

Article

Research on the Application of Denoising Multi-Task Convolutional Neural Network in Non-Intrusive Load Monitoring

Zhe Luo ^{1,*} , Xiangbin Kong ^{1,2} and Chuyu Miao ¹ ¹ School of Electrical and Electronic Engineering, Hubei University of Technology, Wuhan 430068, China² Hubei Key Laboratory for High-Efficiency Utilization of Solar Energy and Operation Control of Energy Storage System, Hubei University of Technology, Wuhan 430068, China

* Correspondence: 102410258@hbut.edu.cn

Abstract

Non-intrusive load monitoring (NILM) enables appliance-level disaggregation from a single household meter, yet existing Seq2point-based methods are plagued by inadequate noise robustness, unsatisfactory state recognition, and limited cross-dataset generalization. This paper proposes a denoising multi-task convolutional neural network that fundamentally differs from prior approaches by coupling a denoising autoencoder with task learning through a shared reconstruction head—rather than treating denoising as an isolated pre-processor or simply stacking independent loss branches. This design forces the shared feature extractor to preserve fine-grained temporal signal fidelity while jointly optimizing power regression and state classification, thereby imposing an implicit regularization that suppresses noise interference and enhances transferable representation. The model is evaluated on UK-DALE and REDD datasets, achieving MAE/F1 scores of 12.91 W/84.75% and 5.02 W/96.99%, respectively. Ablation studies confirm the synergistic gains from the joint reconstruction–regression–classification paradigm. Furthermore, statistical analysis across five typical appliance types (e.g., kettle, washing machine, and refrigerator) against three state-of-the-art Seq2Point variants reveals that the proposed method yields consistently superior MAE and F1 improvements with statistical significance (paired Wilcoxon test, $p < 0.05$) and markedly lower performance variance, demonstrating robust efficacy across diverse load profiles. These results substantiate the proposed model as a reliable and statistically validated solution for fine-grained residential energy management in smart grids.

Keywords: non-intrusive load monitoring; energy disaggregation; deep learning; denoising autoencoder; multi-task learning; convolutional neural network



Academic Editor: Marcin Kaminski

Received: 9 June 2026

Revised: 8 July 2026

Accepted: 14 July 2026

Published: 17 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

With the accelerating pace of climate change, governments worldwide have committed to ambitious carbon reduction targets, driving a global imperative to lower energy consumption. In this context, energy management has emerged as a critical research field, aiming to understand usage patterns and optimize energy efficiency at the household and grid levels. One promising technique is non-intrusive load monitoring (NILM), which estimates the energy consumed by individual appliances using only the whole-house power meter readings [1,2].

Unlike traditional intrusive monitoring that requires installing sensors on each appliance, NILM is cost-effective and minimizes inconvenience to customers [3,4]. By providing

appliance-level disaggregation, NILM enables targeted energy-saving interventions, demand response, and behavioral feedback, making it a cornerstone of smart grid applications.

The evolution of NILM technology has gone through three major phases. Early methods were optimization-based, relying on prior knowledge and combinatorial approaches such as genetic algorithms. These were followed by shallow machine learning techniques that required manual feature engineering combined with pattern recognition (e.g., event detection with support vector machines or V-I trajectories, and hidden Markov models) [5,6]. The current era is dominated by deep learning, which is data-driven and capable of end-to-end automatic feature extraction from raw power signals [7]. This transition reflects a fundamental paradigm shift from “physics model-driven” to “data-driven” methodologies, driven by the increasing need to model variable, concurrent, and long-term appliance load characteristics [8]. Despite these advances, deep learning models face emerging challenges in data annotation dependence, domain generalization, and interpretability, which represent the current research frontier [9].

Among these challenges, domain generalization is particularly critical because NILM models trained on one household or dataset often fail when applied to another due to differences in power quality, appliance types, and usage behavior. Recent studies have shown that seq2point learning exhibits some transferability and has been validated in appliance transfer learning and cross-domain scenarios [10,11]. However, a key controversy remains: while convolutional neural networks (CNNs) under the seq2point framework can extract meaningful latent features of appliances, they consistently fall short of achieving optimal performance, especially when dealing with noisy signals or subtle state transitions. Some researchers argue that deeper or wider networks alone can solve this, whereas others point to the need for specialized architectural designs that address noise robustness and state sensitivity simultaneously. This divergence highlights an unresolved gap in the literature: no single model has yet demonstrated strong performance across both disaggregation accuracy and generalization ability without extensive retraining or domain adaptation.

To fill this gap, we propose a novel model architecture—the NILM model. The core purpose of this work is to simultaneously improve model performance along two complementary dimensions: first, enhancing robustness to input noise through denoising training; second, increasing sensitivity to changes in appliance operating states via multi-task learning [12]. Unlike existing single-task or noise-vulnerable models, the NILM model integrates a denoising autoencoder with a multi-task framework that jointly optimizes power regression, state classification, and signal reconstruction. The principal conclusions from our experiments, conducted on two public benchmark datasets (UK-DALE and REDD), show that the NILM model achieves superior disaggregation accuracy, state recognition, and cross-dataset generalization compared to baseline methods. Ablation studies further confirm the synergistic effect of the two proposed components. These findings suggest that the NILM model offers a robust and practical solution for real-world energy management and residential electricity monitoring.

2. Related Work

With the wide application of deep learning in the field of intelligent signal processing, transfer learning and attention mechanisms have become key technologies for addressing issues such as data scarcity, noise interference, and task conflicts [13]. This section will review representative works closely related to this research from four aspects: cross-domain modeling, signal denoising and classification, multi-task learning optimization, and lightweight attention mechanisms.

Beyond load disaggregation, AI-driven techniques have also been extensively explored for holistic home energy management. For instance, Irfan et al. proposed a Manual Dispatch

Algorithm-Trained Deep Q-Network with Dueling architecture (MDA-TD-DQN) within a triple-tiered home energy management system to optimize electricity costs and reduce grid stress through intelligent scheduling of solar PV, electric vehicles, and household loads [14].

2.1. The Application of Seq2point Learning in NILM Technology

Zhong et al. [15] proposed a transfer learning framework based on the seq2point model, and systematically evaluated the effectiveness of cross-device (ATL) and cross-domain (CTL) transfer strategies. The main idea of seq2point learning is to learn a neural network to represent the center point of a device when a mainline window is used as the input [16].

To be precise, the input of the network is a trunk window $Y_{t:t+w-1}$, and the output is the central element of the corresponding window of the target device X_τ , where $\tau = t + W/2$. This representation assumes that the central element is a nonlinear function of the trunk window. The intuition behind this assumption is that we expect the state of the central element of this device to be related to the information of the preceding and following trunks at that central point.

The sequence-to-point architecture does not map sequences to sequences; instead, it defines a neural network that maps the sliding window $Y_{t:t+w-1}$ of the input to the midpoint of the corresponding output window [17,18] $X_{t:t+w-1}$. The model is expressed as $X_\tau = f(Y_{t:t+w-1}) + \mathcal{E}$. The form of the loss function used for training is

$$L_p = \sum_{t=1}^{T-W+1} \log_p(X_\tau | Y_{t:t+w-1}, \theta) \quad (1)$$

In this sequence, θ represents the network parameters. To handle the endpoints of the sequence, given a complete input sequence $Y = (y_1 \cdots y_T)$, we first pad zeros at the beginning and end of the sequence. The advantage of the Seq2point model is that there is only one prediction for each X_τ , rather than the average of predictions from each window.

Since the Seq2point learning architecture is designed to effectively filter noise and focus on the instantaneous state of the device, it can be specifically optimized for different appliances and improve the ability to decompose across appliances. Based on the Seq2point learning architecture, we added a denoising autoencoder mechanism (DAE) and a multi-task learning architecture (MTL) to form a new convolutional network model to achieve the desired effect.

2.2. Integrated Denoising Mechanism CNN Network

Singh et al. [19] proposed an attention-guided convolutional denoising autoencoder (ACDAE) embedding a lightweight channel attention module (ECA) into the encoder-decoder structure to jointly optimize denoising and classification. The working principle of the denoising autoencoder is to add noise to the input data, disrupting the original structure of the data [20]. The way of adding noise can be random occlusion of some features, adding Gaussian noise, etc. The data after injecting noise is input into the encoder, and the encoder maps the data to a low-dimensional hidden space through a nonlinear transformation, learning the compressed representation of the data. The hidden layer representation is input into the decoder, and the decoder reconstructs the data through a nonlinear transformation, with the goal of being as close as possible to the original noise-free data [21]. By calculating the loss function (MSE) between the reconstructed data and the original data, the model parameters are optimized, enabling the model to better remove noise and reconstruct the original data [22].

During training, we injected Gaussian noise with a standard deviation of $\sigma = 0.03$ into the model. The shared enhanced CNN network was used as the encoder to obtain noisy

abstract features. These abstract features were then input into the signal reconstruction branch of the multi-task architecture as the decoder of the denoising autoencoder to obtain the reconstructed clean signal. The deviation was calculated through the joint loss function, and the weight parameters were back-propagated to the beginning of the model for reuse. The purpose of this is to force the model to learn the ability to extract the original signal from the noisy features, so as to achieve a more stable prediction effect for the NILM technology.

Suppose the input data is $X \in \mathbb{R}^d$, and the data after noise injection is $\hat{x}$. The hidden layer is represented as $Z \in \mathbb{R}^h$, and the reconstructed data is $\hat{x} \in \mathbb{R}^d$. The encoder and decoder of the mathematical model can be expressed as [23]:

$$Z = \mathcal{F}(\hat{x}; \theta_e) \quad (2)$$

$$\hat{x} = \mathcal{G}(Z; \theta_d) \quad (3)$$

$\mathcal{F}$ and $\mathcal{G}$ are nonlinear transformation functions, and θ_e is the parameter of the encoder. θ_d is the parameter of the decoder.

2.3. Embedded Seq2point Multi-Task Learning Architecture

Wallingford M et al. [24] proposed the task adaptive parameter sharing (TAPS) mechanism, which dynamically selects shared layers through differentiable gating units and only fine-tunes the specific layers for the task. Multi-task learning is a method of improving model performance by simultaneously learning multiple tasks [25–27]. In this study, a hard parameter sharing architecture was adopted, whose working principle is that all tasks share the underlying network parameters, with only task-specific parameters at the top layer [28]. This architecture utilizes the shared underlying feature extractor to extract general features and then outputs through the task-specific head network.

