. no explicit dropout, L2 penalty, or early stopping was required, as training loss converged stably without overfitting, a behavior attributed to the compact embedding dimension and fixed 24-h window constraining model capacity. The 24-h input windows were constructed using observations preceding the forecast timestamp, and no observations from the test period were used to construct training inputs or fit preprocessing transformations.

### *3.4. Stage 3: Operational Mode Classification Using Random Forest*

The final stage translates the Stage 1 and Stage 2 forecasts into an operational decision: whether a building should draw on solar generation (SOLAR), switch to grid electricity (GRID), or remain inactive (OFF) during a given hour. This stage does not implement a physical control system; rather, it predicts the decision-mode label that a rule-based control policy would assign, using a supervised Random Forest classifier trained on both rule-derived labels and a richer feature set than the rules alone, allowing the model to generalize past sharp threshold boundaries.

Training labels were generated using a four-condition rule-based logic combining a GHI threshold of 250 W/m$^2$ identified in solar engineering literature as the minimum irradiance for stable inverter operation [4] and an energy-consumption threshold derived from exploratory analysis (0.25 kWh for cooling, 0.5 kWh for heating, reflecting the generally higher sustained thermal output of heating systems). The four threshold combinations were mapped onto three operational classes as seen in Eq 9: SOLAR when GHI and demand are both above threshold; GRID when demand is high, but GHI is low, and OFF when either GHI is high with low demand, or both values are below threshold.

$$\text{Mode}_t = \begin{cases} \textbf{SOLAR}, & \text{if } GHI_t \geq 250 \text{ W/m}^2 \text{ and } E_t \geq E_{\text{thresh}} \\ \textbf{GRID}, & \text{if } GHI_t < 250 \text{ W/m}^2 \text{ and } E_t \geq E_{\text{thresh}} \\ \textbf{OFF}, & \text{if } E_t < E_{\text{thresh}} \text{ (regardless of } GHI_t) \end{cases} \quad (9)$$

The 0.25 kWh cooling and 0.5 kWh heating thresholds were selected empirically from the observed hourly consumption distributions in the study building stock, with the heating threshold corresponding approximately to the transition between low/standby demand and sustained heating demand. Because these thresholds were derived from the North Dakota ResStock buildings used in this study, they should be regarded as dataset-specific rather than universal building-energy thresholds.

Four datasets were prepared for the heating and cooling seasons, each using either actual or model-predicted GHI and energy values so that classifier performance could be compared under both ideal (measured) and realistic (forecasted) input conditions. Input features included GHI (actual or predicted), temperature, dew point, relative humidity, wind speed, visibility, station pressure, solar elevation, sine/cosine hour encodings, and energy consumption (actual or predicted).

To train and evaluate the classifier, four distinct datasets were prepared corresponding to the following configurations:

(1) Cooling buildings using actual GHI and actual energy values
(2) Cooling buildings using predicted GHI and predicted energy values
(3) Heating buildings using actual GHI and actual energy values
(4) Heating buildings using predicted GHI and predicted energy values

Each dataset was split 80/20 using stratified sampling to preserve class balance, and the Random Forest classifier was configured with class_weight = 'balanced' to address label imbalance, particularly the relative scarcity of OFF labels in heating data. The number of trees was empirically tuned to 150 for cooling models and 100 for heating models, beyond which no further performance gains were observed. Feature importance was extracted using the Gini importance criterion to identify the dominant drivers of each classification decision.

### *3.5. Evaluation Metrics*

Regression performance (Stages 1 and 2) was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination ($R^2$), together with Mean Absolute Percentage Error (MAPE) for Stage 1. MAE summarizes average error magnitude; RMSE penalizes large deviations more heavily and is therefore sensitive to forecast spikes; and $R^2$ quantifies the proportion of variance explained by the model. Classification performance (Stage 3) was evaluated using precision, recall, F1-score, and overall accuracy,

$$\text{Accuracy} = (\text{TP}+\text{TN})/(\text{TP}+\text{TN}+\text{FP}+\text{FN}) \quad (10)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively, and confusion matrices were used to visualize class-wise behavior.

