frontiers in Energy Research

TYPE Original Research
PUBLISHED 05 November 2026
DOI 10.3389/fenrg.2026.1813402

Check for updates

OPEN ACCESS

EDITED BY
Paul Arévalo-Cordero,
University of Cuenca, Ecuador

REVIEWED BY
Kenneth E. Okedu,
Melbourne Institute of Technology, Australia
Joshuva Arockia Dhanraj,
Dayananda Sagar University, India

*CORRESPONDENCE
Muhammad Waseem,
mwaseem148@yahoo.com

RECEIVED 18 February 2026
REVISED 13 August 2026
ACCEPTED 31 August 2026
PUBLISHED 05 November 2026

CITATION
Abbas MZ, Sajjad MIA, Ashraf MM, Waseem M, Khan FH, Ghafoor MZ and Hassan U (2026) DeepANFIS: an interpretable and uncertainty-aware neuro-fuzzy framework for energy disaggregation in smart microgrids. *Front. Energy Res.* 14:1813402. doi: 10.3389/fenrg.2026.1813402

COPYRIGHT
© 2026 Abbas, Sajjad, Ashraf, Waseem, Khan, Ghafoor and Hassan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

DeepANFIS: an interpretable and uncertainty-aware neuro-fuzzy framework for energy disaggregation in smart microgrids

Muhammad Zaigham Abbas[1], Malik Intisar Ali Sajjad[1], Muhammad Mansoor Ashraf[1], Muhammad Waseem[2]*, Faisal Hammad Khan[1], Muhammad Zubair Ghafoor[1] and Usman Hassan[1]

[1]Electrical Engineering Department University of Engineering and Technology Taxila, Taxila, Pakistan, [2]International Research on Energy System Integration, Education, and Environment for Sustainability and Innovation (IRESI) Centre, School of Business, Maynooth University, Maynooth, Ireland

Non-intrusive load monitoring (NILM) facilitates appliance-level energy disambiguation based on aggregate power consumption measurements and is an important facilitator of energy efficiency and demand-side management. However, state-of-the-art NILM methods tend to have low interpretability, low sensitivity to the sampling rate, and difficulty in providing reliable uncertainty information, which delay their effectiveness in practical implementations. To address these issues, we present a new deep adaptive neuro-fuzzy inference system (DeepANFIS) for joint appliance-state identification and power estimation with enhanced interpretability and uncertainty awareness. The proposed system comprehensively incorporates modular convolutional feature learning with multiple resolutions, fuzzy memory networks with temporal dependencies, hierarchical fuzzy inference, and evidential deep learning. Extensive experiments were conducted on both the UK-DALE dataset and REDD to illustrate the efficacy of DeepANFIS under high-frequency and low-frequency sampling schemes. On the UK-DALE dataset, DeepANFIS attained an F1-score of 0.7122 with a mean absolute error (MAE) of 25.63 W at a sampling frequency of 6 s, along with an F1-score of 0.8371 and a root mean-squared error (RMSE) of 78.03 W at a sampling frequency of 15 min. On the REDD, DeepANFIS obtained a near-perfect F1-score of 0.9869 with an accuracy of 0.9992 and a low MAE of 2.56 W. As part of the cross-dataset analysis, evaluation on REDD showed excellent F1-score classification capability with a slight drop of 2.4%, but the power estimation transfer remained a challenge. An ablation analysis was further performed to validate the significances of both temporal memory and fuzzy processing in DeepANFIS. Thus, the efficiency of DeepANFIS was established with respect to a non-interpretable uncertainty model for NILM.

KEYWORDS
deep neuro-fuzzy learning, non-intrusive load monitoring, power disaggregation, temporal memory network, uncertainty-aware

1 Introduction

The load on the energy sector is growing rapidly, irrespective of whether it is residential, commercial, or industrial load. Industrial expansion of a country for national advancement and commercial growth to satisfy human aspirations are the main causes of such load increments. Energy conservation and energy efficiency are some of the main challenges in the energy sector. In developing countries, where smart grids are in the pipeline for progress and conventional energy costs are higher, it is necessary to start energy conservation and energy efficiency programs to meet customer requirements. In this context, demand-side management is one of the approaches for providing programs, technologies, and actions to control electrical energy usage.

Several strategies have been developed and adopted for load control, especially in the residential sector that accounts for roughly 40% of the energy consumption globally (Owais et al., 2018). Strategies such as price-based and incentive-based response programs are used to ensure active customer participation both directly and indirectly. However, energy conservation programs such as non-intrusive load monitoring (NILM) are highly useful for providing feedback to customers about their energy usage patterns at the individual appliance level. Programs such as demand response and NILM are some of the topics of interest in the smart building domain. Figure 1 shows that demand-side management in smart grids requires information on customer behaviors, which can be obtained through NILM; moreover, information on variable load patterns is required, which can be achieved through demand response programs.

The NILM technique aims to disaggregate the total power consumption of a household or building into its constituent appliances without the need for individual metering. It was originally introduced by Hart (1992) and leverages machine-learning and signal processing techniques to analyze the aggregate electrical signal from a single point, such as a smart meter, and to identify the operational patterns of various electrical devices within the premises. Thus, NILM helps conserve energy by providing meaningful consumption data with improved efficiency and supported demand management. The evolution of NILM has been remarkable owing to advancements in high-resolution consumption data and the availability of advanced computational techniques. The latest NILM technology utilizes deep-learning architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to increase the precision and resolution of appliance identification. These architectures can process and handle the non-linear features of consumption signals while adjusting to different household settings. Notably, the adoption of Internet of Things (IoT) technology has made it possible to offer instant feedback and control in NILM. Thus, NILM has emerged as a technological bedrock for smart grids, enabling optimal management of energy resources while encouraging sustainable living practices.

2 Related studies

NILM is a vital technology in the context of smart grids and energy management; it refers to the process of disaggregating the total energy consumption of a household into those of the constituent appliances to achieve better energy monitoring, efficiency, and management. Therefore, the following review is an attempt to collate recent advances and methodologies in NILM, with specific focus on deep-learning approaches, optimization techniques, and practical implementations.

Li et al. (2024) proposed a dynamic-time-warping-optimization-based NILM, where dynamic time warping was used to optimize NILM for monitoring several household appliances; this work emphasized improving the accuracy of load identification through enhanced alignment techniques. Rafiq et al. (2024) reviewed the current methods and challenges of advanced deep-learning-based NILM; they noted the need for better models that can capture the complexity and variability of residential data. Chen et al. (2024) proposed a new approach based on time-series self-supervised learning that uses temporal CNNs to improve the accuracy of NILM, especially when the labeled data are limited. Varanasi and Karri (2024) reported a channel-attention-guided bidirectional temporal convolutional network (TCN) for Seq2Point learning; this approach enhances NILM through a bidirectional TCN guided by channel attention mechanisms to improve the accuracy of sequence-to-point learning models. Dai et al. (2024) presented a distributed and privacy-preserving framework called DP2-NILM for maintaining high accuracy while ensuring privacy; this framework focuses on privacy preservation in NILM by distributing the computational load and data across multiple nodes. By combining multiple regression models, Akbar et al. (2024a) proposed a Bayes-ensemble regressor technique for NILM with improved performance. Liu et al. (2024) introduced a novel NILM methodology to identify appliance operational status accurately, thereby enhancing residential electrical safety and management.