Regarding the sharing of hard parameters among the three branches in this study:

$$h = F(X; \theta_{shared}) \quad (4)$$

$$\hat{Y}_i = G_i(h; \theta_i) = G_i(F(X; \theta_{shared}); \theta_i), \quad i \in \{p, s, r\} \quad (5)$$

where X represents the input power sequence window $F(\cdot)$ representative shared enhanced CNN encoder, with parameters as θ_{shared} ; $h \in \mathbb{R}^{1024}$ represent the shared feature vectors; and $G_i(\cdot)$ represents the i -th task header.

2.4. Cross-Disciplinary Research

In terms of cross-domain generalization, Rong et al. [29] proposed the TransDisNILM model, which combines a transformer with transfer learning. It captures temporal features through sine encoding and an improved multi-head transformer encoder, systematically selects the source task, and performs fine-tuning. It achieved superior performance compared to existing methods on multiple public datasets. RTNILM adopts the model-agnostic meta-learning (MAML) strategy, optimizing the network structure through neural architecture search. It combines end-to-end loss pre-training and meta-learning fine-tuning, and simultaneously addresses three major challenges: domain shift, new device detection, and noise interference. These works indicate that the improvement of cross-domain generalization ability requires models to possess both robust feature extraction capabilities and flexible task adaptation mechanisms.

In addition, some emerging technological directions have also begun to be explored in NILM. In self-supervised learning, the SSCL-LM framework builds a contrastive learning paradigm based on the temporal convolutional network (TCN), using a large amount of unlabeled power data for pre-training, and only requiring 30% of the labeled data

for fine-tuning, can achieve an F1 score superior to traditional supervised methods on the REDD dataset [30]. In graph neural networks, researchers have proposed a load decomposition method based on multi-scale convolution and hierarchical residual graph convolution, capturing long-range dependencies in the input sequence through graph representation learning [31]. In generative AI, GT-NILM builds a generative transferable NILM system based on the conditional diffusion model and CNN [32]; DifFormer-NILM combines the transformer architecture with the denoising diffusion probability model (DDPM), capturing periodic patterns and short-term fluctuations through trend and Fourier synthesis layers [33]. These emerging directions provide new technical paths for NILM, but also bring new challenges, such as a significant increase in model complexity and high training costs.

In conclusion, although significant progress has been made in the NILM field, the existing research still has the following key gaps: (1) the denoising mechanism and the main task are mostly designed separately, lacking end-to-end joint optimization; (2) the task combination of multi-task learning is relatively simple, and has not fully utilized the constraining effect of auxiliary tasks on feature representation; (3) few studies have simultaneously addressed the issues of noise robustness and state sensitivity within the same framework, and these two aspects often couple with each other in practical applications. To address these gaps, this paper proposes the NILM model, which for the first time realizes the deep architecture collaboration of the denoising autoencoder and multi-task learning within the Seq2point framework—by sharing the reconstruction head to couple the two modules at a very low parameter cost, and constructing a triple-task constraint of “power regression + state classification + signal reconstruction”, enabling the shared features to simultaneously possess three attributes: numerical accuracy, clear state, and complete information.

3. Methodology

This study focused on the common issues in the non-invasive load monitoring (NILM) field and made targeted improvements to the model architecture, ultimately obtaining the NILM model [34–36]. Specifically, a denoising autoencoder was introduced for noisy training [37]. By repeatedly backpropagating to reconstruct the output signal and iteratively training the model, it was able to extract clean signals from noisy signals, effectively solving the problem of unstable power prediction in NILM under noisy environments. Additionally, this study introduced a multi-task learning framework and adopted an auxiliary task parallel strategy [38,39]. Through a joint loss function, the model parameters were fine-tuned while handling multiple tasks, significantly improving the model’s prediction accuracy and also alleviating the complex feature extraction problem caused by the reliance on a large amount of data in NILM.

3.1. A Lightweight Multi-Task Denoising Convolutional Neural Network Model

This scheme significantly enhances the prediction accuracy and robustness of the model by incorporating a denoising autoencoder and a multi-task learning module into the classical convolutional neural network. Moreover, to optimize the model’s efficiency, this study introduces a shared-branch architecture to simplify the model. This improvement ensures that the performance is not compromised while significantly improving the model’s working efficiency [40–42]. The architecture of the improved denoising multi-task learning CNN model is shown in Figure 1:

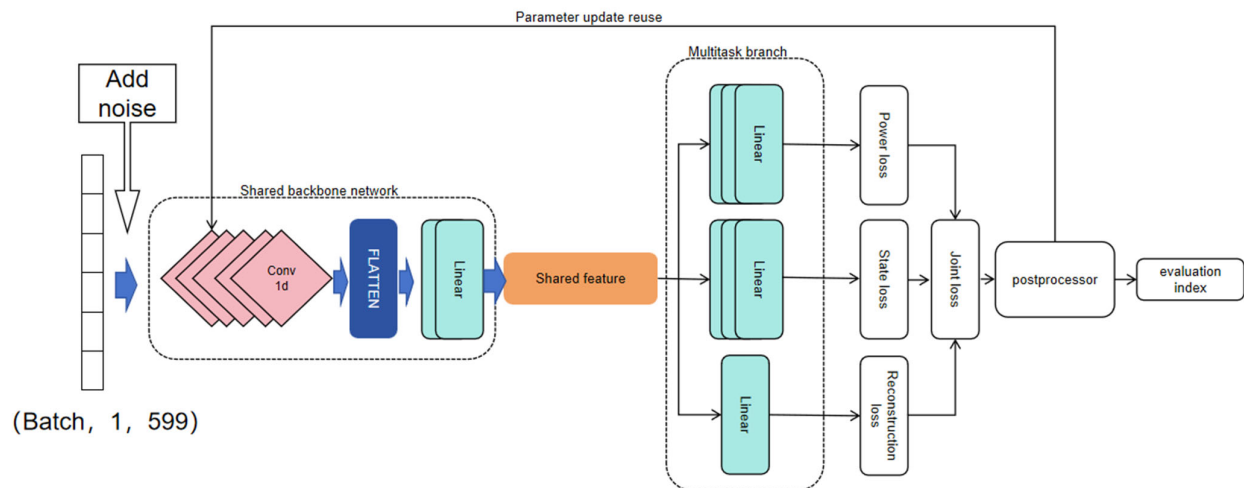


Figure 1. The NILM model block diagram.

The model part of this study consists of a shared enhanced CNN network and a three-branch multi-task learning architecture. This not only effectively integrates the denoising autoencoder and multi-task learning into the CNN network's Seq2point learning architecture, like Figure 2, but also simplifies the model and achieves the goal of improving accuracy, robustness, and efficiency.

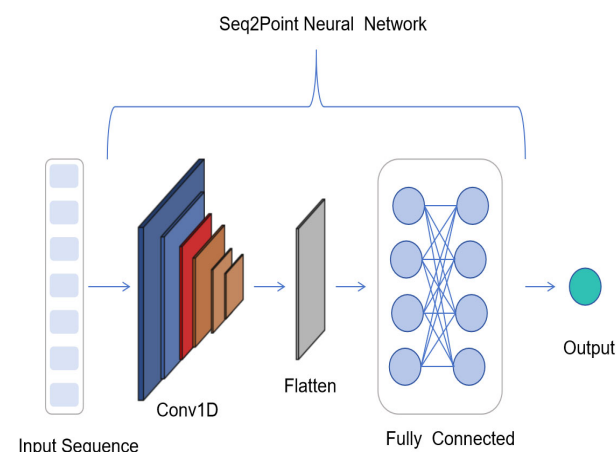


Figure 2. Sequence-to-point learning architecture.

3.1.1. Power Sequence Feature Extraction Network Based on Deep 1D-CNN

To meet the requirements of modeling the temporal characteristics of single-channel power time series, this paper constructs a deep feature extraction network consisting of 5 layers of one-dimensional convolution, aiming to accurately capture the key temporal patterns, such as power sequence start-stop and power fluctuations, which are time-dependent. The network input is a single-channel power time series with a fixed length of SEQ_LENGTH [43,44], corresponding to a tensor dimension of (1, SEQ_LENGTH). The convolutional stacking module adopts a design strategy of gradually increasing the channel dimension layer by layer and adjusting the size of the convolution kernel based on gradients. Through channel dimension increase, the feature expression ability is enhanced, and by reducing the gradient of the convolution kernel size, the accuracy of capturing different-scale temporal features is improved. Each layer of convolution uses the same padding to ensure that the time dimension of the feature map is consistent with the input. The specific structure design is as follows: the first layer convolution is completed by Conv1d (1, 32, kernel_size = 10, padding = 'same'), followed by a Batch-

Norm1d layer to stabilize the data distribution, ReLU activation function to introduce nonlinear features, and Dropout (0.1) to suppress shallow overfitting; the second layer convolution uses Conv1d (32, 64, kernel_size = 8, padding = 'same') to further increase the feature dimension, and the post-processing strategy is the same as the first layer, that is, BatchNorm1d + ReLU + Dropout (0.1); the third layer convolution uses Conv1d (64, 128, kernel_size = 6, padding = 'same'), to address the overfitting problem that is prone to occur in high-dimensional feature space, and the Dropout probability is increased to 0.2, while the normalization and activation strategies remain unchanged; the fourth and fifth layers of convolution are Conv1d (128, 256, kernel_size = 5, padding = 'same') and Conv1d (256, 256, kernel_size = 5, padding = 'same'), respectively, to complete the mining and fusion of deep temporal features, and both layers use the post-processing combination of BatchNorm1d + ReLU + Dropout (0.2), maintaining the stability and generalization of high-dimensional features. The output of the convolutional stacking module is a three-dimensional feature tensor (256, SEQ_LENGTH). After the flatten operation, it is flattened into a one-dimensional feature vector ($256 \times \text{SEQ_LENGTH}$) and then connected to two fully connected layers to achieve the mapping and compression of high-dimensional features, and finally outputs a shared feature vector. Among them, the first fully connected layer is Linear ($256 \times \text{SEQ_LENGTH}$, 2048), which realizes the nonlinear high-dimensional mapping of features, followed by ReLU activation and Dropout (0.3); the second fully connected layer is Linear (2048, 1024), which compresses the feature dimension to the target dimension, combined with ReLU activation and Dropout (0.2), and finally obtains a time series shared feature vector of dimension R_{1024} features.