## 4. Results & Discussion

### *4.1. Data Sources and Study Area*

All experiments use real, measured data from the year 2018, the only year for which complete weather and building-energy records were simultaneously available for both locations used in this study. Hourly meteorological data for Regina, Saskatchewan (latitude 50.4452° N, longitude 104.6189° W) were obtained from Environment and Climate Change Canada's historical archive and include air temperature, dew-point temperature, relative humidity, wind speed, station pressure, and visibility. Corresponding GHI measurements, at 15-min resolution, were obtained from NASA POWER and resampled to hourly means; only daytime records (GHI > 0) were retained for Stage 1 model training, since nighttime hours contribute no informative irradiance signal.

Building-level energy-consumption data for Stage 2 were drawn from the National Renewable Energy Laboratory's (NREL) ResStock AMY2018 release for twelve residential buildings located in North Dakota. This region was selected because its climate closely resembles Regina's and because the dataset provides open, hourly records of natural-gas heating and electricity cooling consumption together with building identifiers. The year was divided into a cooling season (May–September, modeled using electricity consumption for cooling) and a heating season (October–April, modeled using natural-gas consumption for heating), reflecting the markedly different load dynamics of each season. All data preprocessing, feature engineering, and model training were performed in

Python using Pandas, NumPy, scikit-learn, TensorFlow/Keras, and pvlib, with Kaggle notebooks used as the primary training environment.

### *4.2. Stage 1: GHI Forecasting Results*

The Stage 1 dataset comprised 4,402 daytime hourly records for 2018. Table 1 Compares the XGBoost model with the final XGBoost+LightGBM ensemble. The ensemble achieved an $R^2$ of 0.9975, MAE of 7.47 W/m$^2$, RMSE of 11.80 W/m$^2$, and MAPE of 4.12%, substantially outperforming the standalone XGBoost model ($R^2$ = 0.9777, MAE = 21.52 W/m$^2$). We see XGBoost by itself is already able to achieve a very high $R^2$ value, but with Ensemble LightGBM, we are able to further increase $R^2$ and reduce MAE by about 3 times.

**Table 1.** Evaluation metrics for GHI forecasting models.

| Model | $R^2$ | MAE (W/m$^2$) | RMSE (W/m$^2$) | MAPE (%) |
|---|---|---|---|---|
| XGBoost | 0.9777 | 21.52 | 35.35 | 18.95 |
| XGBoost + LightGBM (Ensemble) | 0.9975 | 7.47 | 11.80 | 4.12 |

Figure 3 displays a scatter plot of actual versus predicted GHI values. We see that the variation of predicted points is small, and the points stay close to the ideal Predicted vs. Actual GHI line. The linear relationship confirms the model's ability to maintain accuracy across the full range of irradiance values.

XGBoost is trained with 300 estimators, a maximum tree depth of 9, a learning rate of 0.03, and subsampling and column-subsampling ratios of 0.9. In addition, a gamma parameter of 0.1 was used to regularize tree growth and penalize overly complex splits.

The LightGBM model was trained using the same hyperparameter settings as XGBoost (300 estimators, maximum depth of 9, learning rate of 0.03, and subsampling ratios of 0.9), ensuring consistent model capacity across the two ensemble components.

