energies

MDPI

Article

Multi-Task NILM with Anomaly Detection Using a Hybrid CNN–BiLSTM–Transformer Model

Mihriban Gunay [1,*], Yakup Demir [2] and Marin Zhilevski [3,*]

[1] Department of Electrical and Electronics Engineering, University of Munzur, Tunceli 62000, Türkiye
[2] Department of Electrical and Electronics Engineering, University of Firat, Elazig 23000, Türkiye; ydemir@firat.edu.tr
[3] Faculty of Automatics, Technical University of Sofia, 1000 Sofia, Bulgaria
* Correspondence: mihribangunay@munzur.edu.tr (M.G.); mzhilevski@tu-sofia.bg (M.Z.)

Abstract

Non-Intrusive Load Monitoring (NILM) enables estimation of the energy use of individual appliances in smart buildings from a single aggregate meter. In practice, however, this task is not straightforward. Signals from different appliances can overlap, and the measured data may also include distortions such as spikes, drops, and noise. To address these issues, this study presents a multi-task triple-hybrid deep learning framework that handles appliance classification and anomaly detection together. The model brings together 1D-CNN, BiLSTM, and Transformer Attention so that local patterns, temporal dependencies, and wider contextual information can be learned within the same structure. It also uses a dual-output design to classify appliance categories and detect anomaly types simultaneously. Experiments were carried out on Building 1 of the UK-DALE dataset with four appliances: kettle, microwave, washer dryer, and fridge freezer. For the anomaly task, synthetic disturbances were added to segmented signal windows and grouped as normal, spike, drop, and noise. To check how well the proposed framework handled different scenarios, it was tested on both the UK-DALE and REDD datasets. Looking at the main UK-DALE results, the model correctly identified appliances 99.48% of the time and spotted anomalies with 98.80% accuracy. A secondary test on the REDD dataset yielded an 86.44% classification score. This proves the architecture can adjust to completely new power grid environments without losing its edge. On top of that, when pitted against standard benchmark models like Seq2Point, this triple-hybrid design clearly does a better job of mapping out complex signal changes. As a result, it yields much stronger anomaly detection metrics.

Keywords: non-intrusive load monitoring; NILM; appliance classification; anomaly detection; deep learning; 1D-CNN; BiLSTM; Transformer Attention; UK-DALE

Check for updates

Academic Editor: Sérgio Cruz

Received: 7 May 2026
Revised: 14 June 2026
Accepted: 19 June 2026
Published: 24 June 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) license.

1. Introduction

The growing global demand for energy, together with the push toward carbon-neutral targets, has made energy efficiency a major priority in buildings and industrial facilities. Since residential and commercial buildings account for a large share of overall energy use, understanding appliance-level consumption has become increasingly important for improving efficiency and reducing waste. In this context, Non-Intrusive Load Monitoring (NILM), first introduced by Hart [1], has emerged as a practical and cost-effective approach. By using data from a single smart meter, NILM estimates the energy consumption and operating states of individual appliances without requiring separate sensors for each device.

Early NILM studies mainly relied on statistical and conventional machine learning methods such as Hidden Markov Models (HMMs) and combinatorial optimization [1,2]. Over time, however, advances in computing power have shifted attention toward Deep Learning (DL)-based approaches. Convolutional Neural Networks (CNNs) [3] have proven effective in capturing local transient characteristics in power signals, including sudden spikes and ON/OFF events. Similarly, sequence-based models such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks have been widely used to model temporal dependencies in appliance behavior. Despite these strengths, each approach has its own limitations [4]. CNNs are generally effective at local pattern extraction, but they are less capable of learning long-range temporal relationships. Recurrent models, on the other hand, may suffer from vanishing gradients and overfitting when handling long input sequences. More recently, Transformer architectures and self-attention mechanisms, which have shown remarkable success in Natural Language Processing (NLP), have attracted increasing interest in NILM research because of their ability to highlight the most informative parts of a sequence [5,6].