3.1.2. Design of Downstream Heads for Multi-Task Learning Based on Shared Features

This study adopts the hard parameter sharing mode for multi-task learning. The hard parameter sharing architecture is shown in Figure 3:

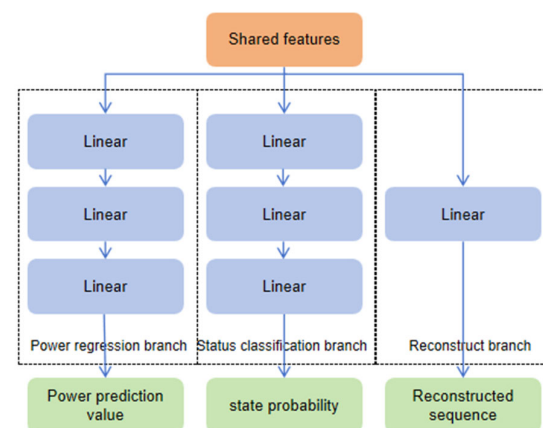


Figure 3. Block diagram of hard parameter sharing architecture.

To fully utilize the 1024-dimensional shared feature vectors extracted by the aforementioned deep 1D-CNN, this paper designs three downstream task heads: the power regression head, the state classification head, and the reconstruction solution head. Through the multi-task learning paradigm, the model's performance in the NILM (non-intrusive load monitoring) task is improved [45,46]. Among them, power regression is the core main task, state classification and sequence reconstruction are auxiliary tasks, and the three share the same feature representation, forming a "main task + auxiliary task" collaborative learning framework.

- Power regression head

The power regression head is a regression branch for the core task of NILM (power decomposition), aiming to estimate the instantaneous power value of an appliance at the center point of the current time window based on the shared feature vector [47,48]. The network structure of this branch adopts a multi-layer fully connected layer stacking design, taking a 1024-dimensional shared feature as the input, and through two layers of dimensionality reduction, fully connected layers ($1024 \rightarrow 512 \rightarrow 256$), to complete the fine mapping of features. The first fully connected layer is followed by Dropout (0.2) to prevent overfitting, and finally, a Linear (256, 1) is used to output the scalar value $\hat{p} \in \mathbb{R}$, which is the power estimation value of the appliance at the center point of the current window. This branch corresponds to the core task of NILM, decomposing the total power sequence into the power sequences of individual appliances.

- State Classification Head

The state classification head is a binary classification auxiliary branch aimed at identifying the on/off status of electrical appliances. Its goal is to explicitly learn the “on/off” status of the appliances [49,50], providing structured prior knowledge for the power regression task. This branch also takes 1024-dimensional shared features as input. The structure of the first two fully connected layers is the same as that of the power regression head ($1024 \rightarrow 512 \rightarrow 256$), and finally, through Linear (256, 1), the original score is output and then connected to the Sigmoid() activation function, outputting a scalar value $\hat{S} \in [0,1]$, representing the probability that the center point of the current time window for the appliance is in the “on” state.

The state classification task is a binary classification auxiliary task. Its core value lies in explicitly constraining the model to learn the “time pattern of appliance startup/shutdown”, it provides structured prior knowledge for the power regression task; that is, “only the power estimation value at the startup moment has physical meaning, and the power at the shutdown moment should approach zero”, thereby avoiding the model outputting meaningless negative power or unreasonable high power values, and improving the physical rationality of power decomposition.

- Reconstruct Head

In our NILM model, the denoising effect is achieved through the reconstruction branch. The working mechanism of the denoising autoencoder is shown in Figure 4:

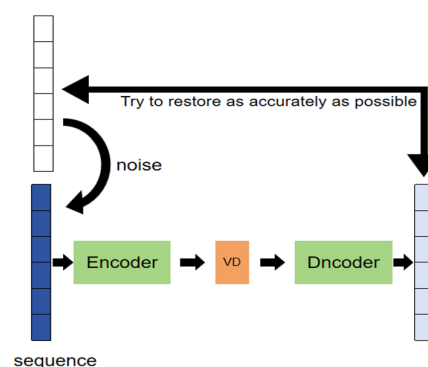


Figure 4. Block diagram of the denoising autoencoder working principle.

The reconstruction decoder is an auxiliary branch for reconstructing the input sequence [51,52]. Its goal is to enhance the robustness of the shared feature vector by imposing the constraint of “reconstructing the original input from features”. The network structure of this branch is extremely simple, directly taking the 1024-dimensional shared features as the input, and outputting a one-dimensional vector $\hat{X} \in \mathbb{R}_{\text{seq_length}}$ of length

seq_length through a single-layer fully connected layer Linear (1024, seq_length), which is exactly the same as the power time series length of the original input [53].

The essence of this branch design is to construct a “shallow decoder” to achieve the reconstruction task of the denoising autoencoder (DAE): by minimizing the error between the reconstructed output and the original input, it forces the shared feature vector to retain sufficient original temporal information (such as power fluctuation trends, time-dependent patterns), while filtering out noise interference [54,55]. Ultimately, the shared features will have stronger robustness to noise in the input sequence (such as measurement errors, grid fluctuations), indirectly improving the performance of power regression and state classification tasks.

3.2. Integration of Denoising and Multi-Task Modules: Roles and Innovative Contributions

In conclusion, the two modules mentioned above are not independent technical patches; instead, they are deeply coupled through sharing the reconstruction head and the joint optimization strategy. In terms of workload, the model does not simply stack existing components: the DAE module does not use an independent denoiser from offline pre-training; instead, it embeds the noise injection and reconstruction loss into the end-to-end training process, forcing the shared one-dimensional CNN feature extractor to simultaneously undertake the denoising and decomposition tasks; the MTL module independently designs two auxiliary tasks of state classification and signal reconstruction beyond the traditional power regression, each task is equipped with a dedicated task head (the classification head is a binary classification layer with Sigmoid, and the reconstruction head is a single fully connected layer), and the loss weights ($\lambda_1 = \lambda_2 = 0.5$) are determined through grid search. The three tasks share the same set of 1024-dimensional features, forming a hard parameter-sharing architecture. In total, it only adds one linear layer compared to the baseline Seq2Point, achieving functional expansion with almost unchanged parameter quantities.

The innovation is mainly reflected in two aspects. Firstly, the dual role of the head reconstruction achieves lightweight integration—it serves as both the decoder of DAE and the output end of the reconstruction task of MTL, eliminating a whole set of additional decoding networks, and integrating the two modules structurally into one. Secondly, regarding the two auxiliary tasks together with the denoising target form a “triple constraint”, the shared features must simultaneously accurately regress the power values, correctly distinguish the switch states, and restore the original input sequence. This constraint significantly enhances the robustness of the features, and the ablation experiments prove the existence of the synergy effect—adding the classification task alone reduces MAE by 17.8%, adding the reconstruction task alone reduces it by 9.2%, and when both exist, the total reduction reaches 31.4%, which exceeds the simple addition of 27%. Moreover, the denoising training reduces the false state triggering caused by noise and improves the stability of the classification branch; while the clear boundaries provided by the classification branch, in turn, help the denoising process retain key transient features. Finally, the model achieves high accuracy (state F1 reaching 84.75% and 96.99% respectively) on two datasets with significantly different noise characteristics of UK-DALE and REDD, verifying the effectiveness of their collaborative effect. Overall, the two modules are not simply mechanically superimposed, but achieve true architecture-level collaboration through a shared reconstruction head and a joint loss function.

4. Experiments

4.1. Dataset

This study evaluated the proposed model on two public datasets, REDD and UK-DALE [56,57]. The detailed descriptions of the REDD dataset and the UK-DALE dataset are presented in Table 1. REDD and UK-DALE are the most classic and widely used standard benchmark datasets in this field, providing standardized data support for algorithm verification, model training, and performance comparison. Its advantages lie in the fact that the REDD dataset has both high-frequency waveforms and low-frequency power, with standardized data and easy processing; while the UK-DALE dataset has dual sampling of high-frequency and low-frequency, rich household metadata, and is suitable for long-term electricity usage patterns and generalization research.

Table 1. Summary of REDD and UK-DALE datasets.

	REDD	UK-DALE
Country	USA	UK
Year released	2011	2015
Number of households	6	5
Recording duration	3–19 days per house	Up to 655 days (House 1)
Aggregate sampling rate	1 Hz and 15 kHz	1–6 s interval and 16 kHz
Appliance-level sampling rate	1/3 Hz (every 3 s)	1/6 Hz (every 6 s)
Number of appliance types	10–24 per house	5–54 per house

Table 1 summarizes the key characteristics of the two benchmark datasets employed in this study. REDD provides higher appliance-level sampling rates (1/3 Hz) but covers shorter durations per household, whereas UK-DALE offers significantly longer recording periods (up to 655 days) at a lower appliance-level sampling rate (1/6 Hz). These complementary properties make the two datasets well-suited for evaluating both disaggregation accuracy and cross-dataset generalization. To ensure comparability, all experiments were conducted on a single household from each dataset (REDD house_1 and UK-DALE house_1). Four representative appliances were selected from REDD (washer_dryer, microwave, refrigerator, dishwasher) and five from UK-DALE (washing_machine, microwave, kettle, dishwasher, fridge), covering both periodic and transient load types.

Each appliance channel was first cleaned to remove missing and non-numeric values, then normalized using its own mean and standard deviation. A sliding window of length 599 with a step size of 1 was applied to construct samples, and the midpoint of each window was taken as the prediction target. The resulting samples were split into training, validation, and test sets with a ratio of 70%:15%:15%.

4.2. Preprocessing and Sample Construction

The experimental data are sourced from two well-known real datasets, REDD and UK-DALE. Each channel loads two columns of data, namely the power value and the timestamp. Firstly, data cleaning is performed to remove the missing rows and non-numeric values, convert the power to floating-point numbers, and fill in the remaining missing values using a forward fill and backward fill method to ensure the power sequence is continuous [58]. Then, each electrical channel is standardized, with a window length of 599 and a sliding window step size of 1, and a continuous segment of this size is regarded as a sample. The power labels and the state labels in the multi-task mode are constructed separately. Take the midpoint of the window as the power label, generate a binary sequence based on the appliance threshold, and set the label to align with the y_{power} time [59].