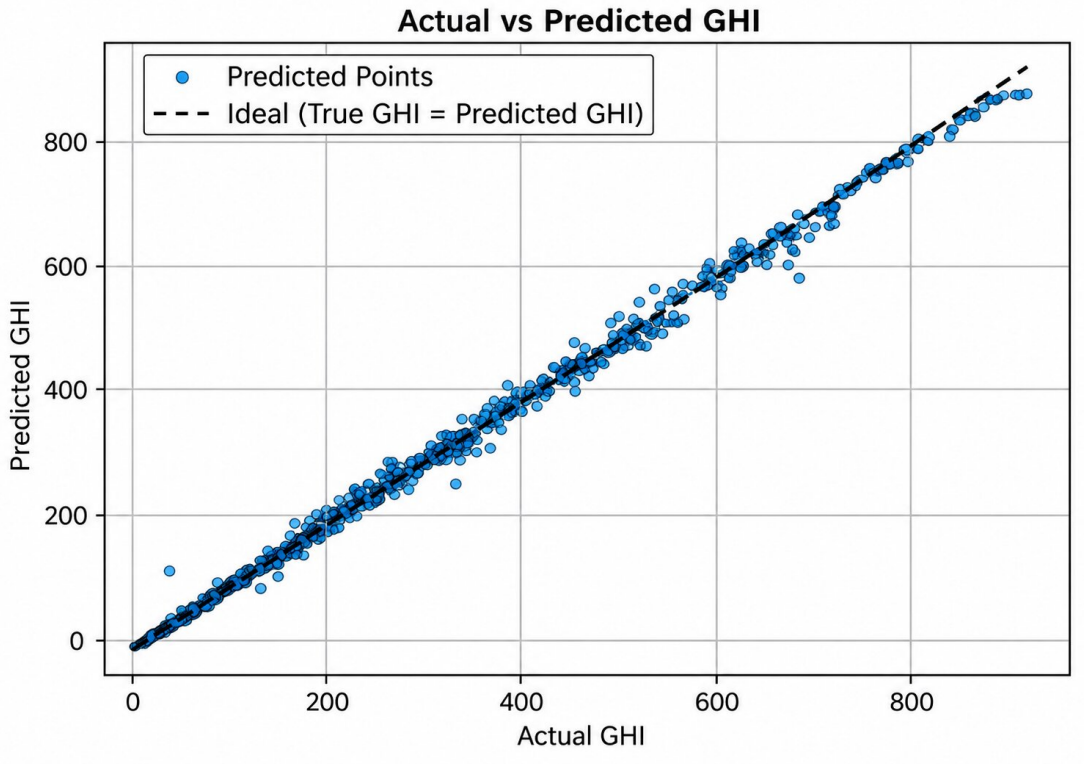



**Figure 3.** Actual vs. predicted GHI.

### *4.3. Stage 2: Energy Consumption Forecasting Results*

Separate CNN+2LSTM+embedding models were trained for the heating season (October–April) and cooling season (May–September). Table 2 summarizes performance: the heating model achieved $R^2$ = 0.9730 (MAE = 0.190 kWh, RMSE = 0.276 kWh), while the cooling model achieved $R^2$ = 0.9140 (MAE = 0.018 kWh, RMSE = 0.030 kWh). Although the cooling model exhibits a lower $R^2$, its absolute error is much smaller in real units, since cooling loads

in this building stock are an order of magnitude smaller than heating loads. The actual vs. predicted energy forecast result are graphed in Figures 4 and 5 for both heating and cooling.

**Table 2.** Performance metrics for heating and cooling energy models.

| CNN+2LSTM+Embedding | $R^2$ | MAE (kWh) | RMSE (kWh) |
|---|---|---|---|
| Cooling | 0.9140 | 0.018 | 0.030 |
| Heating | 0.9730 | 0.190 | 0.276 |