Shabbir et al. (2024) performed a comparative analysis to evaluate different machine-learning techniques, providing insights into their strengths and weaknesses for NILM applications. Yao et al. (2024) introduced a multiscale attention mechanism to improve NILM performance by focusing on multiple temporal scales within the data. Pu et al. (2024) integrated weather and calendar features into denoising autoencoders to enhance NILM performance by accounting for external factors influencing energy consumption. Kommey et al. (2024) proposed a modified K-nearest neighbor algorithm for artificial-intelligence-based NILM and showed better energy consumption monitoring in electrical energy systems. Esmaeili Shayan (2024) discussed the application of IoT and NILM in microgrid energy management systems and showed that they are useful for real-time monitoring and control. A novel unsupervised algorithm for NILM was proposed by Criado-Ramón et al. (2024), who showed its potential to identify appliances from unlabeled data with great accuracy. A few-shot transfer learning approach was proposed by Ding et al. (2024) using meta-learning and relational networks; this work tackled one of the biggest challenges when a small number of training data is available for load recognition. Zhou et al. (2024) presented an efficient deep-learning model with local feature extraction to enhance the precision of NILM in residential settings. A vision-transformer-based technique was discussed by Zhao et al. (2024) for detecting unknown appliances; this study leveraged vision transformers with an additional detection head to identify unknown appliances, thereby improving the

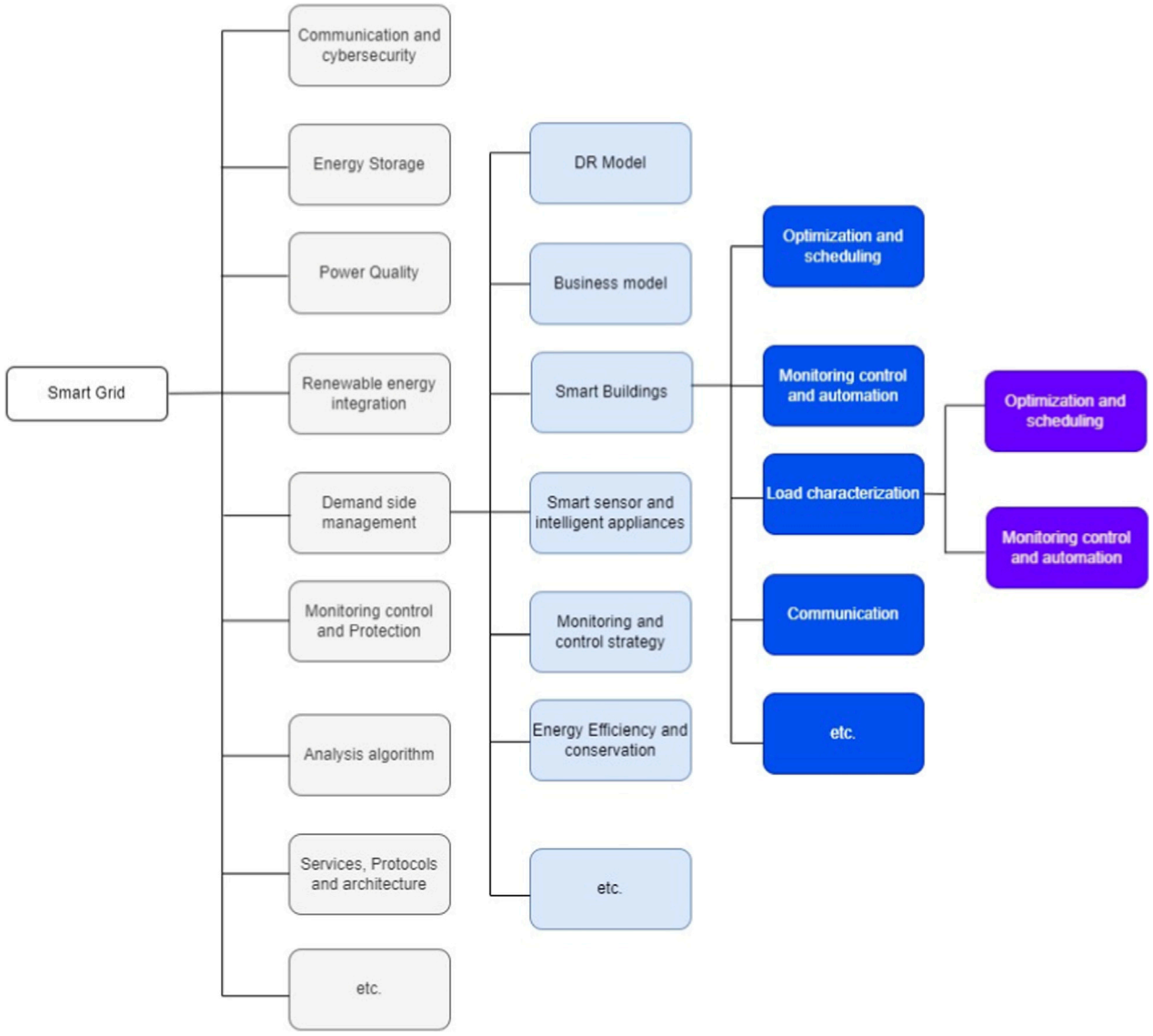


FIGURE 1
Research topics in the context of smart grids.

adaptability and accuracy of NILM. A semi-supervised deep-learning framework was proposed by Akbar et al. (2024b) to enhance NILM in smart grids by combining labeled and unlabeled data for training. A deep-learning-based NILM for three-phase systems was proposed by Gowrienanthan et al. (2023), which extends NILM capabilities beyond single-phase residential applications. An active learning framework was presented by Todic et al. (2023) for low-frequency NILM.

Given these recent reports in literature, our work contributes to NILM as follows:

1. A deep neuro-fuzzy framework is used to jointly achieve power regression and appliance classification; this framework is enhanced using temporal attention and memory mechanisms.
2. Evidential deep learning (EDL) is integrated in the proposed approach to quantify prediction uncertainty, thereby improving robustness in noisy and ambiguous real-world conditions.
3. A multiresolution convolutional module is proposed to capture appliance behaviors across varying temporal scales from transient spikes to prolonged cycles.
4. Extensive experiments are performed on the public United Kingdom domestic appliance-level electricity (UK-DALE) dataset (Kelly and Knottenbelt, 2015) and reference energy disaggregation dataset (REDD) (Kolter and Johnson, 2011) under different sampling rates, along with cross-dataset generalization tests and ablation studies to validate the architectural contributions.

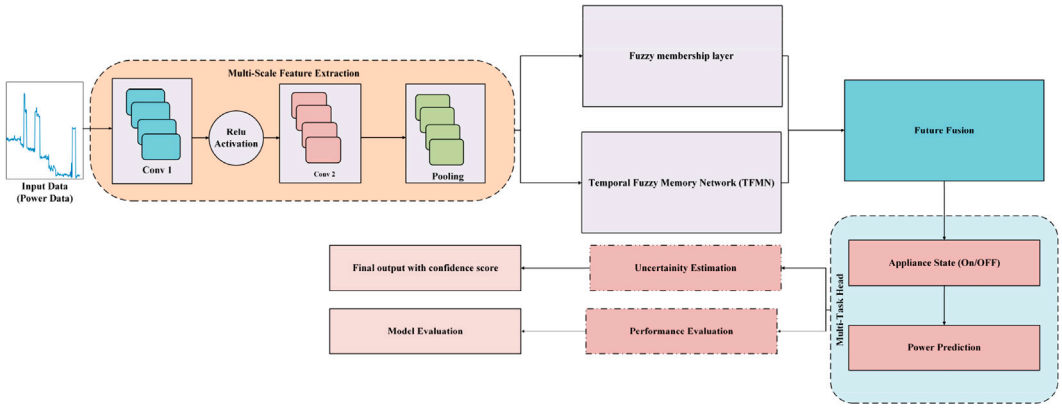


FIGURE 2
Proposed deep adaptive neuro-fuzzy inference system (DeepANFIS) model for non-intrusive load monitoring (NILM).

Unlike other hybrid systems that integrate neural networks and fuzzy inference systems as two separate modules, the proposed deep adaptive neuro-fuzzy inference system (DeepANFIS) is an end-to-end trainable architecture in which the feature extraction, temporal memory, hierarchical fuzzy reasoning, and uncertainty estimation processes are all optimized simultaneously. This means that the proposed approach is a new neuro-fuzzy learning system altogether instead of an integrated set of existing components.

3 Mathematical modeling

This section presents the mathematical formulation of the proposed DeepANFIS model. The model combines multiresolution feature extraction, temporal fuzzy memory, hierarchical fuzzy inference, and evidential uncertainty learning for robust and interpretable power prediction and appliance classification in NILM. The core concepts of DeepANFIS are illustrated in Figure 2.

3.1 Signal representation

The main power signal in NILM is the time series containing overlapping signatures from the household appliances. This time series is decomposed into segments containing both temporal and amplitude modulations that must be retained. The input signal to the model is expressed by Equation 1:

$$X \in \mathbb{R}^{B*C*T}, \tag{1}$$

where B is the batch size, C is the number of input channels, and T is the sequence length that holds the sampled values of the power measurements. The model receives these data in mini batches during the training process.