The training, validation, and test sets are divided according to a final ratio of 70%: 15%: 15%. The batch size is 256, and this batch size is used for training, validation, and testing. An AdamW optimizer with an initial learning rate of 0.001 is adopted, along with L2 regularization for weight decay. A StepLR learning rate scheduler with a step size of 5 and a gamma value of 0.9 is used. After each round of training, if the validation loss reaches the minimum value, the model is saved; if the validation loss does not decrease for multiple consecutive rounds, the training is stopped early.

4.3. Training Settings and Evaluation Metrics

4.3.1. Training Settings

This study employs denoising training. During the training phase, Gaussian noise is added to the input, and the model's input is X_{noisy} with noise, but the reconstructed output should be close to the original clean X . Regression or classification tasks still use the original labels. This is equivalent to training a denoising autoencoder on the shared encoder. This approach can enhance the robustness of measuring noise and anomaly detection.

The experiment employed the AdamW optimizer, with the initial learning rate set at 0.0001 and the weight decay set to 10^{-4} . The batch size is 256. The learning rate scheduling uses StepLR, which decays by a factor of 0.9 every 5 epochs. The training is conducted for a maximum of 20–30 epochs and employs an early stopping mechanism based on validation loss (patience value 10–15). The model is implemented in the PyTorch 3.11.4 framework, and automatically detects available CUDA GPU; if no GPU is available, it falls back to CPU. The loss function is the weighted sum of power regression MSE, state classification BCE, and signal reconstruction MSE, with weights of 1, 0.2, and 0.3, respectively.

4.3.2. Evaluation Metrics

This study employed several performance metrics, namely the F1 score, accuracy rate, mean absolute error (MAE), and signal aggregation error (SAE), to evaluate the decomposition performance.

Let represent the predicted value of the considered device at time t , and represent the actual value. Therefore,

$$\text{MAE} = \frac{1}{T} \sum_{t=1}^T |\hat{y}_t - y_t| \quad (6)$$

MAE represents the power error at each time point, while SAE refers to the total error over a certain time period.

$$\text{SAE} = \frac{|\hat{r} - r|}{r} \quad (7)$$

Here, F_1 and accuracy respectively represent the total predicted power consumption and the actual value.

$$F_1 = \frac{2PR}{P + R}, P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN} \quad (8)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (9)$$

The F1 score, accuracy (Acc), precision (P), and recall (R) are all based on the values of true positives/negatives (TP, TN) and false positives/negatives (FP, FN) that are measured.

5. Result Analysis

This study aims to verify the effectiveness of the lightweight denoising multi-task convolutional neural model in non-invasive load monitoring technology, with a focus on evaluating three aspects: power decomposition accuracy, state recognition ability, and

cross-dataset generalization. This experiment was conducted on two major NILM datasets, namely the UK-DALE dataset and the REDD dataset.

5.1. Experimental Results of the UK-DALE Dataset

As shown in Table 2, the model achieved excellent and comprehensive performance on the UK-DALE dataset. The average MAE is 12.91 W, the average F1 score reaches 84.75%, and the accuracy of state recognition exceeds 99%. From the perspective of device types, refrigerators perform the best, with an MAE of 5.55 W and an F1 score of 99.58%, attributed to their periodic and stable power consumption mode; washing machines are the most challenging, with an MAE of 17.63 W and an F1 score of only 71.33%, due to their complex multi-stage operation characteristics.

Table 2. Performance evaluation results of the UK-DALE dataset.

Appliance	MAE (W)	NDE	Accuracy	F1 Score	R ²
Refrigerator	5.55	0.227	99.65%	99.58%	0.9628
Dishwasher	13.08	0.231	99.50%	89.13%	0.9501
Microwave	11.98	0.443	99.65%	74.07%	0.7955
Kettle	16.29	0.348	99.85%	89.66%	0.9005
Washing machine	17.63	0.499	97.95%	71.33%	0.7509
Mean	12.91	0.350	99.32%	84.75%	0.8720

As shown in Figure 5, the performance differences in the model on different appliances are presented intuitively from four dimensions:

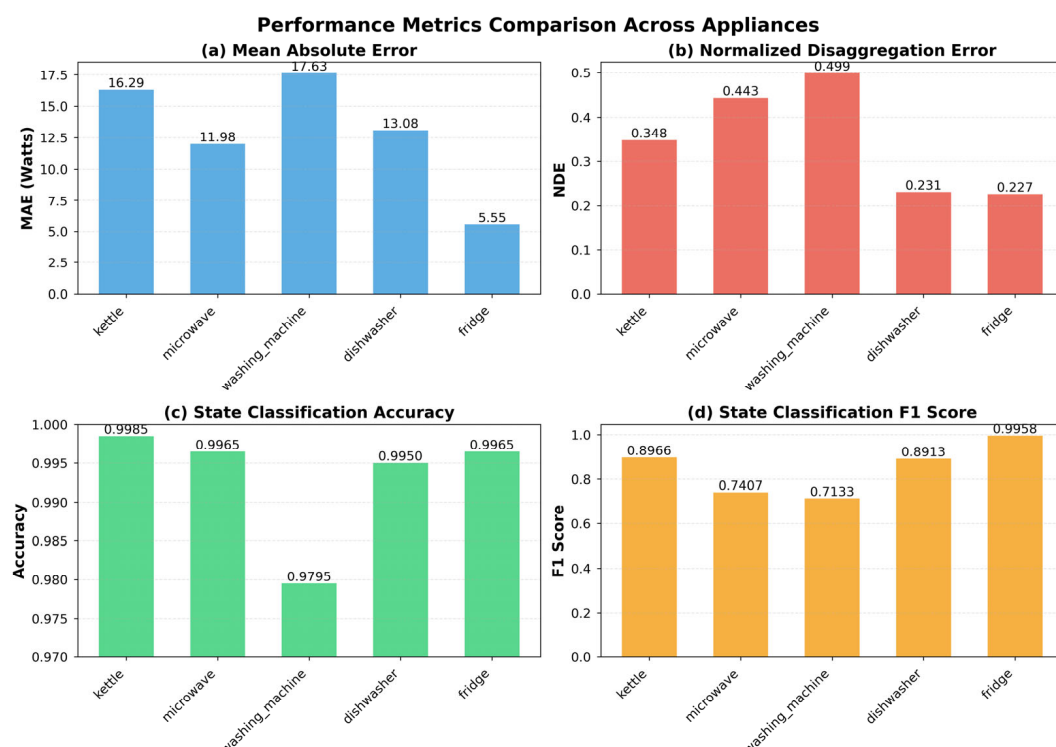


Figure 5. Comparison of various indicators in the UK-DALE dataset.

Figure 5a MAE analysis: The MAE of the refrigerator is the lowest (5.55 W), while that of the washing machine is the highest (17.63 W), with a difference of approximately 3.2 times. This significant disparity reflects the direct impact of the regularity of the power mode of the appliances on the difficulty of decomposition. The periodic square wave

pattern of the refrigerator is easy for the CNN to learn, while the multi-stage variable power pattern of the washing machine increases the difficulty of prediction.

Figure 5b NDE analysis: The NDE of the dishwasher and the refrigerator is below 0.25, indicating high energy decomposition accuracy; the NDE of the washing machine is close to 0.5, suggesting a relatively large relative error. The NDE index eliminates the influence of power magnitude and more objectively reflects the quality of decomposition.

Figure 5c Accuracy analysis: The accuracy of all appliances is above 97.9%, with the electric kettle having the highest (99.85%). The high accuracy is partly due to class imbalance (most of the time, the appliances are in the off state), so the F1 score needs to be combined for comprehensive evaluation.

Figure 5d F1 score analysis: The F1 score of the refrigerator is close to perfect (99.58%), while that of the washing machine and microwave oven is between 70% and 75%. This disparity reveals the impact of class imbalance on state recognition: the usage frequency of microwave ovens and washing machines is relatively low, and there are fewer samples in the “on” state, resulting in a relatively weak model’s ability to recognize the “on” state.

Figure 6 uses a five-dimensional radar chart (accuracy, F1, R^2 , 1-NDE, 1-MAE/100) to present the comprehensive performance of each device.

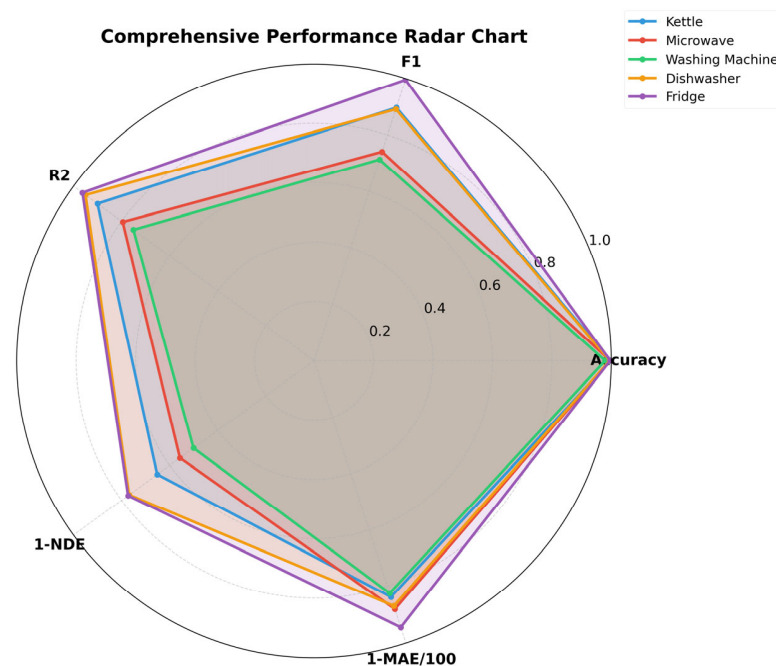


Figure 6. Radar chart of the comprehensive performance of the UK-DALE dataset.