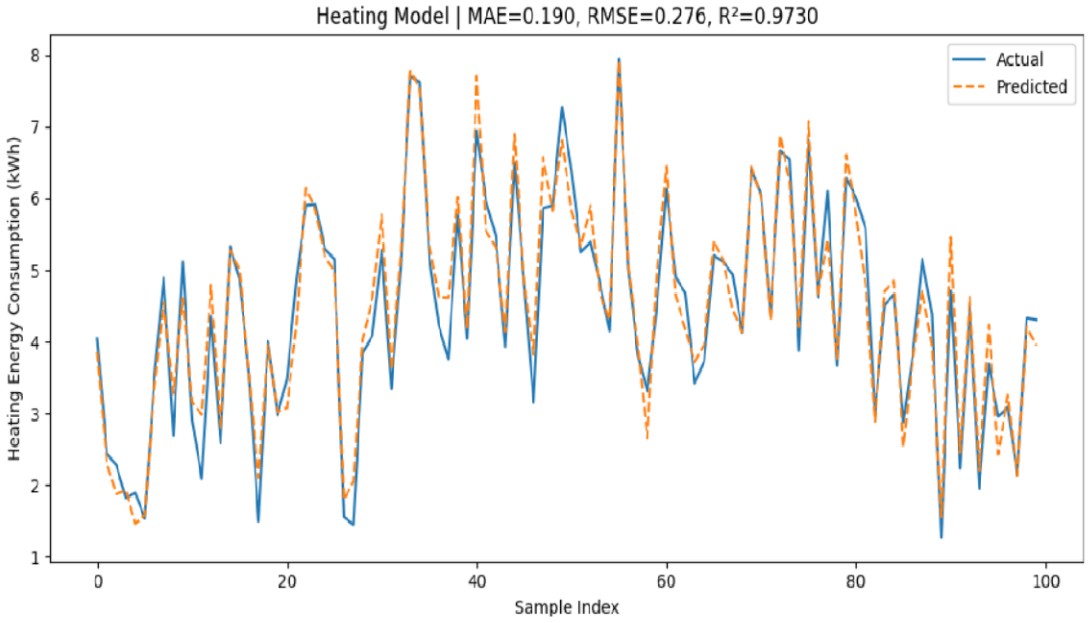



**Figure 4.** Actual vs. predicted heating energy consumption.

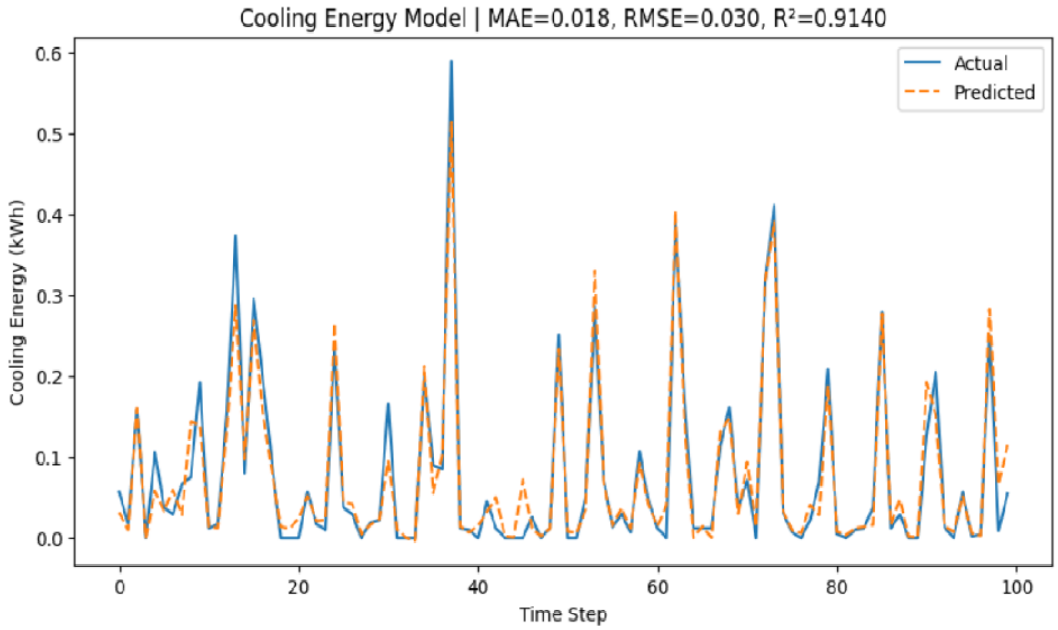



**Figure 5.** Actual vs. predicted cooling energy consumption.

From Figures 4 and 5 we can clearly see that the predicted cooling and heating energy consumption is very similar to the actual cooling and heating energy consumption. This shows that the model is able to capture the temporal aspect of the energy consumption through the modelling of the peaks, troughs, and turning points.

The model was trained for 120 epochs using the Adam optimizer with a learning rate of 0.01 and mean squared error (MSE) loss on an 80/20 train-test split. No explicit regularization techniques such as dropout, L2 penalties, or early stopping were applied during training.

### *4.4. Stage 3: Operational Mode Classification Results*

Four Random Forest classifiers were trained: heating/actual, heating/predicted, cooling/actual, and cooling/predicted, each using GHI, energy consumption, solar elevation, and cyclic time features as inputs. Across all four configurations, classification accuracy exceeded 99%.