3.2 Adaptive Gaussian fuzzy membership layer

The process of fuzzification involves using membership functions to change crisp numbers into linguistic values. The use of Gaussian functions is preferred here because the functions are rendered smooth and differentiable. This layer learns both the location and width of each membership function; it can also be adapted to varying inputs (Equation 2). Each input dimension x_i is fuzzified using Gaussian membership functions that are then adapted in a manner similar to Equation 3:

$$\mu_{i,m}(x_i) = exp\left(-\frac{(x_i - c_{i,m})^2}{2W_{i,m}^2}\right); \quad i = 1,\ldots,d; \quad m = 1,\ldots,M, \tag{2}$$

here, $C_{i,m}$ and $W_{i,m}$ are the learnable center and width of the m^{th} Gaussian membership function for input i, respectively. Each input is thus represented by multiple overlapping fuzzy sets, forming a fuzzy feature space that captures non-linear relationships better.

The resulting membership matrix is given by

$$\mu(x) = \mathbb{R}^{d*M}, \tag{3}$$

where d represents the input dimension and M is the number of membership functions per input.

3.3 Hierarchical fuzzy inference layer

This layer models rule-based reasoning such as the Takagi–Sugeno–Kang (TSK) fuzzy system. Instead of a single-layer fuzzy logic, DeepANFIS uses a hierarchical arrangement to improve its ability to model high-dimensional and non-linear systems without exponential rule explosion.

The activation rule for the DeepANFIS model can be defined as Equation 4:

$$s_r(x) = \sigma(R_r u(x)), \tag{4}$$

where $u(x)$ is the flattened membership vector, R_r are learnable rule premise weights, and $\sigma(\cdot)$ is the sigmoid function. Once the rules are defined, normalization is performed to ensure that the rule activations are summed to one. The normalization is given by Equation 5:

$$\bar{s}_r(x) = \frac{s_r(x)}{\sum_{r'=1}^{R} s_{r'}(x) + \varepsilon}. \tag{5}$$

where ε is a small positive constant added for numerical stability, preventing division by zero when the sum of rule activations approaches zero.

After layer normalization, the TSK consequent is given by Equation 6:

$$y_r(x) = A_r[x;1], \tag{6}$$

where A_r represents the consequent parameters with bias. After defining the consequent parameters, defuzzification is performed to ensure that each fuzzy rule contributes proportionally to the final output, thereby blending neural adaptability with fuzzy interpretability.

One important benefit of this hierarchical formulation is that the Cartesian products of all the membership functions need not be enumerated. In a typical ANFIS framework with D input dimensions and M membership functions for each dimension, the complete construction of rules results in a total of M^D rules. However, the hierarchical representation in the DeepANFIS model maps the discovered membership functions to a reduced number of rules using the premise weights. Thus, the total number of rules is bound by the fuzzy rule dimension and not by the Cartesian product of all input memberships. This means that despite having fewer rules, the interpretability of the rules is achieved in the DeepANFIS framework. This interpretability is possible because the predictions are obtained from the normalized rule activations and their TSK consequents. Contrary to the opacity of neural networks, the contributions of the rules are observable at all times, thus making it possible for experts to analyze the linguistic rules that dominate the predictions.

3.4 Temporal fuzzy memory network

The appliance states and transitions exhibit temporal dependencies; for example, the operations of a washing machine evolve through various phases. The temporal fuzzy memory network (TFMN) combines the memory ability of long short-term memory (LSTM) networks and the interpretability of fuzzy logic to retain historical context while learning the non-linear dependencies. The feature encoding for the TFMN is given by Equation 7:

$$e_t = \phi(Ef_t + b_E), \tag{7}$$

where $\phi(\cdot)$ = ReLU is the non-linear activation function while $b_E \in \mathbb{R}^h$ and $E \in \mathbb{R}^{d_f * h}$ are learnable parameters. This step produces fuzzy-encoded temporal features given by Equation 8:

$$E = [e_1, e_1, e_1, \ldots, e_t'] \in \mathbb{R}^{B*h*T'}. \tag{8}$$

Here, h represents the fuzzy hidden dimension. The fuzzy encoder enables each time step to represent the soft membership degrees of the latent temporal states as fuzzy rules over time. The sequence E is processed by a bidirectional LSTM (BiLSTM) network (Equation 9) where each LSTM cell maintains two forms of memory: cell state (c_t) for long-term memory and hidden state (h_t) for short-term or current contextual information.

$$H = \text{BiLSTM}(E) \in \mathbb{R}^{B*2h*T'}. \tag{9}$$

The bidirectional architecture allows information to flow from both past and future time steps, as shown in Equation 10:

$$H_t = \left[\overrightarrow{h_{lstm}}; \overleftarrow{h_{lstm}}\right]. \tag{10}$$

The BiLSTM captures long-term dependencies and temporal logic that the fuzzy rules alone cannot represent; while the LSTM captures sequential dependencies, not all time steps contribute equally.

The attention mechanism learns to assign a weight α_t to each hidden state as in Equation 11:

$$\alpha_t = \frac{\exp\left(\frac{Q_t * K_t^T}{\sqrt{d_k}}\right)}{\sum_{j=1}^{T'} \exp\left(\frac{Q_t * K_t^T}{\sqrt{d_k}}\right)}, \tag{11}$$

where Q, K, and V are the query, key, and value projections of H. Here, the time steps with more informative features receive higher attention weights, thus improving interpretability and energy event detection. The attended sequence is given by Equation 12:

$$A = \sum_{t=1}^{T'} \alpha_t V_t \tag{12}$$

To preserve the temporal context between consecutive inputs, a fuzzy memory update mechanism (Equations 13, 14) is used to integrate the current attended representation with the previous memory:

$$m_{\text{new}} = U\left(\frac{1}{T'} \sum_{t=1}^{T'} A_t\right), \tag{13}$$

$$m_t = \alpha_m m_{\text{new}} + (1 - \alpha_m) m_{t-1}, \tag{14}$$

where $U(\cdot)$ is a small feed-forward network for non-linear transformation, m_t is the current fuzzy memory state, and $\alpha_m \in (0, 1)$ is the memory update coefficient controlling the balance between the new and old information. This mechanism mimics human cognitive memory, where the old information is blended with new information according to relevance. The final temporal fuzzy representation combines the attended hidden states and updated memory, as shown in Equation 15:

$$h_{TFMN} = [A; m_t] \in \mathbb{R}^{B*(2h+m_s)}. \tag{15}$$

This vector encapsulates both short-term attention and long-term memory; it is then passed to the hierarchical fuzzy inference and evidential reasoning layers. The TFMN is distinguishable from transformer-based temporal models. While transformers capture temporal dependencies through the self-attention mechanism, the proposed TFMN utilizes the combination of recurrent sequential memory with attention mechanism and fuzzy-state modeling. Using the explicit sequential bias in the form of the BiLSTM, the attention mechanism finds the most relevant temporal state information while the fuzzy memory ensures retention of a compact version of the past-state information. This architecture was specially designed to retain the temporal states in an interpretable form rather than using only self-attention in high-dimensional space.

3.5 Multiresolution fuzzy feature extraction

In NILM, different appliances exhibit characteristic behaviors at different time scales; a kettle exhibits very sharp and short transients, while a refrigerator shows periodic spikes, a washing machine has long-duration cycles with multiple stages, and a microwave exhibits mid-range fluctuations. Using a single convolution kernel size forces the model to see the data at only one temporal scale, causing loss of information. Therefore, DeepANFIS introduces multiresolution fuzzy feature extraction, where multiple convolutional pipelines process the same input signal using different temporal receptive fields. This mimics a multiscale fuzzy filter bank similar to wavelets but learned end-to-end. For each resolution level p, the input sequence is processed through the following layers:

- convolution layer-1
- non-linear activation (ReLU)
- convolution layer-2
- pooling layer

For each scale p, the features are extracted as shown in Equation 16:

$$F^{(p)} = pool(\varphi(Conv2_p(\varphi(Conv1_p(X))))), \tag{16}$$

where X is the input sequence or output from the previous layer; $\phi(\cdot)$ is the non-linear activation function called ReLU; $Conv1_p$ and $Conv2_p$ are convolution kernels with sizes corresponding to scale p.