A key challenge in NILM today is not only identifying appliances accurately, but also detecting anomalies in the power signal, such as spikes, drops, and noise, especially in complex and noisy settings where several devices operate simultaneously. In many existing studies, appliance recognition and anomaly detection are treated as separate problems. This separation can increase computational cost and reduce the suitability of these methods for real-time applications [7,8].

To address this gap, this study proposes a multi-task triple-hybrid deep learning architecture that performs appliance classification and anomaly detection within a unified framework. The proposed model combines CNN, BiLSTM, and Transformer Attention modules in a sequential manner. In this structure, 1D-CNN layers extract short-term local patterns from raw power signals, the BiLSTM layer captures temporal dependencies in both forward and backward directions, and the Transformer encoder emphasizes the most informative feature representations through multi-head attention. Supported by a dual-output design, denoted as app_out and anom_out, the model can classify target appliances accurately while also detecting anomalies that affect signal quality.

2. Related Works

Over the last few years, deep learning has clearly changed the direction of NILM research and has led to better performance than many traditional methods. Existing studies show that this progress has come from several lines of work, including convolution-based models, recurrent networks, attention mechanisms, and multi-task learning frameworks.

Deep Learning and Hybrid Architectures: Among the early studies that shaped this area, Kelly and Knottenbelt [9] showed that deep learning could be applied effectively to NILM through Denoising Autoencoders (DAEs) and convolutional networks. Their work opened the way for later architectures designed specifically for load disaggregation. One of the best-known examples is the Sequence-to-Point (Seq2Point) model introduced by Zhang et al., which estimates the midpoint of an input sequence and later became one of the most widely used approaches in NILM. As the field progressed, researchers increasingly started to combine different architectures instead of relying on a single model type. For example, Huzzat et al. [10] proposed a GRU-BERT hybrid model, where Bi-GRU layers were used to learn temporal load behavior and the BERT component helped capture wider contextual information. He et al. [11] followed a similar idea by integrating Transformer and CNN modules in a sequence-to-sequence framework for noisy industrial settings such as cement factories. In another study, Ouzine et al. [12] introduced the Parallel Hybrid PHCNN-GRU model to strengthen feature extraction and improve multi-target regression forecasting.

Multi-Task Learning: Multi-task learning has become an important direction in NILM because it allows a model to learn more than one related objective from the same dataset, often improving generalization performance. Shin et al. [13] introduced the Subtask Gated Networks architecture, which jointly performs energy disaggregation and load on/off state detection. Similarly, Çimen et al. [14] applied a multi-task GRU (M-GRU) framework to microgrid energy management and showed that learning appliance consumption and load states together produced better results than single-task models. Dash and Sahoo [15] extended this idea by proposing a Transformer-based, attention-driven, multi-task probabilistic network that simultaneously handles appliance state detection and energy disaggregation.

Anomaly Detection and Comprehensive Feature Extraction: Another important issue in NILM is the detection of noise, abnormal patterns, and appliance malfunctions, since such disturbances can directly affect model reliability. De Diego-Otón et al. [16] used MobileNet-based convolutional feature extractors for appliance classification and for detecting unidentified appliances or anomalous patterns in limited-data settings. Saha et al. [17] proposed a Capsule Network framework that simultaneously identifies appliance type and activity state by taking advantage of stronger feature representation capabilities. Hu et al. [18] introduced a Multi-Scale Multi-Label Classification Model (MMCM) to identify multiple appliance load states simultaneously.