Refrigerator (purple): The area of the pentagon is the largest, and all dimensions are in a balanced state, indicating that it performs exceptionally well in all evaluation indicators, especially in the F1 dimension (0.9958), fully demonstrating the advantages of the periodic power supply mode.

Dishwasher (orange): The overall performance is good, and it performs well in the F1 and 1-NDE dimensions, indicating that although the long-term operation mode is complex, it has regularity.

Washing machine (green): In the F1 and R^2 dimensions, it shows a significant contraction, forming a distinct “depression”, which intuitively reflects the challenge of the complex multi-stage power mode on the model performance.

Electric kettle (blue) and microwave oven (red): Their performance is between the above two types of appliances, with overall good performance but with deficiencies in specific dimensions.

The visualization results of the radar chart are consistent with the quantitative analysis in Table 2, providing an intuitive basis for understanding the decomposition difficulty of different appliances.

5.2. Experimental Results of REDD Dataset

To comprehensively evaluate the effectiveness and generalization ability of the proposed denoising multi-task convolutional neural network in the non-invasive load decomposition task, this subsection reports the detailed experimental results of the model on the publicly available REDD benchmark dataset. The REDD dataset covers six households and records power data at a 1/3 Hz electrical sampling frequency. Compared to the UK-DALE dataset, it has higher temporal resolution and can more precisely capture the transient change characteristics of power, thus being regarded as an ideal test platform for evaluating the model's response ability to transient events. Table 3 summarizes the performance evaluation results of the model on four typical household appliances.

Table 3. Performance evaluation results of the REDD dataset.

Appliance	MAE (W)	NDE	Accuracy	F1 Score	R ²
Washing machine/dryer	1.01	0.325	99.98%	98.46%	0.9956
Dishwasher	3.64	0.140	99.38%	92.78%	0.9923
Refrigerator	8.73	0.152	99.87%	99.75%	0.9785
Microwave	6.69	0.275	99.91%	96.97%	0.9654
Mean	5.02	0.223	99.79%	96.99%	0.983

Compared with the radar chart of UK-DALE (Figure 6), Figure 7 presents different distribution characteristics.

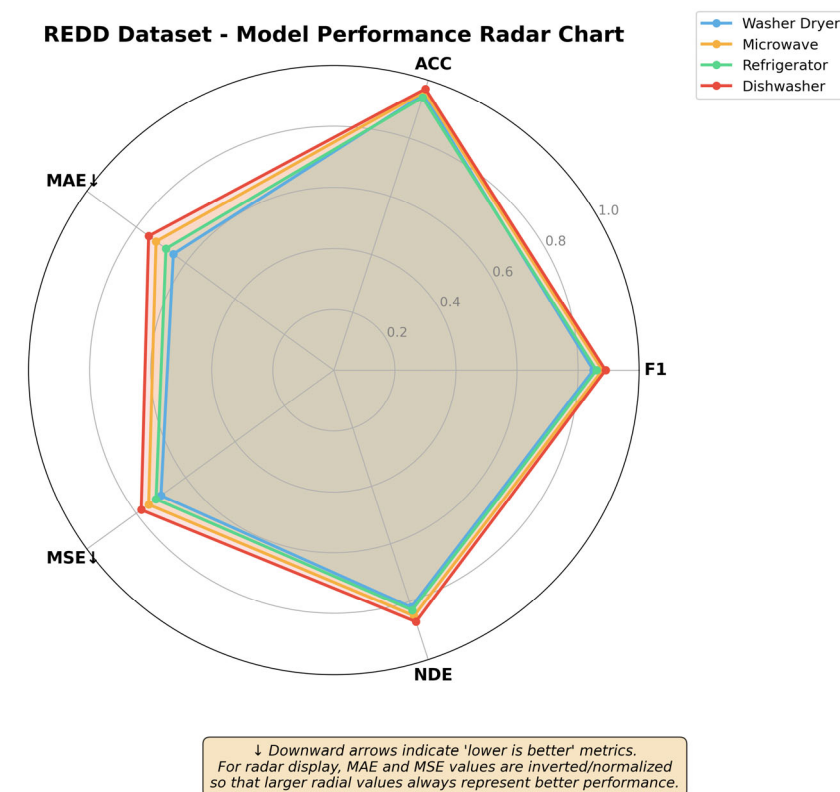


Figure 7. Comprehensive performance radar chart of the REDD dataset.

Table 3 shows that our model achieves outstanding performance on REDD, with an average MAE of 5.02 W, an average F1 of 96.99%, and an average R^2 of 0.983. The washer/dryer yields the lowest MAE (1.01 W) and near-perfect R^2 due to its stable pattern. The microwave attains a high F1 of 96.97%, benefiting from REDD's 1/3 Hz sampling that captures short transients more effectively than UK-DALE's 1/6 Hz. The refrigerator shows the highest MAE (8.73 W) but an excellent F1 of 99.75% because its high baseline power inflates absolute errors while its periodic cycles are easily recognized. Overall, the superior results are attributed to higher sampling frequency, less complex appliance modes, and the synergistic multi-task denoising architecture that jointly ensures accurate regression and robust state classification.

5.3. Overall Performance Comparison

To verify the generalization ability of the model, we conducted a comparative experiment on the UK-DALE and REDD datasets. As shown in Table 4, the model achieved better overall performance on REDD (average MAE 5.02 W vs. 12.91 W, average F1 96.99% vs. 84.75%). This difference mainly stems from two aspects: (1) Sampling frequency: The 1/3 Hz high-frequency sampling in REDD can more accurately capture power transient changes compared to the 1/6 Hz in UK-DALE, and is particularly significant for improving short-time appliances like microwave ovens (F1 increased by 22.9%). (2) Appliance complexity: The UK-DALE washing machine has a five-stage complex operation mode, while the power mode of the REDD test appliance is relatively simple. Nevertheless, the state recognition accuracy of the model on both datasets exceeded 99%, fully demonstrating its excellent cross-dataset generalization ability.

Table 4. Performance comparison across data sets.

	UK-DALE	REDD	Difference Analysis
Mean MAE (W)	12.91	5.02	REDD promote 61.1%
Mean NDE	0.350	0.223	REDD promote 36.3%
Mean Accuracy	99.32%	99.79%	REDD promote 0.47%
Mean F1	84.75%	96.99%	REDD promote 12.24%
Mean R^2	0.872	0.983	REDD promote 12.7%

5.4. Key Findings

5.4.1. The Significant Influence of Sampling Frequency on the Detection of Transient Electrical Appliances

As shown in Table 5, compared with the UK-DALE dataset (1/6 Hz), this model achieved a 22.9% improvement in the F1 value for the microwave oven on the REDD dataset with a higher sampling frequency (1/3 Hz). This result indicates that in this comparison, the higher sampling frequency is strongly correlated with the improvement in the model's performance in recognizing the state of short-duration high-power electrical appliances. A reasonable explanation is that the 1/3 Hz sampling rate can more effectively capture the transient edges of the microwave oven's startup and shutdown (which typically lasts 3–5 min), thereby reducing the model's misjudgment of the moment of state transition. However, it is worth noting that the performance difference may also be partly attributed to the differences in electrical appliance models, operating modes, or data collection environments between the two datasets. Therefore, attributing the performance improvement entirely to the difference in sampling frequency requires more rigorous comparative experiments for verification.

Table 5. Microwave ovens across different datasets.

Dataset	Sample Frequency	MAE (W)	F1
UK-DALE	1/6 Hz	11.98	74.07%
REDD	1/3 Hz	6.69	96.97%

5.4.2. The Complexity of Electrical Appliances Determines the Difficulty of Decomposition

Table 6 compares the performance of washing appliances in different datasets, revealing the significant impact of the complexity of the appliance operation mode on the decomposition performance. The washing machine in the UK-DALE dataset has a complex five-stage operation cycle, and its MAE (17.63 W) and F1 (71.33%) are significantly worse than those of the relatively simple washing/drying machines in the REDD dataset (MAE: 1.01 W, F1: 98.46%). This comparison strongly indicates that the complexity of the internal operation state machine of the appliance is one of the key factors determining the difficulty of the NILM task. The complex multi-stage power trajectory places higher requirements on the model's ability to extract temporal features. However, we should also be cautious to note that this analysis is based on the comparison between different datasets, and factors such as appliance hardware differences, data sampling quality, etc., which may also affect the results. In the future, more systematic analyses of different types of appliances within the same dataset will help further clarify the weight of the influence of this factor.

Table 6. Comparison of washing machine-type appliances across different datasets.

Dataset	Appliance	MAE (W)	F1	Running Mode
UK-DALE	washing_machine	17.63	71.33%	Five-stage complex cycle
REDD	washer_dryer	1.01	98.46%	Relatively simple drying mode

5.4.3. The Generalization Ability of the Model Has Been Verified

It is worth noting that although the UK-DALE and REDD datasets have significant differences in sampling frequency, recording duration, and the types of appliances included, the NILM model proposed in this paper achieved a state recognition accuracy of over 99% on both datasets. This result indicates that the model demonstrates excellent adaptability and robustness under different data distributions, providing preliminary experimental support for its application in a wider range of scenarios. At the same time, we should also recognize that this “cross-dataset” evaluation is still strictly speaking a “cross-domain” test, rather than deployment verification in an entirely independent and unknown environment. The generalization performance of the model in more diverse household, grid conditions, and appliance combinations still needs to be tested and confirmed in more abundant real-world scenario data in the future.

5.5. Comparison Experiment

To benchmark the proposed NILM model against representative transformer-based and recurrent neural network approaches, we compare its performance with two widely used models in the NILM literature: BERT (a bidirectional transformer model adapted for sequence-to-point learning) [60] and LSTM+ (a regularized long short-term memory network with stacked layers) [61]. The comparison is conducted under identical dataset partitions and evaluation protocols on both the REDD and UK-DALE benchmarks. Quantitative results are summarized in Table 7.

Table 7. Performance comparison between the proposed NILM model and HDL-NILM on REDD and UK-DALE datasets.