After extracting features at each scale, they are fused as shown in Equation 17:

$$F_{fused} = concat_p(F^{(p)}). \tag{17}$$

Finally, a 1×1 convolution is performed to merge the features from all scales, as shown in Equation 18:

$$F_{merged} = Conv_{1*1}(F_{fused}). \tag{18}$$

The use of various convolution kernels with varying receptive fields allows DeepANFIS to learn the features at different time resolutions; shorter kernels learn high-frequency transients while longer kernels retain the long-duration cycles of the appliances, which are visible even upon coarse sampling. Therefore, using multiscale learning reduces the dependency on any single sampling resolution by retaining discriminative information at various time resolutions.

3.6 Evidential fuzzy uncertainty estimation

Traditional neural networks produce point estimates and cannot express confidence. The proposed DeepANFIS integrates EDL to model prediction uncertainty; here, evidential theory is used to assume that the model outputs evidence for each class rather than a probability.

The evidence values g_k are converted to Dirichlet parameters a_k as shown in Equation 19:

$$a_k = g_k + 1 \tag{19}$$

The total evidence is then computed as Equation 20:

$$S_k = \sum_{k=1}^{K} a_k \tag{20}$$

The probability for the predicted class is given by Equation 21:

$$\rho_k = \frac{a_k}{S} \tag{21}$$

The overall uncertainty is given by Equation 22:

$$\mu = \frac{K}{S}. \tag{22}$$

where K is the number of classes, a_k is the non-negative evidence (Dirichlet parameter) for class k, S is the total evidence (Dirichlet strength), and T is the total number of time steps in one sequence.

The confidence smoothing is then performed as shown in Equation 23:

$$\tilde{\rho_k} = \gamma\rho_k + (1-\gamma)\frac{1}{K}. \tag{23}$$

Although Bayesian neural networks and Monte–Carlo dropout provide uncertainty estimates based on stochastic forward propagation of the model, the chosen EDL technique provides uncertainty estimates by directly learning the parameters of a Dirichlet distribution during a single forward pass through the network. This ensures that the uncertainty estimation has a negligible effect on inference performance and allows the suggested method to be used in real-time NILM tasks. Although the evidential method yields an uncertainty estimate that can help determine any ambiguous predictions made under measurement noise or distributional shifts, the uncertainty estimation alone cannot be considered as being robust to arbitrary measurement noise. The present study focused on assessing the natural variability in publicly available datasets, and noise stress testing was beyond the scope of our experiments.

3.7 Task heads

DeepANFIS uses two specialized task heads to achieve power disaggregation: a regression head to predict continuous appliance power values and a classification head to determine appliance ON/OFF states. Both heads share the same fuzzy–temporal representation, allowing the system to learn richer features. The joint learning of regression and classification improves stability, enhances generalization, and enables the model to better understand the relationships between appliance-state transitions and corresponding power behaviors. If h_{task} is the shared fuzzy–temporal latent representation, the power regression output is given by Equation 24:

$$\hat{y}_{power} = R(h_{task}), \tag{24}$$

Where $R(.)$ is a fully connected neural network. The classification head is then given by Equation 25:

$$\hat{y}_{cls} = softmax(C(h_{task})), \tag{25}$$

where $C(.)$ is another fully connected layer.

3.8 Loss function

The model was optimized by developing a combined loss function integrating classification, regression, and uncertainty modeling. Cross-entropy was used to ensure that the appliance state was detected accurately, while mean-squared error was used to minimize the power prediction errors. An evidential loss term was thus added to enable the model to estimate uncertainty under noisy and ambiguous conditions while avoiding overconfident predictions. This combination yields a robust, reliable, and uncertainty-aware learning framework. The losses associated with classification, regression, and evidential uncertainty are given by Equations 26–28, respectively, while the total loss is given by Equation 29:

$$\mathcal{L}_{cls} = \frac{1}{B}\sum_{i=1}^{B} y_i \log\left(\tilde{y}_i\right), \tag{26}$$

$$\mathcal{L}_{reg} = \frac{1}{BT}\left(\sum_{i,t}\left(\hat{y}_{i,t}^{power} - y_{i,t}^{power}\right)\right)^2, \tag{27}$$

$$\mathcal{L}_{evid} = \frac{1}{B}\sum_{i,k} y_{i,k}\left(log S_i - log p_{i,k}\right), \tag{28}$$

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_r \mathcal{L}_{reg} + \lambda_e \mathcal{L}_{evid}. \tag{29}$$

As the entire objective function has now been optimized over the neural, recurrent, attention, fuzzy, and evidential parameters, the optimization process has become non-convex in nature. Hence, we cannot provide any formal guarantee regarding the global convergence property of the proposed method. The stability of the optimization process was empirically analyzed using training and validation trajectories, and the loss coefficients were selected empirically using preliminary validation experiments. The final values λ_r = 1.0 and λ_e = 0.01 provided the best compromise between regression accuracy and uncertainty calibration.

3.9 Performance metrics

The performance metrics for DeepANFIS are defined to enable evaluation of both the accuracy of event detection and accuracy of power estimation. In classification problems, the accuracy, precision, recall, and F1-score parameters are used to determine the appliance state correctly. In regression problems, the average percentage error between the result and actual power consumed is used to define the mean absolute error (MAE) metric. These performance evaluation metrics are described by Equations 30–34:

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N}, \tag{30}$$

$$Precision = \frac{T_P}{T_P + F_P}, \tag{31}$$

$$Recall = \frac{T_P}{T_P + F_N}, \tag{32}$$

These are parameters can be defined after (Equation 32) as where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative counts, respectively.

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall}, \tag{33}$$

$$MAE = \frac{1}{NT}\sum_{i,t}\left|\hat{y}_{i,t} - y_{i,t}\right|. \tag{34}$$

where MAE is the mean absolute error, N is the number of samples, T is the number of time steps, and $\hat{y}_{i,t}$ and $y_{i,t}$ are the predicted and actual power values, respectively.

4 Results and discussion

4.1 UK-DALE dataset

4.1.1 Overall performance comparison

The performance of the DeepANFIS model was evaluated under two distinct sampling regimes in the UK-DALE dataset, namely, high-frequency (6 s) and low-frequency (15 min) sampling. We performed three independent runs with different random seeds to ensure robustness. Table 1 summarizes the aggregate overall metrics for the UK-DALE dataset. The model parameters and values used are shown in Appendix A.

The results in Table 1 show that there is a tradeoff between recall and precision across different sampling frequencies. The data from the 6-s samples had an almost perfect recall score of 0.9805 but a moderate precision of 0.5646, indicating that almost all appliance events could be detected but included many false positives; the 15-min data showed an inverse relationship, where the high precision of 0.9691 and lower recall of 0.7368 indicated that the predictions were highly accurate but missed more true events. The 15-min sampling provided a better overall F1-score of 0.8371 against the F1-score of 0.7122 at 6-s sampling. Therefore, the 15-min sampling provided better harmonic means of precision and recall for the aggregated task.

The MAE of the 6-s data was lower (average value of 25.6 W) than that of the 15-min data (average value of 31.3 W); this can be interpreted to mean that the former sampling rate allowed better per-point power estimation. However, the root mean-squared error (RMSE) was better for the 15-min data (average value of 77.7 W) that for the 6-s data (average value of 87.4 W). The 15-min model showed a higher average prediction confidence score of 0.6649 against a score of 0.5885 for the 6-s model. The comparable stabilities of the results for the two significantly different sampling rates suggest that the proposed multiresolution representation effectively captures complementary information from the temporal dimension. Despite the fact that low-frequency sampling will smooth out transient events, the large receptive fields coupled with temporal memory enable capture of slow appliance dynamics.

4.1.2 Per-appliance performance analysis

The performances differ greatly among different appliances, emphasizing the distinct difficulty imposed by each appliance when being turned on or during operation. Table 2 highlights the F1-score and MAE of each appliance averaged over the seeds.

From the results in Table 2, we see that the refrigerator has a perfect F1-score at 6-s sampling and retains this outstanding performance with an F1-score of 0.999 at 15-min sampling. The constant and periodic power consumption pattern of the refrigerator can be simply learned by DeepANFIS based on the temporal memory part. The dishwasher shows excellent results for both

TABLE 1 Overall performance metrics across the UK-DALE dataset.