The model proposed in this study is designed to overcome key limitations of existing hybrid and multi-task approaches in the NILM literature. Recent works, such as the GRU-BERT model of Huzzat et al. [10] and the Tra-CNN framework proposed by Cheng and Zhong [19], have shown that combining different neural network components can improve load identification performance. Nevertheless, these studies generally integrate only two major architectural components. In contrast, the proposed framework introduces a more comprehensive triple-hybrid architecture that combines 1D-CNN layers for extracting local transient patterns, a BiLSTM layer for learning bidirectional temporal dependencies, and Transformer Attention modules for enhancing feature representation through global contextual modeling. In addition, the proposed model extends beyond conventional appliance recognition by jointly performing appliance classification and anomaly detection for spike, drop, and noise patterns. While hybrid configurations combining convolutional layers and attention mechanisms have been successfully deployed in other domains, their application in NILM has predominantly been restricted to dual-component setups (e.g., CNN-Transformer or GRU-BERT) focusing solely on load disaggregation. The architecture proposed in this study represents a multi-task triple-hybrid framework specifically engineered to capture low-level micro-transients via 1D-CNN, recursive sequential evolution via BiLSTM, and global contextual attention via a Transformer block simultaneously for both appliance classification and multi-class anomaly identification. This unified design distinguishes the proposed approach from many existing NILM methods and enhances its practical relevance for real-world smart energy monitoring applications.

3. Problem Formulation and Theoretical Background

3.1. Problem Formulation

Traditional NILM methods are generally used to estimate the energy consumption of individual appliances from aggregated smart meter data [1]. In most earlier studies, this task was handled as a regression problem, with the main goal being to separate the total household load into appliance-level consumption signals. This view is useful, but it does not fully capture what happens in real smart grid settings. In practice, power data are often affected by abnormal appliance behavior, noise in the measurements, communication issues, or unexpected user actions [20]. These irregularities can distort the input signal; as a result, the reliability of NILMs is reduced when they are ignored.

For this reason, the present study treats NILM differently and formulates it as a multi-task classification problem [21]. Instead of focusing only on appliance-level estimation, the proposed framework is designed to determine which appliance is operating and whether the observed consumption pattern is abnormal within the same time window. In this way, appliance recognition and anomaly detection are handled together in a single learning framework. Specifically, this framework models anomaly detection as a fully supervised classification task rather than a time-series forecasting or prediction problem. Since the ground-truth labels for both normal and anomalous operations are explicitly defined and known during the offline training stage (as detailed in Section 4.2), the model learns to map input sequences directly into discrete categorical classes.

Let the time-series power consumption signal be represented as $\mathbf{x} \in \mathbb{R}^{\mathtt{T}}$ where $\mathtt{T}$ denotes the length of the observation window.

The objective of the proposed model is to learn an optimal mapping function $f(\cdot)$ that simultaneously predicts two target variables [21]:

$$f(\mathbf{x}) = \left(\hat{y}_{app}, \hat{y}_{anom}\right) \tag{1}$$

where

- $\hat{y}_{app} \in 0,\ 1,\ \ldots,\ N-1$ denotes the predicted appliance class among N different appliance categories;
- $\hat{y}_{anom} \in 0,\ 1,\ 2,\ 3$ represents the anomaly class of the input signal, corresponding to 0: Normal, 1: Spike, 2: Drop, and 3: Noise patterns.

To jointly optimize these tasks, the proposed multi-task learning framework minimizes a combined loss function defined as [21]:

$$\boldsymbol{\mathcal{L}}_{total} = \lambda_1 \boldsymbol{\mathcal{L}}_{app}\left(\hat{y}_{app}, y_{app}\right) \ + \ \lambda_2 \boldsymbol{\mathcal{L}}_{anom}\left(\hat{y}_{anom}, y_{anom}\right) \tag{2}$$

where $\boldsymbol{\mathcal{L}}_{app}$ and $\boldsymbol{\mathcal{L}}_{anom}$ denote the cross-entropy loss functions corresponding to appliance classification and anomaly detection tasks, respectively. The coefficients λ_1 and λ_2 are weighting parameters that balance the contribution of each task during model training.

Through this joint optimization strategy, the neural network learns shared feature representations that capture both appliance-specific power signatures and abnormal signal characteristics, thereby improving the overall generalization capability of the system.

3.2. Theoretical Background

The proposed triple-hybrid architecture, consisting of CNN, BiLSTM, and Transformer components, is designed to capture both local and global dependencies in electricity consumption signals. The main theoretical components of the model are briefly summarized below.