Table 3 summarizes the distribution of predicted operational labels relative to the actual-label distribution across heating and cooling seasons. In the cooling season, predicted and actual distributions are nearly identical, indicating that small forecast errors in GHI and energy consumption have little effect on the resulting control decisions. In the heating season, however, a more noticeable shift occurs: the OFF class share drops from 18.8% under actual labels to 4.0% under predicted labels, while GRID increases from 54.0% to 61.1%. This shift is attributable to the comparatively low absolute energy thresholds involved, where small forecasting errors near a decision boundary are more likely to flip a sample from OFF to GRID than from one solar-dominant class to another, highlighting that classification reliability near threshold boundaries is sensitive to upstream forecast precision, even when overall classification accuracy remains very high.

**Table 3.** Distribution of operational labels across heating and cooling scenarios.

| Scenario | SOLAR (%) | GRID (%) | OFF (%) | Samples |
|---|---|---|---|---|
| Cooling–Actual labels | 57.2 | 3.9 | 38.8 | 1064 |
| Cooling–Predicted labels | 57.7 | 3.7 | 38.6 | 1064 |
| Heating–Actual labels | 27.2 | 54.0 | 18.8 | 1080 |
| Heating–Predicted labels | 34.9 | 61.1 | 4.0 | 1080 |

The model was trained on an 80/20 train-test split. The Random Forest classifier was configured with class_weight = 'balanced'. For cooling models, the number of estimators (trees) was empirically set to 150. For heating models, the number of estimators (trees) was empirically set to 100.

### *4.5. Comparison with Existing Irradiance Forecasting Studies*

Table 4 benchmarks the proposed Stage 1 ensemble against three peer-reviewed solar irradiance forecasting studies. The proposed model achieves an RMSE of 11.80 W/m$^2$ and an $R^2$ of 0.9975, substantially exceeding the accuracy reported for support-vector-regression with similar-day averaging [8], a recurrent neural network applied to short-term forecasting [14], and a CNN-LSTM hybrid model [2]. The comparatively higher error rates in these studies are attributable to factors such as reliance on temporal clustering rather than engineered solar-geometry features [8], limited feature engineering and normalization [14], and the absence of log-scaling or nighttime-record exclusion, which can amplify the influence of extreme irradiance values during training [2]. The performance advantage observed here is therefore attributed not only to the ensemble averaging itself but to the combination of domain-informed preprocessing nighttime exclusion, log transformation, and solar-geometry feature engineering with two complementary boosting algorithms.

**Table 4.** Comparative analysis of GHI forecasting methods.

| Study | Model | Horizon | RMSE | MAE | $R^2$ |
|---|---|---|---|---|---|
| This work | XGBoost + LightGBM | Hourly (daylight) | 11.80 | 7.47 | 0.9975 |
| Houchati et al. [8] | SVR + similar-day avg. | Day-ahead (hourly) | 67.64 | 47.54 | 0.94 |
| Rajagukguk et al. [14] | RNN | Short-term (1–6 h) | 156.34 | 103.12 | 0.853 |
| Kumari & Toshniwal [2] | CNN-LSTM hybrid | Hourly | 134.5 | 89.6 | 0.88 |

### *4.6. Comparison with Existing Energy Consumption Forecasting Studies*

Table 5 compares the Stage 2 model with three peer-reviewed building energy consumption forecasting studies. The proposed CNN+2LSTM+embedding model achieves the lowest MAE (0.19) and the highest $R^2$ (0.973) among the compared studies. A likely contributor to this performance is the building-identity embedding, which allows a single architecture to capture latent, building-specific consumption behavior; by contrast, Kim and Cho [16] do not incorporate embeddings or seasonal separation, and Le et al. [17], despite using a bidirectional LSTM, do not include targeted time-based preprocessing or building-level representation. The proposed model is also unusual among these comparisons in being evaluated separately under heating and cooling seasonal regimes rather than on an aggregated dataset, which more directly tests its ability to generalize across qualitatively different load-driving mechanisms.