Dataset	Accuracy	Precision	Recall	F1-score	MAE (W)	RMSE (W)	Average confidence
UK-DALE (6 s)	0.8106 ± 0.0723	0.5646 ± 0.1085	0.9805 ± 0.0021	0.7122 ± 0.0911	25.63 ± 2.50	87.36 ± 5.59	0.5885 ± 0.084
UK-DALE (15 min)	0.8704 ± 0.0069	0.9691 ± 0.0040	0.7368 ± 0.0126	0.8371 ± 0.0096	30.94 ± 0.14	78.03 ± 0.44	0.6649 ± 0.001

TABLE 2 Per-appliance F1-score and MAE (mean across seeds).

Appliance	F1-score		MAE (W)	
	6-s sampling	15-min sampling	6-s sampling	15-min sampling
Kettle	**0.9998 ± 0.001**	0.0000 ± **0**	8.72 ± 0.82	19.50 ± 0.13
Microwave	0.5787 ± 0.424	0.0431 ± 0.058	40.91 ± 1.46	42.76 ± 0.32
Refrigerator	**1.0000 ± 0**	**0.9988 ± 0.001**	17.02 ± 1.40	36.92 ± 0.20
Dishwasher	**0.9368 ± 0.002**	**0.9865 ± 0.001**	7.68 ± 0.70	20.17 ± 0.25
Washer dryer	**0.9014 ± 0.008**	0.4174 ± 0.070	34.83 ± 0.61	34.01 ± 0.32
Washing machine	0.2021 ± 0.000	**0.8561 ± 0.011**	36.36 ± 15.07	26.62 ± 0.22

Results displayed in bold are better compared to others.

sampling cases, with the F1-score increasing to 0.987 from 0.937 at the lower sampling rate; the lengthy and multistep but distinct operation cycle of the dishwasher is duly recorded with very accurate parameters, especially by the 15-min model. Conversely, washing machines show better results at lower sampling rates, with the F1-score increasing from 0.202 to 0.856, which is good for the 15-min model but not excellent; the output of the 15-min model shows that it successfully identifies the presence of a washing machine with accurate aggregated parameters. Appliances such as kettles and microwave ovens have short-term heavy power consumption, which gives poor results at the 15-min sampling rate; their short-term event activations are smoothed or reduced in the 15-min aggregates and are nearly eradicated. The performance of the washer dryer declines heavily at low-frequency sampling, with the F1-score reducing from 0.901 to 0.417; here, the complex and time-varying cycles representing both the washing and drying operations of the washer dryer may become confused with the operations of other devices in the aggregates. The MAE is higher at the 15-min sampling rate for appliances such as kettle, refrigerator, and dishwasher; this result is expected since the detailed power consumption variations are averaged. The MAE outputs for the washer dryer and washing machine remain same or show minimal increments, authenticating the total energy consumption calculated by the model in each time frame as correct despite the poor event identification performance.

4.2 Performance analysis of the proposed DeepANFIS model for REDD

The results of the DeepANFIS model for REDD are presented in Tables 3, 4. The performance of DeepANFIS on REDD far exceeds that on the UK-DALE dataset, showing the outstanding capability of the model. Here, the model showed an outstanding mean F1-score of 0.9869 with a near-perfect accuracy of 0.9992, precision of 0.9869, and recall of 0.9871. The accuracy of power estimation was equally outstanding, with a low MAE of 2.56 W and RMSE of 20.19 W. These metrics indicate a dramatic performance improvement compared to the UK-DALE dataset, which had a best F1-score of 0.8371 at 15-min sampling. However, the significant error reductions for the REDD should be considered carefully. Some of these improvements may be attributed to the fact that REDD contains high-resolution information, clear appliance signatures, and lower household heterogeneity.

TABLE 3 Overall performance metrics for REDD.

Metric	Average value
Accuracy	0.9992 ± 0.0007
Precision	0.9869 ± 0.0221
Recall	0.9871 ± 0.0032
F1-score	0.9869 ± 0.0104
MAE (W)	2.5636 ± 0.8069
RMSE (W)	20.1874 ± 3.4945
Average confidence	0.678 ± 0.0331

TABLE 4 Appliance-level performance for REDD.

Appliance	F1-score	MAE (W)
Microwave	0.4203 ± 0.2848	1.030 ± 0.155
Refrigerator	0.9973 ± 0.0008	5.610 ± 2.378
Dishwasher	0.9639 ± 0.0127	1.047 ± 0.085

At the appliance level, there was a wide variation in performance and some interesting trends. The refrigerator continued to be a strongly performing appliance, with a near-perfect F1-score of 0.9973 and very low MAE of 5.61 W. This proves that the temporal memory design in DeepANFIS is highly effective at capturing

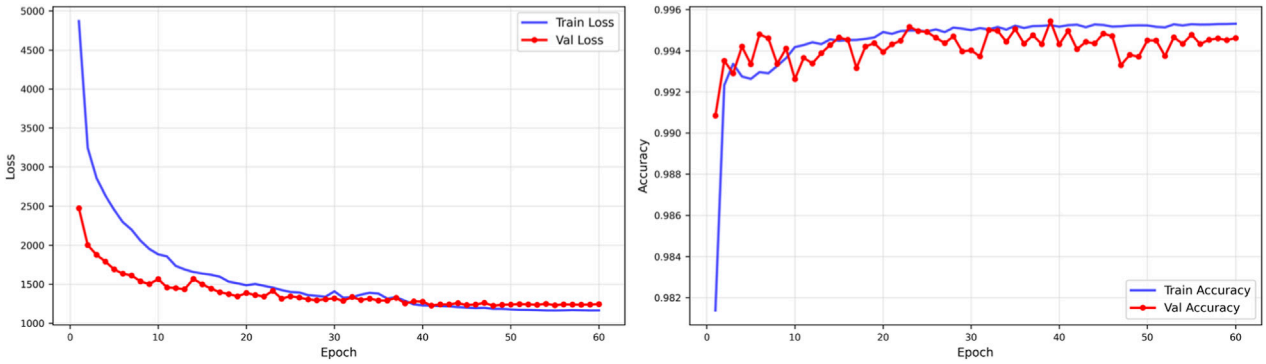


FIGURE 3
Training and validation losses (left) and accuracy curves (right) from the cross-dataset experiments.

the behaviors of periodic continuous appliances with predictable consumption patterns. Moreover, the dishwasher identification performance is good, with an F1-score of 0.9639 and very low estimation error with MAE of 1.05 W, confirming that the multifunctional cycles of this appliance can be distinctly identified in REDD. The microwave showed the most variable behavior and was difficult to model, with a greatly varying F1-score based on random seed initialization; the F1-score varied greatly for different random seeds, from a low of 0.2500 (for random seed 456) to a high of 0.7500 (for random seed 123). Even with this classification variability, all seeds produced very low power estimation errors for the microwave with an average MAE of 1.03 W, indicating that they could adequately estimate the power consumption when identified correctly. One of the major developments in the REDD experiments was the fixed evidential fuzzy uncertainty quantification module, which involved both evidence clipping and Kullback–Leibler regularization; this allowed a sensible average confidence of 0.678 for all seeds, showing that the model could successfully capture important confidence attributes in real-world settings. In addition to assessing whether there could be overfitting, we assessed the evolution of the training and validation curves during optimization. The convergence trends of the DeepANFIS model are presented in Figure 3. Here, the training error decreases quickly in the early epochs and then remains constant; meanwhile, the validation error also experiences a sharp decrease and then becomes stable at a certain point. Moreover, the accuracy curves for both training and validation are almost overlapped during the later epochs. These results do not show any signs of severe overfitting even though they cannot exclude the risk of dataset bias. Thus, the remaining performance differences between the UK-DALE dataset and REDD can be explained by the household and device differences.

4.3 Cross-dataset study

From an architectural point of view, cross-household robustness can be achieved through the combination of three different mechanisms. First, the adaptive Gaussian membership functions can dynamically adapt to changes in the distribution of appliance features. Second, the TFMN recognizes temporal patterns that are not sensitive to variation between households. Third, multiresolution convolutional feature extraction allows learning of the appliance features at different time resolutions rather than memorizing signal patterns specific to the households. The cross-dataset generalization experiments bring out both the strengths and limitations of DeepANFIS when evaluated in unseen environments. We trained the model on the UK-DALE dataset but tested it on the completely different REDD, which showed excellent classification generalization while suffering from significant estimation accuracy degradation.