3.2.1. 1D Convolutional Neural Networks (1D-CNN)

One-dimensional Convolutional Neural Networks (1D-CNNs) are commonly used to learn local patterns from time-series data [22]. For NILM, this is especially useful because short-duration changes in the power signal often reflect appliance switching behavior. CNN layers can capture such local variations directly from raw input and provide a strong starting point for feature extraction.

Given an input signal $\mathbf{x}$ and a convolutional filter $\mathbf{w}$ of size k, the convolution operation can be expressed as [23]:

$$y[n] = \sum_{i=0}^{k-1} x[n-i] \cdot w[i] \ + \ b \tag{3}$$

where b denotes the bias term. By applying multiple convolutional filters, the CNN layers learn discriminative local features that represent appliance-specific power signatures.

These characteristics make CNN layers particularly suitable for capturing short-term transient patterns and sudden fluctuations in power consumption signals, which are critical for identifying appliance switching events and abnormal signal variations.

3.2.2. Bidirectional Long Short-Term Memory (BiLSTM)

Local patterns alone, however, are not enough to describe how appliances operate over time. Many devices follow temporal patterns that extend beyond short signal fragments, which makes sequence modeling necessary. Long Short-Term Memory (LSTM) networks are well suited to this task because they are designed to learn temporal dependencies from sequential data [24].

In BiLSTM networks, the sequence is processed in both forward and backward directions, producing two hidden states for each time step: a forward hidden state $\overrightarrow{h}_t$ and a backward hidden state $\overleftarrow{h}_t$ [25]. These representations are concatenated to obtain the final contextual representation [25]:

$$h_t = \left[\overrightarrow{h}_t, \overleftarrow{h}_t\right] \tag{4}$$

In this study, a bidirectional LSTM (BiLSTM) is used so that the model can consider both earlier and later parts of the sequence. This helps represent appliance operating cycles more clearly and supports the learning of more complex temporal behavior.

3.2.3. Transformer and Multi-Head Attention

To capture wider relationships across the full sequence, the model also includes a Transformer-based self-attention mechanism [26]. Unlike recurrent models, self-attention does not process the sequence step by step. Instead, it learns how different parts of the signal relate to one another directly.

The attention mechanism computes the relevance between query (Q), key (K) and value (V) matrices as [26]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5}$$

where d_k is the scaling factor corresponding to the dimensionality of the key vectors.

The Multi-Head Attention mechanism applies this attention operation in multiple parallel subspaces, enabling the model to capture different types of dependencies within the time-series signal. In the context of NILM, this mechanism helps the network focus on the most informative segments of the power signal, improving both appliance classification and anomaly detection performance.

By capturing global contextual relationships across the entire sequence, the Transformer attention mechanism further enhances the model's ability to focus on the most informative regions of the power signal for both appliance classification and anomaly detection tasks.

4. Materials and Methods

4.1. Dataset and Data Preprocessing

This study is based on the UK Domestic Appliance-Level Electricity (UK-DALE) dataset [27], which is widely used in Non-Intrusive Load Monitoring (NILM) research. The dataset contains high-resolution electricity measurements collected from residential buildings in the United Kingdom and has become a common reference point for evaluating appliance identification and load disaggregation methods.

For the present analysis, data from Building 1 were used. Four household appliances were selected: kettle, microwave, washer dryer and fridge freezer. These appliances were included because they show clearly different operating behaviors and are frequently examined in NILM studies. Their power consumption signals were obtained through the NILM Toolkit (NILMTK, Version 0.4.0) using Python 3.12.13, with a sampling interval of 6 s, so that all measurements were handled at the same temporal resolution.

Figure 1 shows the characteristic power consumption patterns of the selected appliances. The figure makes it clear that each appliance produces a different signal shape over time. Kettles and microwaves, for example, tend to generate short but sharp peaks because of their high power demand. Washer dryers, by contrast, usually operate over longer periods and exhibit more complex patterns. The fridge freezer shows a more regular and lower-amplitude consumption profile. Taken together, these differences indicate that NILM models need to capture both short transient events and longer temporal behavior in power signals.