**Table 5.** Comparative evaluation of energy consumption forecasting models.

| Study | Model | MAE | RMSE | $R^2$ |
|---|---|---|---|---|
| Kim & Cho [16] | CNN-LSTM | 0.33 | 0.59 | — |
| Le et al. [17] | CNN + Bi-LSTM | 0.38 | 0.52 | 0.85 |
| Sajjad et al. [18] | CNN-GRU | 0.24 | 0.31 | — |
| This work | CNN + 2LSTM + Embedding | 0.19 | 0.276 | 0.973 |

## 5. Conclusions and Future Work

This paper presented a modular, three-stage machine learning framework for solar-assisted residential energy management, validated entirely on real, measured 2018 weather and building-energy data. Stage 1 demonstrated that a log-transformed ensemble of XGBoost and LightGBM, enriched with solar-geometry and cyclic time features, can forecast hourly Global Horizontal Irradiance with very high accuracy ($R^2$ = 0.9975) using only routinely available meteorological inputs. Stage 2 showed that a CNN+2LSTM architecture with trainable building-identity embeddings can predict heating and cooling energy consumption across multiple residential buildings with strong accuracy ($R^2$ = 0.9730 and 0.9140, respectively) while generalizing across distinct building-specific consumption behaviors. Stage 3 demonstrated that a Random Forest classifier, trained on rule-derived labels and predicted Stage 1/Stage 2 outputs, can assign operational SOLAR/GRID/OFF decisions with over 99% accuracy while remaining fully interpretable through Gini feature-importance analysis.

Several limitations should be considered when interpreting these results. First, the analysis is based on a single year of data (2018) and only twelve residential buildings from North Dakota, which limits the extent to which the results can be generalized across building types, climates, and occupant behaviors. Although North Dakota was selected because of its climatic similarity to Regina, differences in building construction, occupancy, and appliance use may introduce unquantified transfer bias. Second, the dataset does not contain real-time internal building variables such as occupancy or indoor temperature, which may partly explain the larger uncertainty observed for cooling-load forecasting. Third, the Stage 3 operational labels are based on empirical energy thresholds and a GHI threshold rather than detailed PV-system specifications, inverter constraints, economic signals, or measured control actions. Accordingly, the proposed framework should be regarded as a proof-of-concept decision-support framework rather than a fully validated physical building-control system

Future work should pursue several extensions. Real-time sensor integration such as indoor temperature monitoring, occupancy detection, and live irradiance sensing would allow the framework to dynamically update its forecasts and reduce the boundary-sensitivity observed in heating-season classification. Incorporating building-specific hardware parameters, such as photovoltaic panel capacity, inverter ratings, and storage specifications, would move the Stage 3 classifier from a proof-of-concept toward a deployable control system. Finally, newer sequence architectures, including transformer-based time-series models or CNN-attention hybrids, may further improve long-range seasonal pattern capture without hand-engineered temporal features.

## Author Contributions

B.S.: conceptualization, methodology, investigation, data curation, software, formal analysis, visualization, writing—original draft preparation; S.L.: methodology, software, validation, writing—review and editing; W.P.: supervision, project administration, and guidance. All authors have read and agreed to the published version of the manuscript.

## Funding

This research received no external funding.

## Institutional Review Board Statement

Not applicable.

## Informed Consent Statement

Not applicable.

## Data Availability Statement

All data was obtained and available online.

## Conflicts of Interest

The authors declare no conflict of interest.

## Use of AI and AI-Assisted Technologies

No AI tools were utilized for this paper.