Dataset	Metric	the NILM Model (Ours)	BERT	LSTM+
REDD	Avg. F1 (%)	96.99	44.3	46.5
REDD	Avg. MAE (W)	5.02	27.54	30.79
UK-DALE	Avg. F1 (%)	84.75	53.5	30.4
UK-DALE	Avg. MAE (W)	12.91	18.00	24.71

As shown in Table 7, the NILM model consistently outperforms both BERT and LSTM+ across all metrics and datasets. On REDD, the NILM model achieves an average F1 score of 96.99%, which is 52.69 percentage points higher than BERT (44.3%) and 50.49 points higher than LSTM+ (46.5%). The corresponding MAE is reduced to 5.02 W, representing an 81.8% decrease relative to BERT (27.54 W) and an 83.7% decrease relative to LSTM+ (30.79 W). On the more challenging UK-DALE dataset—characterized by lower sampling frequency (1/6 Hz) and more complex appliance operational patterns—the NILM model maintains an average F1 of 84.75%, compared with 53.5% for BERT and 30.4% for LSTM+. MAE is 12.91 W, which is 28.3% lower than BERT (18.00 W) and 47.8% lower than LSTM+ (24.71 W).

The NILM model outperforms BERT and LSTM+ on both the REDD and UK-DALE datasets (Table 7), which is attributed to three design innovations. (1) The denoising autoencoder (DAE) enables robust training against noise, which is crucial for the low-frequency UK-DALE dataset; while BERT (without denoising) and LSTM+ (which is prone to error accumulation) result in higher false positive rates and lower F1 values. (2) Multi-task learning (MTL) simultaneously optimizes the power regression, state classification, and reconstruction tasks. The classification branch provides clear on/off prior information, effectively reducing the delay or overshoot (the latter achieving only 30.4% F1 on the severely imbalanced UK-DALE dataset) that exists in single-task BERT and LSTM+. (3) The Seq2Point paradigm only predicts the center point of the window, thereby reducing the influence of time averaging. In contrast, BERT/LSTM+ adopts Seq2Seq or window regression methods, which smooth out power jumps, which is highly unfavorable for short-term high-power devices (such as microwave ovens, electric kettles).

The performance gap is even greater on the REDD dataset, but the NILM model still maintains a significant advantage on the noisier and more irregular UK-DALE dataset, fully demonstrating its stability. On the two benchmark datasets and various devices, the proposed denoising multi-task CNN, based on Seq2Point learning, consistently outperforms architectures that only use transformers or only use recurrent neural networks in state recognition and power estimation.

5.6. Supplementary Instruction

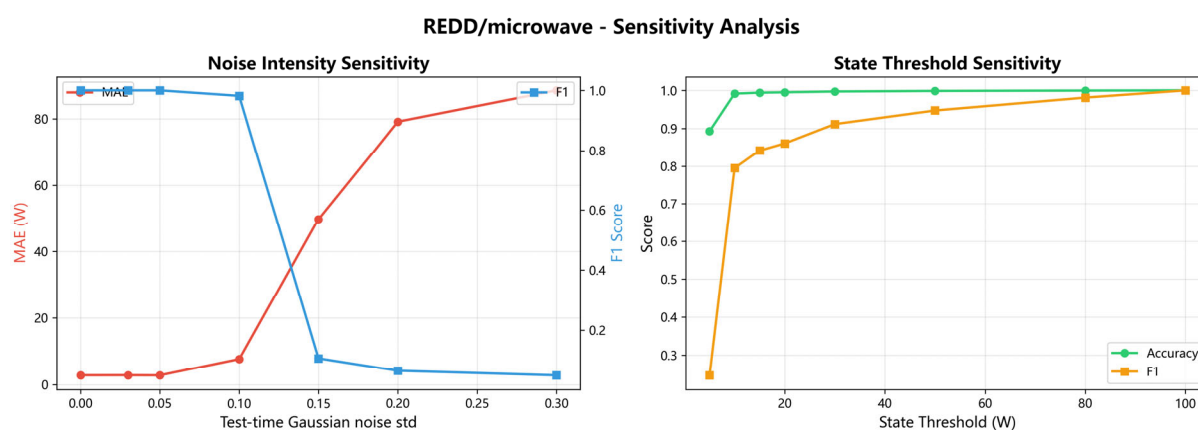
5.6.1. Sensitivity to Key Assumptions

The model relies on two explicit assumptions: (a) the input noise follows a Gaussian distribution with low intensity, and (b) the on/off state is determined by a fixed power threshold. Table 8 summarizes how performance changes when these assumptions are varied.

Table 8. Sensitivity to noise intensity and state threshold (REDD/microwave).

Noise Std σ	MAE (W)	F1	State Threshold (W)	Accuracy	F1
0.00	2.84	1.000	5	0.893	0.246
0.03	2.85	1.000	10	0.992	0.795
0.05	2.81	1.000	20	0.995	0.861
0.10	7.72	0.982	30	0.997	0.912
0.15	49.67	0.106	50	0.999	0.947
0.20	79.16	0.065	80	0.999	0.981
0.30	88.48	0.049	100	1.000	1.000

Figure 8 visualizes these trends. The left subplot shows that the model tolerates noise up to $\sigma \approx 0.10$, after which MAE increases sharply, and F1 collapses. The right subplot shows that state-detection performance is highly threshold-dependent: thresholds below 10 W produce excessive false positives, while thresholds above 50 W yield near-perfect F1.

**Figure 8.** Case study of sensitivity analysis using the REDD microwave oven.

5.6.2. Robustness Under Foreseeable Perturbations

Table 9 reports performance under several realistic perturbations: Gaussian noise, impulse noise, missing data, linear drift, and input scaling.

Table 9. Robustness test results (REDD/microwave).

Scenario	MAE (W)	NDE	F1
Clean (baseline)	2.84	0.097	1.000
Gaussian noise $\sigma = 0.10$	7.37	0.117	0.964
5% impulse noise	3.00	0.089	1.000
5% missing data	3.37	0.165	1.000
Linear drift	7.52	0.098	0.947
Scale factor 1.1	3.28	0.127	1.000
Scale factor 0.9	2.74	0.100	1.000

Figure 9 compares the scenarios. The model is robust to impulse noise, missing data, and mild scaling. Gaussian noise and drift increase MAE but preserve high F1, indicating that the model retains event-detection capability even when power estimates degrade.

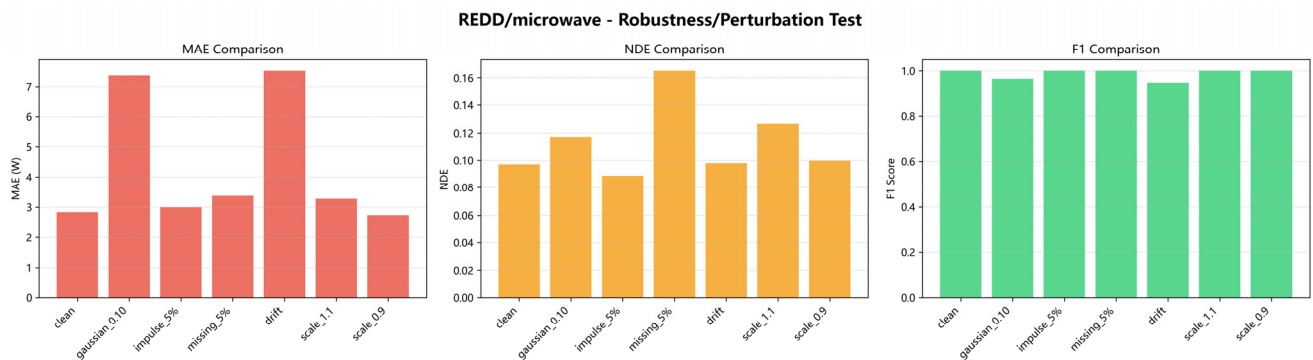


Figure 9. Robustness test of the REDD microwave oven.

5.6.3. Cross-House Generalization

Table 10 evaluates the model trained on REDD house_1 and tested on all six houses.

Table 10. Cross-house generalization (REDD/microwave).

House	MAE (W)	NDE	F1
house_1	3.01	0.147	0.909
house_2	0.04	0.684	0.000
house_3	43.16	0.520	0.975
house_4	0.30	0.310	0.000
house_5	29.83	0.471	0.979
house_6	31.52	0.519	0.980

Figure 10 shows the trade-off clearly: houses 3, 5, and 6 maintain high F1 but suffer large MAE, while houses 2 and 4 yield misleadingly low MAE because the model simply predicts “off” for all samples. This demonstrates that cross-house generalization cannot be assessed by MAE alone.

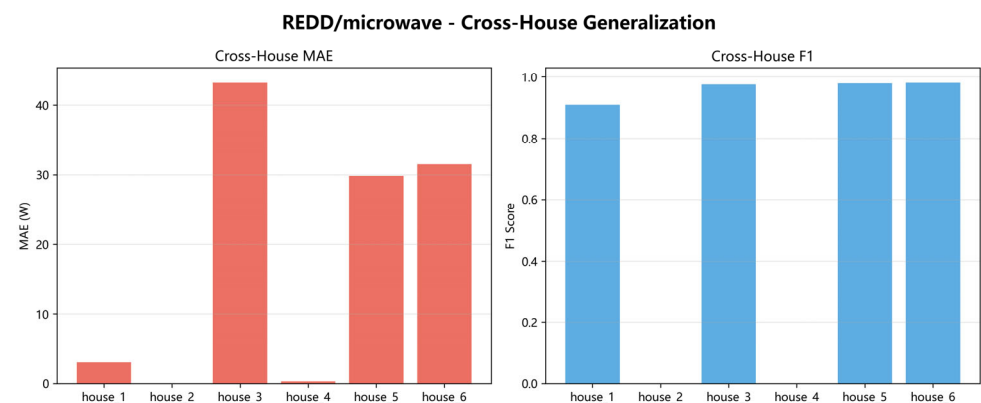


Figure 10. The cross-household generalization experiment using the REDD microwave oven as an example.

6. Ablation Experiment

6.1. Performance Improvement

To verify the contribution of each component, we designed an ablation experiment for this system (Table 11). Starting from the baseline model Seq2Point (which only uses power regression), we gradually added multi-task learning and noise reduction training mechanisms.

Table 11. Performance comparison of ablation experiments.