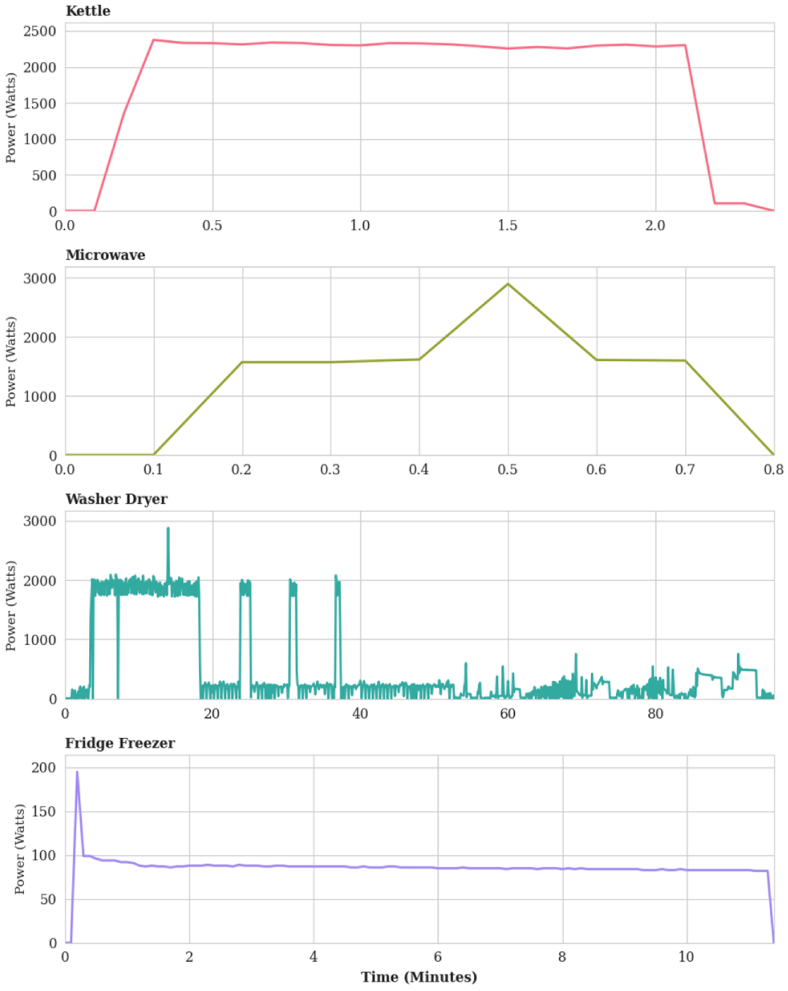


Figure 1. Real power consumption signatures of target appliances.

To comprehensively evaluate the multi-task model's cross-dataset generalization capability and ensure that the feature abstraction pipeline is not overfitted to a specific geographical grid configuration, the Reference Energy Disaggregation Data Set (REDD) [28] was additionally utilized. Developed by MIT, the REDD dataset contains whole-house and appliance-level electricity consumption monitoring from real-world residential homes in the United States. For the cross-dataset validation framework, active power measurements from Building 1 were extracted. To maintain evaluation consistency and align closely with

the primary UK-DALE experiments, the target configuration was mapped to four comparable representative appliances: refrigerator, oven (utilized as a high-power surrogate for the microwave due to missing channel records in the specific building), washer dryer, and dish washer. The sampling resolution was uniformly aligned to 6 s to provide an identical input environment across different smart grid definitions without modifying the core model layout.

Before model training, the raw power signals were preprocessed to improve consistency and numerical stability. Missing values were replaced with zeros so that continuity in the time-series data could be preserved. After that, each appliance-specific signal was normalized using StandardScaler, which rescales the data to zero mean and unit variance. This step helps reduce scale differences among appliances and supports more stable and efficient model training.

The normalized signals were then divided into segments using a sliding-window procedure. Each sample was defined as a fixed-length time-series window with $T = 128$, and consecutive windows were extracted with a stride of 32 samples, resulting in overlapping segments. This segmentation scheme allows the model to capture short-term signal changes while still preserving temporal continuity across neighboring windows.