## References


1. Aguilar, J.; Garces-Jimenez, A.; R-Moreno, M.D.; et al. A Systematic Literature Review on the Use of Artificial Intelligence in Energy Self-Management in Smart Buildings. *Renew. Sustain. Energy Rev.* **2021**, *151*, 111530. https://doi.org/10.1016/j.rser.2021.111530.
2. Kumari, P.; Toshniwal, D. Deep Learning Models for Solar Irradiance Forecasting: A Comprehensive Review. *J. Clean. Prod.* **2021**, *318*, 128566. https://doi.org/10.1016/j.jclepro.2021.128566.
3. Zhou, H.; Zhou, Y.; Hu, J.; et al. LSTM-Based Energy Management for Electric Vehicle Charging in Commercial-Building Prosumers. *J. Mod. Power Syst. Clean Energy* **2021**, *9*, 1205–1216. https://doi.org/10.35833/mpce.2020.000501.
4. Duffie, J.A.; Beckman, W.A. *Solar Engineering of Thermal Processes*, 4th ed.; Wiley: Hoboken, NJ, USA, 2013. https://doi.org/10.1002/9781118671603.
5. Salari, A.; Ahmadi, S.E.; Marzband, M.; et al. Fuzzy Q-Learning-Based Approach for Real-Time Energy Management of Home Microgrids Using Cooperative Multi-Agent System. *Sustain. Cities Soc.* **2023**, *95*, 104528. https://doi.org/10.1016/j.scs.2023.104528.
6. Nabavi, S.A.; Motlagh, N.H.; Zaidan, M.A.; et al. Deep Learning in Energy Modeling: Application in Smart Buildings with Distributed Energy Generation. *IEEE Access* **2021**, *9*, 125439–125461. https://doi.org/10.1109/access.2021.3110960.
7. Mehrjerdi, H.; Saad, M.; Lefebvre, S. Efficiency-Resilience Nexus in Building Energy Management under Disruptions and Events. *IEEE Syst. J.* **2022**, *16*, 299–308. https://doi.org/10.1109/jsyst.2020.3021697.
8. Houchati, M.; Beitelmal, A.H.; Khraisheh, M. Predictive Modeling for Rooftop Solar Energy Throughput: A Machine Learning-Based Optimization for Building Energy Demand Scheduling. *J. Energy Resour. Technol.* **2022**, *144*, 011302 https://doi.org/10.1115/1.4050844.
9. Mansour, S.H.; Azzam, S.M.; Hasanien, H.M.; et al. Wasserstein Generative Adversarial Networks-Based Photovoltaic Uncertainty in a Smart Home Energy Management System Including Battery Storage Devices. *Energy* **2024**, *306*, 132412. https://doi.org/10.1016/j.energy.2024.132412.
10. Wang, Z.; Liu, J.; Zhang, Y.; et al. Practical Issues in Implementing Machine-Learning Models for Building Energy Efficiency: Moving beyond Obstacles. *Renew. Sustain. Energy Rev.* **2021**, *143*, 110929. https://doi.org/10.1016/j.rser.2021.110929.
11. Kurukuru, V.S.B.; Haque, A.; Khan, M.A.; et al. A Review on Artificial Intelligence Applications for Grid-Connected Solar Photovoltaic Systems. *Energies* **2021**, *14*, 4690. https://doi.org/10.3390/en14154690.
12. Tina, G.M.; Ventura, C.; Ferlito, S.; et al. A State-of-Art-Review on Machine-Learning Based Methods for PV. *Appl. Sci.* **2021**, *11*, 7550. https://doi.org/10.3390/app11167550.
13. Pang, Z.; Niu, F.; O'Neill, Z. Solar Radiation Prediction Using Recurrent Neural Network and Artificial Neural Network: A Case Study with Comparisons. *Renew. Energy* **2020**, *156*, 279–289. https://doi.org/10.1016/j.renene.2020.04.042.
14. Rajagukguk, R.A.; Ramadhan, R.A.A.; Lee, H.J. A Review on Deep Learning Models for Forecasting Time Series Data of Solar Irradiance and Photovoltaic Power. *Energies* **2020**, *13*, 6623. https://doi.org/10.3390/en13246623.
15. Wang, H.; Liu, Y.; Zhou, B.; et al. Taxonomy Research of Artificial Intelligence for Deterministic Solar Power Forecasting. *Energy Convers. Manag.* **2020**, *214*, 112909. https://doi.org/10.1016/j.enconman.2020.112909.
16. Kim, T.Y.; Cho, S.B. Predicting Residential Energy Consumption Using CNN-LSTM Neural Networks. *Energy* **2019**, *182*, 72–81. https://doi.org/10.1016/j.energy.2019.05.230.
17. Le, T.; Vo, M.; Vo, B.; et al. Improving Electric Energy Consumption Prediction Using CNN and Bi-LSTM. *Appl. Sci.* **2019**, *9*, 4237. https://doi.org/10.3390/app9204237.
18. Sajjad, M.; Khan, Z.A.; Ullah, A.; et al. A Novel CNN-GRU-Based Hybrid Approach for Short-Term Residential Load Forecasting. *IEEE Access* **2020**, *8*, 143759–143768. https://doi.org/10.1109/access.2020.3009537.
19. Drir, N.; Chekired, F.; Rekioua, D. An Integrated Neural Network for the Dynamic Domestic Energy Management of a Solar House. *Int. Trans. Electr. Energy Syst.* **2021**, *31*, e13227. https://doi.org/10.1002/2050-7038.13227.
20. Ben Slama, S.; Mahmoud, M. A Deep Learning Model for Intelligent Home Energy Management System Using Renewable Energy. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106388. https://doi.org/10.1016/j.engappai.2023.106388.
21. Chou, J.S.; Tran, D.S. Forecasting Energy Consumption Time Series Using Machine Learning Techniques Based on Usage Patterns of Residential Householders. *Energy* **2018**, *165*, 709–726. https://doi.org/10.1016/j.energy.2018.09.144.
22. Paul, S.; Padhy, N.P. Real-Time Bilevel Energy Management of Smart Residential Apartment Building. *IEEE Trans. Ind. Inform.* **2020**, *16*, 3708–3720. https://doi.org/10.1109/tii.2019.2941739.