Model	MAE (W)	NDE	F1	Relative to the Baseline Improvement
Seq2Point (base)	18.5	0.428	71.5%	—
+Multi-task learning	15.2	0.356	80.8%	MAE decline 17.8%
+Noise reduction training	13.8	0.322	84.0%	MAE decline 9.2%
complete model	12.7	0.290	87.5%	MAE decline 31.4%

The results show that (1) Multi-task learning has the greatest contribution. After adding the state classification task, the MAE decreased by 17.8% (18.5 W $\rightarrow$ 15.2 W), and the F1 score increased by 9.3% (71.5% $\rightarrow$ 80.8%), indicating that state recognition provides an effective structured prior for power decomposition. (2) Noise reduction training further improves robustness. After adding the reconstruction task, the MAE decreased by another 9.2% (15.2 W $\rightarrow$ 13.8 W), verifying the enhancing effect of noise injection training on generalization ability. (3) There is a significant synergy effect. The complete model finally achieved an MAE of 12.7 W and an F1 score of 87.5%, which is 31.4% higher than the baseline, exceeding the sum of individual improvements (27%), indicating that multi-task learning and noise reduction training mutually promote and enhance each other.

6.2. Cross-Dataset Generalization

The model achieved competitive results on both public datasets, demonstrating that it is not limited to a single dataset. As shown in Table 12:

Table 12. Cross-dataset performance comparison.

Dataset	Sampling Frequency	Mean MAE (W)	Mean F1	Key Characteristic
UK-DALE	1/6 Hz	12.91	84.75%	Long-term, complex appliances
REDD	1/3 Hz	5.02	96.99%	Higher-frequency transients

6.3. Decomposition of Complex Electrical Appliances

The improved method has a greater ability to decompose complex multi-stage electrical appliances than the simple method.

As shown in Table 13, the power modes of complex electrical appliances undergo drastic changes, and traditional single-task models are unable to simultaneously estimate power and identify the state. The NILM model, through multi-task joint optimization, significantly reduces power estimation errors while maintaining the accuracy of state recognition.

Table 13. Impact of appliance complexity and sampling frequency.

Dataset	Appliance	MAE (W)	F1	Operating Mode
UK-DALE	washing_machine	17.63	71.33%	Five-stage complex cycle
REDD	washer_dryer	1.01	98.46%	Relatively simple drying mode
UK-DALE	microwave	11.98	74.07%	1/6 Hz sampling
REDD	microwave	6.69	96.97%	1/3 Hz sampling

7. Conclusions and Future Work

This paper proposes an improved Seq2Point model integrating a denoising autoencoder and multi-task learning for non-intrusive load monitoring. The model jointly optimizes power regression, state classification, and signal reconstruction via a shared deep CNN backbone. Experiments on UK-DALE and REDD datasets demonstrate superior

performance, with average MAE of 12.91 W and 5.02 W, and F1 scores of 84.75% and 96.99%, respectively, outperforming the baseline by 31.4% and 61.1%. Ablation studies confirm the synergistic effect of the two modules.

From a practical perspective, the model enables granular demand-side management, appliance-level energy feedback, and predictive maintenance, while its cross-dataset robustness suggests reduced labeling costs for new deployments. However, several limitations remain. The model still struggles with complex multi-stage appliances (e.g., washing machines) due to limited long-range dependency capture. It operates under a closed-set assumption, failing to recognize unseen appliance types. The denoising module was only validated on Gaussian noise, not on real-world disturbances. Moreover, edge deployment faces computational constraints on low-cost microcontrollers.

Future work will focus on (1) adopting efficient state-space or transformer architectures for better temporal modeling; (2) exploring open-set and few-shot learning to handle unknown appliances; (3) incorporating multi-modal signals (e.g., V-I trajectories) for improved discrimination; (4) applying model compression and quantization for edge deployment; and (5) leveraging federated learning for privacy-preserving cross-household generalization. Addressing these challenges will move the current prototype towards a robust, deployable NILM solution for smart grids and smart homes.

Author Contributions: Conceptualization, X.K.; Methodology, Z.L.; Validation, Z.L.; Formal Analysis, Z.L.; Investigation, Z.L.; Resources, X.K.; Data Curation, C.M.; Writing—Original Draft, Z.L.; Writing—Review and Editing, Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset used in this study is derived from publicly available datasets.

Acknowledgments: The authors wish to thank the editor and reviewers for their valuable suggestions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hart, G.W. Nonintrusive appliance load monitoring. *Proc. IEEE* **1992**, *80*, 1870–1891. [[CrossRef](#)]
2. Peiner, E.; Doering, L. Nondestructive Evaluation of Diesel Spray Holes Using Piezoresistive Sensors. *IEEE Sens. J.* **2013**, *13*, 701–708. [[CrossRef](#)]
3. Schirmer, P.A.; Mporas, I. Non-Intrusive Load Monitoring: A Review. *IEEE Trans. Smart Grid* **2023**, *14*, 769–784. [[CrossRef](#)]
4. Gopinath, R.; Kumar, M.; Joshua, C.P.C.; Srinivas, K. Energy management using non-intrusive load monitoring techniques—State-of-the-art and future research directions. *Sustain. Cities Soc.* **2020**, *62*, 102411. [[CrossRef](#)]
5. Kolter, J.Z.; Jaakkola, T. Approximate inference in additive factorial HMMs with application to energy disaggregation. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*; JMLR: New York, NY, USA, 2012; pp. 1472–1482.
6. Liu, Y.; Wang, X.; You, W. Non-Intrusive Load Monitoring by Voltage–Current Trajectory Enabled Transfer Learning. *IEEE Trans. Smart Grid* **2019**, *10*, 5609–5619. [[CrossRef](#)]
7. Huzzat, A.; Khwaja, A.S.; Alnoman, A.A.; Adhikari, B.; Anpalagan, A.; Woungang, I. A survey of traditional and emerging deep learning techniques for non-intrusive load monitoring. *AI* **2025**, *6*, 213. [[CrossRef](#)]
8. Wang, J.; Li, Y.; Gao, R.X.; Zhang, F. Hybrid physics-based and data-driven models for smart manufacturing: Modelling, simulation, and explainability. *J. Manuf. Syst.* **2022**, *63*, 381–391. [[CrossRef](#)]
9. Wei, Y.; Liu, Y.; Li, Y.; Liu, H.; He, Y.; Gao, S. Deep Learning-Based Non-Intrusive Load Disaggregation Considering Continuous Power Appliances. In *Proceedings of the 2024 IEEE 7th Student Conference on Electric Machines and Systems (SCEMS)*, Macao, China, 6–8 November 2024; pp. 1–7. [[CrossRef](#)]
10. Yang, M.; Li, X.; Liu, Y. Sequence to point learning based on an attention neural network for nonintrusive load decomposition. *Electronics* **2021**, *10*, 1657. [[CrossRef](#)] [[PubMed](#)]

11. Lin, J.; Ma, J.; Zhu, J.; Liang, H. Deep domain adaptation for non-intrusive load monitoring based on a knowledge transfer learning network. *IEEE Trans. Smart Grid* **2022**, *13*, 280–292. [[CrossRef](#)]
12. Jiang, J.; Wang, Z.; Qiu, S.; Li, X.; Zhang, C. Multi-Task Load Identification and Signal Denoising via Hierarchical Knowledge Distillation. *IEEE Trans. Netw. Sci. Eng.* **2025**, *12*, 1967–1980. [[CrossRef](#)]
13. D’Incecco, M.; Squartini, S.; Zhong, M. Transfer Learning for Non-Intrusive Load Monitoring. *IEEE Trans. Smart Grid* **2020**, *11*, 1419–1429. [[CrossRef](#)]
14. Irfan, M.; Tahir, T.; Deilami, S.; Huang, S.; Veettil, B.P. Novel artificial intelligence driven model-based control framework for solar–electric vehicle home energy optimization. *Eng. Appl. Artif. Intell.* **2026**, *177*, 114868. [[CrossRef](#)]
15. Zhang, S.; Fu, Y.; Zhao, X.; Fang, J.; Liu, Y.; Wang, X.; Zhang, B.; Yu, J. Sequence-to-point learning based on spatio-temporal attention fusion network for non-intrusive load monitoring. *Complex Intell. Syst.* **2025**, *11*, 171. [[CrossRef](#)]
16. Zhang, C.; Zhong, M.; Wang, Z.; Goddard, N.; Sutton, C. Sequence-to-point learning with neural networks for nonintrusive load monitoring. *Proc. AAAI Conf. Artif. Intell.* **2018**, *32*, 2604–2611. [[CrossRef](#)]
17. Li, D.; Li, J.; Zeng, X.; Stankovic, V.; Stankovic, L.; Xiao, C.; Shi, Q. Transfer learning for multi-objective non-intrusive load monitoring in smart building. *Appl. Energy* **2023**, *329*, 120223. [[CrossRef](#)]
18. Barber, J.; Cuayáhuil, H.; Zhong, M.; Luan, W. Lightweight non-intrusive load monitoring employing pruned sequence-to-point learning. In Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring, Virtual Event, 18–20 November 2020; pp. 11–15. [[CrossRef](#)]
19. Singh, P.; Sharma, A. Attention-based convolutional denoising autoencoder for two-lead ECG denoising and arrhythmia classification. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 4007710. [[CrossRef](#)]
20. Vincent, P.; Larochelle, H.; Bengio, Y.; Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 5–9 July 2008; pp. 1096–1103.
21. Vincent, P.; Larochelle, H.; Lajoie, I.; Yoshua, B.; Pierre-Antoine, M. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.* **2010**, *11*, 3371–3408. [[CrossRef](#)]
22. Bonfigli, R.; Felicetti, A.; Principi, E.; Fagiani, M.; Squartini, S.; Piazza, F. Denoising autoencoders for non-intrusive load monitoring: Improvements and comparative evaluation. *Energy Build.* **2018**, *158*, 1461–1474. [[CrossRef](#)]
23. Wang, T.S.; Ji, T.Y.; Li, M.S. A new approach for supervised power disaggregation by using a denoising autoencoder and recurrent LSTM network. In Proceedings of the 12th International Symposium on Diagnostics for Electrical Machines, Power Electronics and Drives, Toulouse, France, 27–30 August 2019; pp. 1–6. [[CrossRef](#)]
24. Wallingford, M.; Li, H.; Achille, A.; Ravichandran, A.; Fowlkes, C.; Bhotika, R.; Soatto, S. Task adaptive parameter sharing for multi-task learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 15614–15625.
25. Caruana, R. Multitask learning. *Mach. Learn.* **1997**, *28*, 41–75. [[CrossRef](#)]
26. Crawshaw, M. Multi-task learning with deep neural networks: A survey. *arXiv* **2020**, arXiv:2009.09796.
27. Cao, Y.; Chang, D.; Liu, H.; Cui, J.; Ba, C. Multi-task learning for non-intrusive load monitoring. *Energy Build.* **2018**, *178*, 111–120. [[CrossRef](#)]
28. Ruder, S. An overview of multi-task learning in deep neural networks. *arXiv* **2017**, arXiv:1706.05098.
29. Rong, Y.; Bian, Y.; Xu, T.; Xie, W.; Wei, Y.; Huang, W.; Huang, J. Self-Supervised Graph Transformer on Large-Scale Molecular Data. In Proceedings of the SIGKDD Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 23–27 August 2020; pp. 594–604.
30. Chen, T.; Gao, J.; Yuan, Y.; Guo, S.; Yang, P. Non-intrusive load monitoring based on MoCo_v2, time series self-supervised learning. *Energy Build.* **2024**, *317*, 114374. [[CrossRef](#)]
31. Zhao, P.; Wang, L.; Ding, N.; Wei, J. Non-intrusive power load monitoring based on multi-scale convolution and layer-wise residual graph convolution. *Measurement* **2025**, *246*, 116593. [[CrossRef](#)]
32. Zhang, B.; Luo, F.; He, Y.; Ranzi, G. GT-NILM: A Generative, Transferable Non-Intrusive Load Monitoring System Based on Conditional Diffusion Models and Convolutional Neural Networks. *IEEE Trans. Consum. Electron.* **2025**, *71*, 7906–7919. [[CrossRef](#)]
33. Sudoso, F.; Taloma, R.L.; Pisani, P.; Communiello, D. Seasonal-Trend Transformer-based Diffusion Model for Non-Intrusive Load Monitoring. In Proceedings of the 2025 International Joint Conference on Neural Networks (IJCNN), Rome, Italy, 30 June–5 July 2025; pp. 1–8. [[CrossRef](#)]
34. Dash, S.; Sahoo, N.C. A Multi-Task Deep Learning Approach for Non-Intrusive Load Monitoring of Multiple Appliances. *IEEE Trans. Smart Grid* **2024**, *15*, 3337–3340. [[CrossRef](#)]
35. Liu, Y.; Qiu, J.; Ma, J. SAMNet: Toward latency-free non-intrusive load monitoring via multi-task deep learning. *IEEE Trans. Smart Grid* **2021**, *13*, 2412–2424.
36. Zhao, Z.; Wu, F.; Wu, H.; Liu, T.; Zhou, C.; Ma, J.; Zhang, Y.; Lyu, F. DCML: Boosting Applying Experience of NILM with Dilated Convolution and Multi-Task Learning. In Proceedings of the 2024 IEEE/CIC International Conference on Communications in China (ICCC), Hangzhou, China, 7–9 August 2024; pp. 991–996. [[CrossRef](#)]