Overall, the accuracy shown in Table 5 indicates a slight improvement from training to testing: the accuracy increased from 0.9954 to 0.9981, suggesting that the test set (REDD) may contain cleaner or more separable appliance signatures than the UK-DALE validation set. However, the F1-score was a more informative metric, revealing a moderate performance drop of 2.4%; this suggested that even though the model retained excellent overall classification ability, its performance in terms of balancing precision and recall degraded slightly in the new environment. The worst impact was observed on the power estimation accuracy, with the MAE increasing close to three times from 6.82 W for UK-DALE to 18.28 W for REDD. This substantial increase clearly indicated that while the model acquired the generic features for electrical device identification during the learning process, the general task of appliance on/off classification and the generic properties of electricity consumption were not transferred efficiently across datasets. This is a natural consequence as households in the United Kingdom and United States have distinct appliance models, usage patterns, electricity infrastructure, and measurement practices.

The results can vary greatly at the device level, providing information on the kinds of device signatures that work best for data transfer. At the device level, the dishwasher has an impeccable level of generalization in terms of the overall results (Table 6), with a near-perfect F1-score of 0.9929, an accuracy of 1.0000, and a low power consumption estimation error of 1.35 W. This implies that the operational signatures of the dishwasher are similar and distinct for various geographic locations and designs. This characteristic qualifies the dishwasher signatures as appropriate candidates for cross-validation tasks. Moreover, the results for refrigerators show

TABLE 5 Overall results for the cross-dataset study using the proposed model.

Metric	Training (UK-DALE)	Testing (REDD)	Performance change	Generalization ratio
Accuracy	0.9954 ± 0.0001	0.9981 ± 0.0003	−0.0027	1.003
F1-score	0.9938 ± 0.0001	0.9698 ± 0.0040	+0.0240	0.976
MAE (W)	6.82 ± 0.26	18.28 ± 0.44	+11.46 W	2.68
Confidence	-	0.5024 ± 0.0223	-	-

TABLE 6 Appliance-level performance for cross-dataset generalization.

Appliance	F1-score (testing)	Accuracy (testing)	MAE (W) (testing)
Microwave	0.2263 ± 0.0435	0.9945 ± 0.0009	1.38 ± 0.02
Refrigerator	0.9987 ± 0.0005	0.9998 ± 0.0001	52.10 ± 1.25
Dish washer	0.9929 ± 0.0122	1.0000 ± 0.0001	1.35 ± 0.06

a low level of generalization, with a near-perfect F1-score of 0.9987 for device identification but a poor estimate of 52.10 W for power measurement error. While this indicates that the model can accurately determine the operational phases of a device, the actual active and inactive periods of refrigerator operation in the United Kingdom differ greatly than those in the United States. The most difficult case of cross-dataset generalization was observed for the microwave whose F1-score reduced to 0.2263 for the test condition; this is significantly lower than the within-dataset UK-DALE score of 0.5787 at 6-s sampling rate and REDD appliance-level score of 0.4203. This large degradation suggests that the usage patterns, power signatures, or load interactions of the microwave are strongly specific to the dataset. Combined with the near-perfect test accuracy of 0.9945, this implies that the model defaults to predicting no microwave usage for most of the time steps and achieves a high accuracy simply because microwave usage is considered as sparse without identification of actual usage events. This forms a negative transfer where knowledge from the source domain actively affects performance on the target domain for some appliance classes.

To analyze the classification results for the cross-dataset validation, the confusion matrices achieved with the UK-DALE dataset for training and REDD for testing are shown in Figure 4. For the UK-DALE dataset, the model classified 487,001 OFF samples and 291,476 ON samples correctly; moreover, 41 OFF samples were incorrectly classified as ON, and 3,512 ON samples were wrongly classified as OFF. For the REDD, the model classified 36,606 OFF samples and 1,159 ON samples correctly; additionally, 63 OFF samples were wrongly classified as ON, and 8 ON samples were classified as OFF.

4.4 Ablation study

The results of the ablation study are shown in Figures 5, 6. The results reveal complex and somewhat counterintuitive interactions between the core components of DeepANFIS. Surprisingly, removing the hierarchical fuzzy system improved the overall performance metrics, with the F1-score increasing by 7.4% and accuracy improving by 6.3%. This suggests that the fuzzy logic component implemented in the system may be introducing constraints or suboptimal decision boundaries that could be overcome with a simpler multilayer perceptron scheme. From the results of the ablation study, it is clear that there exists a significant tradeoff between interpretability and performance; this is evident from the increased F1-score of approximately 7.4% for classification and accuracy improvement of about 6.3% upon removal of the hierarchical fuzzy layer, suggesting that the unconstrained neural layer offers flexibility in defining the decision boundary. Therefore, the fuzzy layer should not be seen as an element whose purpose is to enhance accuracy. The significance of the fuzzy layer is to enable fuzzy reasoning using adaptive membership functions, activation of rules, normalization of rule contributions, and TSK consequents.

Critically, when the temporal memory network was disconnected, we observed dramatic performance degradations in all classification performance measures; the F1-score decreased by 15.0% to 0.6054, and the accuracy declined by 12.7% to 0.7076. Such steep performance reductions highlight the core role of modeling the temporal aspects in NILM, where all appliance operations involve sequences with temporal dependencies. Appliances such as refrigerators with compressor operations, dishwashers with sequence operations, and washing machines with multiple phases demand memory of the previous states for accurate NILM. The observation that the recall metric is not significantly affected at 0.9779 in the absence of temporal memory indicates that the appliance activity can be detected accurately but the precision reduces to 0.4384, hinting that incorrect appliances are being assigned for power consumption variations in instances where the temporal information can be resolved. Curiously, both ablation settings showed slightly better power estimation performances than the complete model, with the MAE decreasing from 25.64 W to approximately 23.7 W and RMSE improving slightly in both cases. This counterintuitive outcome, where a simpler architecture provides better regression performance, points to a possible overparameterization of the complete DeepANFIS model on the power estimation problem or presence of problematic interactions among the model elements with regard to regression head optimization. The roles of the temporal memory elements and fuzzy systems are more critical in the classification problem; they can

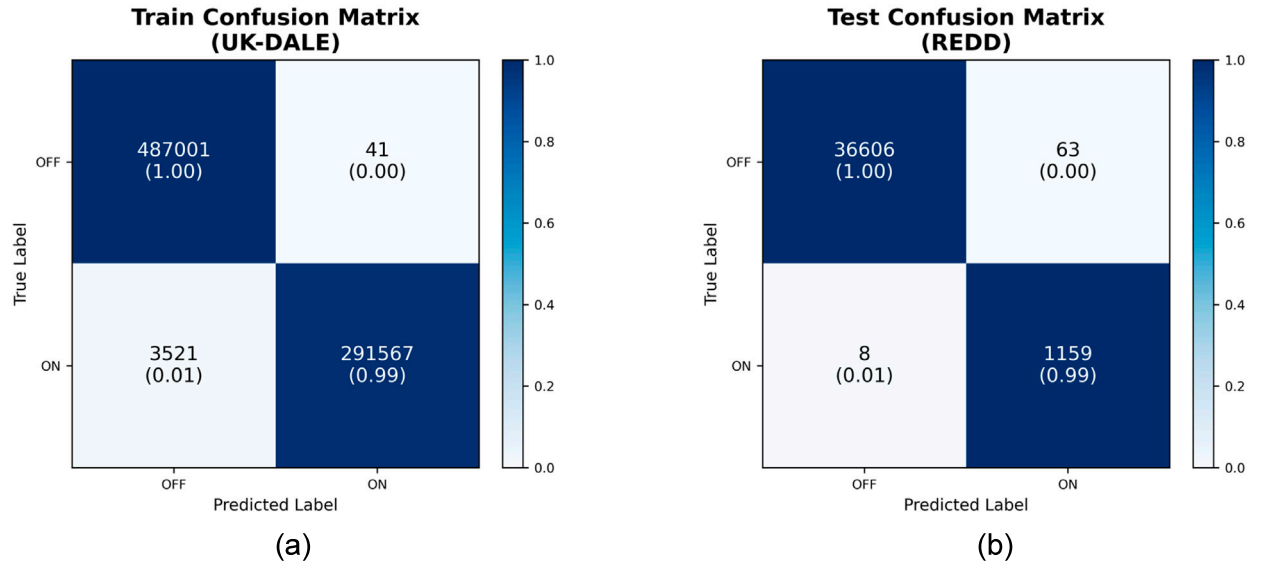


FIGURE 4
Cross-dataset confusion matrices for **(a)** training on the UK-DALE dataset and **(b)** testing on the REDD.