To improve data quality and limit the effect of long inactive periods, an additional filtering rule was applied during window generation. More specifically, windows were retained when their maximum value satisfied $\max(x) > 0.1$. Windows with lower activity were not removed completely; instead, they were kept with a small probability of 5% so that low-power operating states would still be represented in the dataset. In this way, the dataset preserves diversity while reducing the influence of less informative samples. Each resulting window was then assigned its corresponding appliance label and used as inputs for the anomaly generation stage and the subsequent multi-task learning framework.

4.2. Synthetic Anomaly Injection

Since publicly available NILM datasets usually do not include sufficiently labeled anomaly categories, a synthetic anomaly injection procedure was added to the data preparation pipeline. The main purpose of this step was to generate anomalous signal patterns in a controlled way and to make it possible to train the model for anomaly classification together with appliance recognition.

The anomaly generation procedure was applied to each segmented window after preprocessing. However, anomalies were not injected into every window without restriction. To avoid unrealistic modifications in nearly inactive signals, a threshold-based rule was introduced in the implementation. If the maximum value of a window was below 0.1, that window was left unchanged and directly assigned the normal label. In this way, anomaly generation was limited to windows containing sufficiently informative signal activity.

For the windows that satisfied this condition, synthetic anomalies were injected with a probability of 0.15. Once anomaly injection was activated, one of three anomaly types was selected at random: spike, drop, or noise. The anomalous region within the window was then determined by randomly choosing a start index and a short segment length ranging from 2 to 10 samples.

For spike anomalies, the selected segment was replaced by an amplified value derived from the maximum amplitude of the window. More specifically, the signal in that region was multiplied by a random factor between 2 and 5, simulating a sudden increase in power consumption. For drop anomalies, the selected segment was scaled by a random coefficient between 0.0 and 0.2, producing a sharp decrease in signal magnitude. For noise anomalies, zero-mean Gaussian noise with a standard deviation of 0.5 was added to the selected segment in order to mimic irregular perturbations and measurement distortions.

As a result, each input window was assigned one of four anomaly labels: Normal (0), Spike (1), Drop (2), or Noise (3). This procedure enriched the training set with a wider range of abnormal patterns and enabled the proposed framework to learn more discriminative features for anomaly-aware NILM analysis.

The strategic rationale for introducing these synthetic mathematical anomalies into the post-standardization domain, rather than applying them to raw power profiles (in Watts) before normalization, is deeply rooted in preventing statistical bias across heterogeneous appliance profiles. For instance, an anomaly introduced in absolute raw units—such as a fixed 300 W spike—would induce a catastrophic and unrealistic deformation on a low-power device like a fridge freezer (nominal base load ≈ 90 W), while remaining entirely unnoticeable within the active operational cycle of a high-power device like a kettle (≈2400 W). Aligning unique raw thresholds for every specific operational mode would exponentially inflate rule-based hyperparameter complexity. By enforcing the StandardScaler protocol first, all electrical signatures are mapped onto a uniform zero-mean and unit-variance sequence space. Consequently, injecting stochastic geometry modifications (multiplying by a factor of 2–5 or adding Gaussian noise) within this standardized boundary ensures that the structural deformation scales relative to each device's natural variance profile. This invariant representation prevents the multi-task learning heads from being blinded by absolute power domains, encouraging stable gradient synchronization.

4.3. Proposed Triple-Hybrid Architecture

The proposed model employs a triple-hybrid deep learning architecture that integrates convolutional, recurrent, and attention-based components within a unified multi-task framework. The overall structure of the model is illustrated in Figure 2.