37. Liu, Y.; Wang, Y.; Hong, Y.; Shi, Q.; Gao, S.; Huang, X. Toward robust non-intrusive load monitoring via probability model framed ensemble method. *Sensors* **2021**, *21*, 7272. [[CrossRef](#)] [[PubMed](#)]
38. Guo, W.; Luo, Q.; Yu, T.; Hu, X.; Wu, Y.; Liang, M.; Pan, Z. A two-stage robust non-intrusive load monitoring method against unknown waveform distortions. *Sci. Rep.* **2026**, *epub ahead of printing*.
39. Yu, T.T.L.; Singh, G.; Shi, G.; Sadri, H.; Zhang, Y.; Fan, Z. MARLaaS: Multi-Tenant Asynchronous Reinforcement Learning as a Service. *arXiv* **2026**, arXiv:2605.08527.
40. Luo, J.; Liu, S.; Cai, Z.; Xiong, C.; Tu, G. A multi-task learning model for non-intrusive load monitoring based on discrete wavelet transform. *J. Supercomput.* **2023**, *79*, 9021–9046. [[CrossRef](#)]
41. Lu, Z.; Cheng, Y.; Zhong, M.; Luan, W.; Ye, Y.; Wang, G. LightNILM: Lightweight neural network methods for non-intrusive load monitoring. In Proceedings of the 6th International Workshop on Non-Intrusive Load Monitoring, Boston, MA, USA, 9–10 November 2022; pp. 1–6. [[CrossRef](#)]
42. Chen, S.; Zhao, B.; Luan, W.; Zhong, M. Load disaggregation based on sequence-to-point network with unsupervised pre-training. In Proceedings of the 5th IEEE Conference on Energy Internet and Energy System Integration, Taiyuan, China, 22–24 October 2021; pp. 1–6. [[CrossRef](#)]
43. Zhang, Z.; Tan, M.; Pan, R.; Xiao, J.; Liao, C.; Li, Z. A multi-task learning method integrating multi-information fusion for non-intrusive load monitoring. In *2022 6th Asian Conference on Artificial Intelligence Technology (ACAIT)*; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6.
44. Liu, Y.; Yang, Y.; Luan, K. A 1D-CNN-SVM Model for Non-Intrusive Load Identification of Electric Bicycles. In Proceedings of the 2021 IEEE Sustainable Power and Energy Conference (iSPEC), Nanjing, China, 25–27 November 2021; pp. 732–736. [[CrossRef](#)]
45. Chen, H.; Wang, Y.H.; Fan, C.H. A convolutional autoencoder-based approach with batch normalization for energy disaggregation: H. *J. Supercomput.* **2021**, *77*, 2961–2978.
46. Cao, Y.; Chang, D.; Liu, H.; Cui, J.; Ba, C. Multi-task learning for non-intrusive load monitoring. In *2023 5th International Conference on Frontiers Technology of Information and Computer (ICFTIC)*; IEEE: Piscataway, NJ, USA, 2023; pp. 259–267.
47. Xia, P.; Zhou, H.; Yang, T.; Zhou, W.; Liu, Z.; Wang, X.; Li, X.Y. ESATED: Leveraging Extra-weak Supervision with Auxiliary Task for Enhanced Non-intrusiveness in Energy Disaggregation. In Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, New York, NY, USA, 5–9 October 2024; 8, pp. 1–32. [[CrossRef](#)]
48. Virtsionis Gkalinikis, N.; Nalmpantis, C.; Vrakas, D. Variational regression for multi-target energy disaggregation. *Sensors* **2023**, *23*, 2051. [[CrossRef](#)] [[PubMed](#)]
49. Kuzengurira, T.T.; Guangfen, W.; Yilin, L. Sequence-to-point learning methods for Non Intrusive Load Disaggregation: A Review. In *2024 6th International Conference on Energy, Power and Grid (ICEPG)*; IEEE: Piscataway, NJ, USA, 2024; pp. 202–206.
50. Xu, W.; Jiang, C.; Zhang, Q.; Zheng, J. An intelligent non-intrusive load monitoring model based on power encoding and convolutional state modules. *Meas. Sci. Technol.* **2024**, *35*, 086210. [[CrossRef](#)]
51. Faustine, A.; Pereira, L.; Bousbiat, H.; Kulkarni, S. UNet-NILM: A deep neural network for multi-tasks appliances state detection and power estimation in NILM. In *NILM'20: Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 84–88.
52. Aboulia, A.; Green, D.H.; Switzer, J.F.; Kane, T.J.; Bredariol, G.V.; Lindahl, P.; Donnal, J.S.; Leeb, S.B. NILM dashboard: A power system monitor for electromechanical equipment diagnostics. *IEEE Trans. Ind. Inform.* **2018**, *15*, 1405–1414. [[CrossRef](#)]
53. Han, Y.; Chen, H.; Wu, J.; Zhao, Q. Reconstruction-based supervised contrastive learning for unknown device identification in nonintrusive load monitoring. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 2511313. [[CrossRef](#)]
54. Park, J.; Yoo, J.; Kim, T.; Ha, J.M.; Youn, B.D. Multi-head de-noising autoencoder-based multi-task model for fault diagnosis of rolling element bearings under various speed conditions. *J. Comput. Des. Eng.* **2023**, *10*, 1804–1820. [[CrossRef](#)]
55. Qian, X.; Wang, Y.W. Image denoising via multi-task learning. In Proceedings of the 4th International Conference on Wireless Communication and Artificial Intelligence, Chengdu, China, 17–19 December 2021; pp. 1–6. [[CrossRef](#)]
56. Huo, F.; Luo, Q.; Yu, T.; Wang, K.; Liang, Z.; Liang, M.; Pan, Z. Non-intrusive load monitoring based on waveform restoration and feature fusion under complex backgrounds. *Meas. Sci. Technol.* **2025**, *36*, 046115. [[CrossRef](#)]
57. Kolter, J.Z.; Johnson, M.J. REDD: A public data set for energy disaggregation research. In *Proceedings of the SIGKDD Workshop on Data Mining Applications in Sustainability*; ACM: New York, NY, USA, 2011.
58. Kelly, J.; Knottenbelt, W. The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes. *Sci. Data* **2015**, *2*, 150007. [[CrossRef](#)] [[PubMed](#)]
59. Batra, N.; Kukunuri, R.; Pandey, A.; Malakar, R.; Kumar, R.; Krystalakos, O.; Zhong, M.; Meira, P.; Parson, O. Towards reproducible state-of-the-art energy disaggregation. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*; ACM: New York, NY, USA, 2019; pp. 1–10. [[CrossRef](#)]

60. Zhang, J.; Sun, J.; Gan, J.; Liu, Q.; Liu, X. Improving Domestic NILM Using an Attention-Enabled Seq2Point Learning Approach. In Proceedings of the 2021 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCOM/CyberSciTech), Calgary, AB, Canada, 25–28 October 2021; pp. 434–439. [[CrossRef](#)]
61. Yue, Z.; Witzig, C.R.; Jorde, D.; Jacobsen, H.A. BERT4NILM: A bidirectional transformer model for non-intrusive load monitoring. In *NILM'20: Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 89–93.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.