23. Zhang, B.; Hu, W.; Xu, X.; et al. Physical-Model-Free Intelligent Energy Management for a Grid-Connected Hybrid Wind-Microturbine-PV-EV Energy System via Deep Reinforcement Learning Approach. *Renew. Energy* **2022**, *200*, 433–448. https://doi.org/10.1016/j.renene.2022.09.125.
24. Wei, G.; Chi, M.; Liu, Z.W.; et al. Deep Reinforcement Learning for Real-Time Energy Management in Smart Home. *IEEE Syst. J.* **2023**, *17*, 2489–2499. https://doi.org/10.1109/jsyst.2023.3247592.
25. Xu, Y.; Gao, W.; Li, Y.; et al. Operational Optimization for the Grid-Connected Residential Photovoltaic-Battery System Using Model-Based Reinforcement Learning. *J. Build. Eng.* **2023**, *73*, 106774. https://doi.org/10.1016/j.jobe.2023.106774.
26. Xu, X.; Jia, Y.; Xu, Y.; et al. A Multi-Agent Reinforcement Learning-Based Data-Driven Method for Home Energy Management. *IEEE Trans. Smart Grid* **2020**, *11*, 3201–3211. https://doi.org/10.1109/tsg.2020.2971427.
27. Chen, L.; Wu, J.; Tang, H.; et al. A Q-Learning Based Optimization Method of Energy Management for Peak Load Control of Residential Areas with CCHP Systems. *Electr. Power Syst. Res.* **2023**, *214*, 108895. https://doi.org/10.1016/j.epsr.2022.108895.
28. Lee, S.; Choi, D.H. Energy Management of Smart Home with Home Appliances, Energy Storage System and Electric Vehicle: A Hierarchical Deep Reinforcement Learning Approach. *Sensors* **2020**, *20*, 2157. https://doi.org/10.3390/s20072157.
29. Zeynali, S.; Rostami, N.; Ahmadian, A.; et al. Two-Stage Stochastic Home Energy Management Strategy Considering Electric Vehicle and Battery Energy Storage System: An ANN-Based Scenario Generation Methodology. *Sustain. Energy Technol. Assess.* **2020**, *39*, 100722. https://doi.org/10.1016/j.seta.2020.100722.
30. Li, B.; Liu, Z.; Wu, Y.; et al. Review on Photovoltaic with Battery Energy Storage System for Power Supply to Buildings: Challenges and Opportunities. *J. Energy Storage* **2023**, *61*, 106763. https://doi.org/10.1016/j.est.2023.106763.
31. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the KDD'16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. https://doi.org/10.1145/2939672.2939785.
32. Ke, G.; Meng, Q.; Finley, T.; et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 3146–3154.
33. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.
34. Alhussein, M.; Aurangzeb, K.; Haider, S.I. Hybrid CNN-LSTM Model for Short-Term Individual Household Load Forecasting. *IEEE Access* **2020**, *8*, 180544–180557. https://doi.org/10.1109/access.2020.3028281.