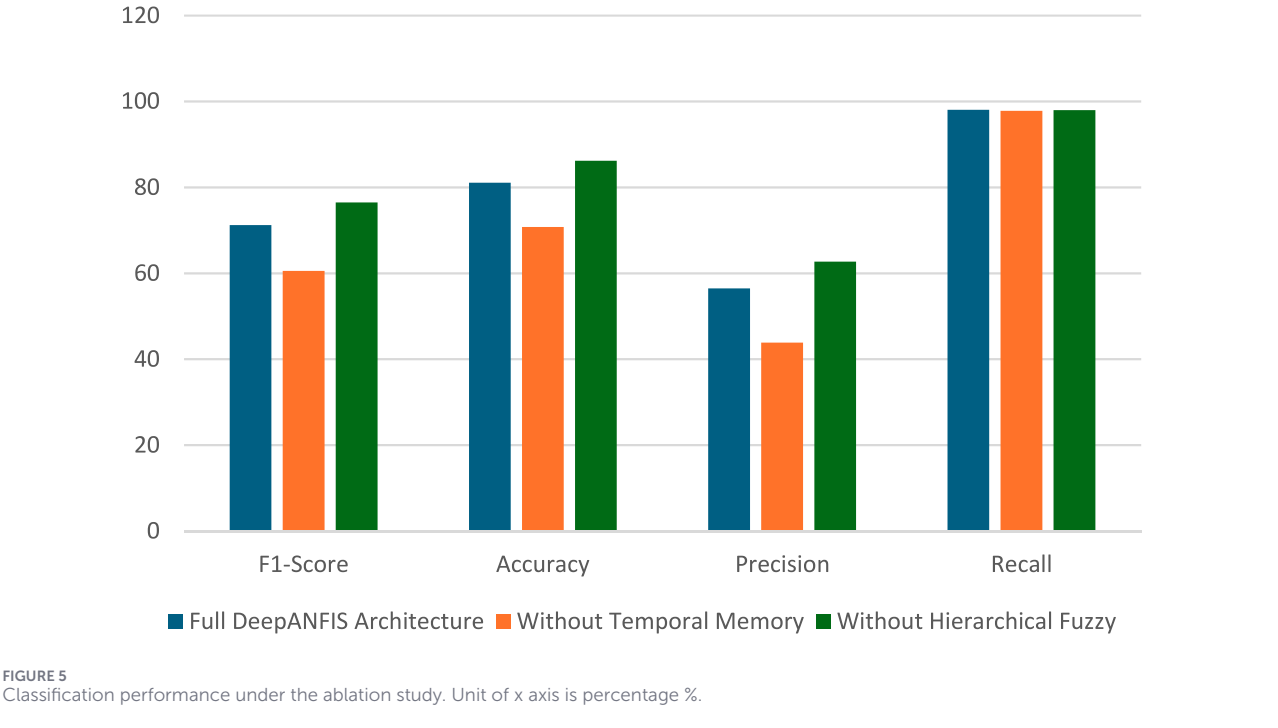


FIGURE 5
Classification performance under the ablation study. Unit of x axis is percentage %.

introduce noise or complexity, both of which impact the power prediction accuracy negatively.

4.5 Comparative analysis

To further validate the usefulness of the DeepANFIS architecture, a comparative analysis was carried out with respect to a traditional CNN-based NILM model for the same 6-s UK-DALE dataset. It is evident from Table 7 that DeepANFIS produces a higher average F1-score (0.7435) than the CNN (0.528), which is an improvement of about 40.8%. This is especially true for appliances such as the dishwasher, kettle, and refrigerator, which achieved F1-scores of 0.9368, 0.9998, and 1.0000, respectively. However, the CNN model produced a lower average MAE of 16.51 W than DeepANFIS (MAE of 22.14 W); the CNN also produced lower MAEs with regard to the power estimates for the washing machine and microwave. From these findings, we

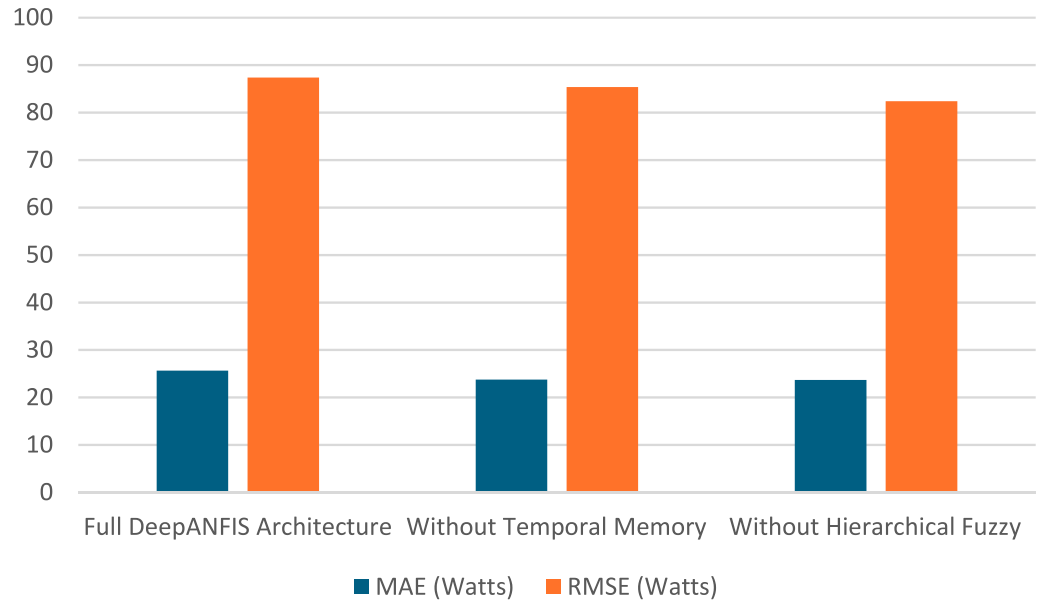


FIGURE 6
Power prediction performance under the ablation study. Unit of y axis is watts (w).

TABLE 7 **Comparison of CNN and DeepANFIS on the UK-DALE dataset at 6-s sampling rate.**

Appliance	F1-score		MAE (W)	
	CNN	DeepANFIS	CNN	DeepANFIS
Dishwasher	0.510	**0.9368**	25.43	**7.68**
Kettle	0.850	**0.9998**	9.64	**8.72**
Refrigerator	0.718	**1.0000**	29.20	**17.02**
Washing machine	0.173	**0.2021**	11.90	36.36
Microwave	0.341	**0.5787**	6.36	40.91

Results displayed in bold are better compared to others.

can conclude that the main strength of the DeepANFIS architecture over CNN is in discriminating the appliance state and not necessarily reducing the power estimation error.

5 Conclusion

We present a comprehensive neuro-fuzzy architecture incorporating multiresolution temporal feature learning, fuzzy hierarchical inference, temporal fuzzy memory, and evidential uncertainty estimation within a single optimization process for simultaneous appliance classification and power regression. The proposed framework is expected to offer better performance than a mere hybrid architecture involving CNN, BiLSTM, and ANFIS. Experiments performed on the UK-DALE dataset showed that DeepANFIS successfully balanced precision and recall across the sampling regimes, producing an F1-score of 0.7122 at 6-s resolution and up to 0.8371 at 15-min resolution. The framework was also found to be competitive in terms of the power estimation accuracy, with MAE values of 25.63 W for high-frequency sampling and 30.94 W for low-frequency sampling. On the REDD, we obtained exceptional performance, with an F1-score of 0.9869, an accuracy of 0.9992, and a very low MAE of 2.56 W. These results show the good fit of the proposed model when the appliance signatures are cleaner.

The results of the cross-dataset generalization experiments demonstrated the excellent appliance detection capability of DeepANFIS, with a limited 2.4% drop in the F1-score but an increase in the power estimation error from 6.82 W to 18.28 W due to the dataset-specific nature of appliance power characteristics. The ablation study showed that temporal memory was crucial for accurate predictive classification performance while the hierarchical fuzzy inference layer was responsible for providing interpretability. Interestingly, absence of the fuzzy inference layer enhanced the classification performance, indicating a tradeoff between predictive flexibility and interpretability. DeepANFIS illustrates that integrating a deep temporal model with fuzzy logic and evidential learning can substantially improve the reliability of an NILM system while providing good explainability. We expect that our future efforts will be aimed at robustly enhancing the cross-domain power estimation, reducing the architectural simplicity, and extending the framework to real-time and large-scale smart grid deployment.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding author.

Author contributions

MZA: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review and editing. MS: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review and editing. MMA: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review and editing. MW: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review and editing. FK: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review and editing. MG: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review and editing. UH: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review and editing.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Akbar, M. K., Amayri, M., Bouguila, N., Delinchant, B., and Wurtz, F. (2024a). Evaluation of regression models and Bayes-ensemble regressor technique for non-intrusive load monitoring. *Sustain. Energy Grids Netw.* 38, 101294. doi: 10.1016/j.segan.2024.101294

Akbar, M. K., Amayri, M., and Bouguila, N. (2024b). A novel non-intrusive load monitoring technique using semi-supervised deep learning framework for smart grid. *Build. Simul.* 17 (3), 441–457. doi: 10.1007/s12273-023-1074-5

Chen, T., Gao, J., Yuan, Y., Guo, S., and Yang, P. (2024). Non-intrusive load monitoring based on MoCo_v2, time series self-supervised learning. *Energy Build.* 317, 114374. doi: 10.1016/j.enbuild.2024.114374

Criado-Ramón, D., Ruiz, L. G. B., Iruela, J. R. S., and Pegalajar, M. C. (2024). A novel non-intrusive load monitoring algorithm for unsupervised disaggregation of household appliances. *Information* 15 (2), 87. doi: 10.3390/info15020087

Dai, S., Meng, F., Wang, Q., and Chen, X. (2024). DP2-NILM: a distributed and privacy-preserving framework for non-intrusive load monitoring. *Renew. Sustain. Energy Rev.* 191, 114091. doi: 10.1016/j.rser.2023.114091

Ding, D., Li, J., Wang, H., and Wang, K. (2024). Load recognition with few-shot transfer learning based on meta-learning and relational network in non-intrusive load monitoring. *IEEE Trans. Smart Grid* 15, 4861–4876. doi: 10.1109/TSG.2024.3390441

Esmaeili Shayan, M. (2024). Application of IoT and non-intrusive load monitoring techniques in microgrid energy management and monitoring systems. *Iran. J. Energy Environ.* 15 (3), 287–293. doi: 10.5829/ijee.2024.15.03.07

Gowrienanthan, B., Kiruthihan, N., Rathnayake, K. D. I. S., Kiruthikan, S., Logeeshan, V., Kumarawadu, S., et al. (2023). Deep learning based non-intrusive load monitoring for a three-phase system. *IEEE Access* 11, 49337–49349. doi: 10.1109/ACCESS.2023.3276475

Hart, G. W. (1992). Nonintrusive appliance load monitoring. *Proc. IEEE* 80 (12), 1870–1891. doi: 10.1109/5.192069

Kelly, J., and Knottenbelt, W. (2015). The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes. *Sci. Data* 2, 150007. doi: 10.1038/sdata.2015.7

Kolter, J., and Johnson, M. (2011). REDD: a public data set for energy disaggregation research. *Artif. Intell.* 25.

Kommey, B., Tamakloe, E., Kponyo, J. J., Tchao, E. T., Agbemenu, A. S., and Nunoo-Mensah, H. (2024). An artificial intelligence-based non-intrusive load monitoring of energy consumption in an electrical energy system using a modified K-nearest neighbour algorithm. *IET Smart Cities* 6, 132–155. doi: 10.1049/smc2.12075

Li, M., Tu, Z., Wang, J., Xu, P., and Wang, X. (2024). Dynamic time warping optimization-based non-intrusive load monitoring for multiple household appliances. *Int. J. Electr. Power Energy Syst.* 159, 110002. doi: 10.1016/j.ijepes.2024.110002

Liu, J., Wang, C., Xu, L., Wang, M., and Xu, Y. (2024). Enhancing residential electricity safety and management: a novel non-intrusive load monitoring-based methodology for accurate appliance operational state identification. *Appl. Sci.* 14 (2), 503. doi: 10.3390/app14020503

Owais, S., Gul, S., Akbar, R., Sajid, M. B., and Khalil, A. (2018). Energy consumption in residential sector of Pakistan. Available online at: https://www.researchgate.net/publication/342734707 (Accessed December 1, 2025).

Pu, Z., Huang, Y., Weng, M., Meng, Y., Zhao, Y., and He, G. (2024). Enhancing non-intrusive load monitoring with weather and calendar feature integration in DAE. *Front. Energy Res.* 12, 1361916. doi: 10.3389/fenrg.2024.1361916

Rafiq, H., Manandhar, P., Rodriguez-Ubinas, E., Ahmed Qureshi, O., and Palpanas, T. (2024). *A Review of Current Methods and Challenges of Advanced Deep Learning-Based Non-intrusive Load Monitoring (NILM) in Residential Context*. Elsevier Ltd. doi: 10.1016/j.enbuild.2024.113890

Shabbir, N., Vassiljeva, K., Nourollahi Hokmabad, H., Husev, O., Petlenkov, E., and Belikov, J. (2024). Comparative analysis of machine learning techniques for non-intrusive load monitoring. *Electronics* 13 (8), 1420. doi: 10.3390/electronics13081420

Todic, T., Stankovic, V., and Stankovic, L. (2023). An active learning framework for the low-frequency non-intrusive load monitoring problem. *Appl. Energy* 341, 121078. doi: 10.1016/j.apenergy.2023.121078

Varanasi, L. N. S., and Karri, S. P. K. (2024). Enhancing non-intrusive load monitoring with channel attention guided bi-directional temporal convolutional network for sequence-to-point learning. *Electr. Power Syst. Res.* 228, 110088. doi: 10.1016/j.epsr.2023.110088

Yao, L., Wang, J., and Zhao, C. (2024). Non-intrusive load monitoring based on multiscale attention mechanisms. *Energies (Basel)* 17 (8), 1944. doi: 10.3390/en17081944

Zhao, Q., Liu, W., Li, K., Wei, Y., and Han, Y. (2024). Unknown appliances detection for non-intrusive load monitoring based on vision transformer with an additional detection head. *Heliyon* 10 (9), e30666. doi: 10.1016/j.heliyon.2024.e30666

Zhou, K., Zhang, Z., and Lu, X. (2024). Non-intrusive load monitoring based on an efficient deep learning model with local feature extraction. *IEEE Trans. Ind. Inf.* 20, 9497–9507. doi: 10.1109/TII.2024.3383521

10.3389/fenrg.2026.1813402

Appendix A

Model parameter definitions and values

Symbol	Description	Value
B	Batch size	32
C	Number of input channels	1
T	Sequence length	599
M	Number of membership functions per input	3
R	Number of fuzzy rules	16
D	Input dimension	128
H	Number of hidden units in BiLSTM	64
T'	Output time steps after temporal processing	
Ms	Memory size	64
K	Number of output classes	2
λ_r	Regression loss weight	1.0
λ_e	Evidential loss weight	0.01
η	Learning rate	1e-3
Optimizer	Weight update method	AdamW
Activation	Non-linear function	ReLU, sigmoid
Membership type	Fuzzy membership function	Gaussian
Training/testing split		Training: 1, 2 Validation: 4 Testing: 5
Attention	Temporal feature enhancement	Multihead self-attention
Network depth	Number of hierarchical fuzzy layers	1
Rule type	Takagi–Sugeno–Kang (type-III) fuzzy inference	Implemented in hierarchical fuzzy layer