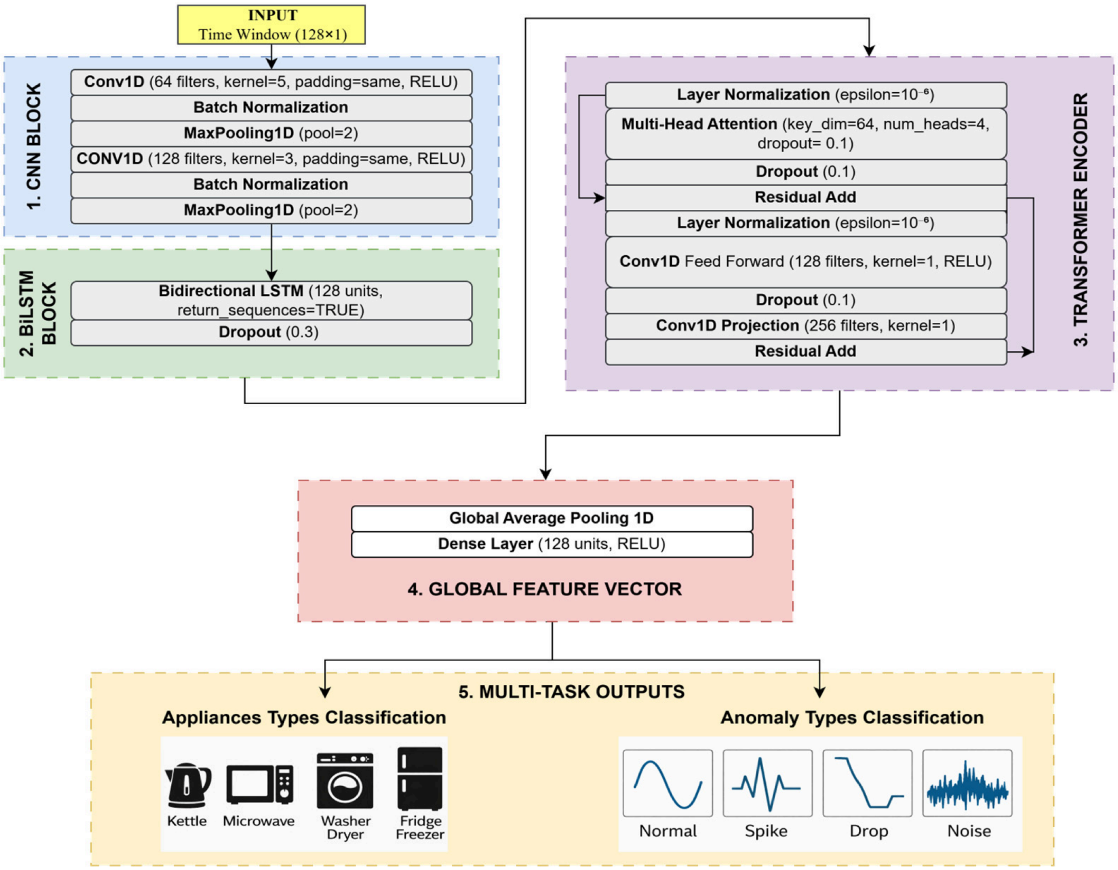


Figure 2. Proposed triple-hybrid architecture integrating CNN, BiLSTM, and Transformer encoder blocks for joint appliance classification and anomaly detection.

Figure 2 shows the overall architecture of the proposed model. The input to the network is a fixed-length window of size 128 × 1. Rather than relying on a single feature extraction strategy, the model is built in a layered way so that different properties of the electricity signal can be learned at different stages. In simple terms, the early part of the network looks for local changes, the middle part follows how the signal evolves over time, and the final part helps the model understand broader relationships across the whole sequence.

It is important to highlight that a sequential (stacked) network topology was chosen over a parallel architecture to achieve hierarchical feature abstraction. If the 1D-CNN, BiLSTM, and Transformer modules were deployed in parallel, each model would be forced to process the raw, noisy, and high-dimensional time-series data independently, which limits their synergistic potential. In our stacked configuration, these modules function complementarily: the early 1D-CNN layers perform localized spatial filtering to extract short-term transient signatures (e.g., abrupt switching events) while reducing dimensionality. The subsequent BiLSTM layer receives these clean, down-sampled feature maps to model bidirectional, medium-term temporal dependencies. Finally, the Transformer Attention encoder operates on top of these temporal representations to capture global, long-range contextual relationships via multi-head self-attention. This progressive refinement pipeline creates a highly discriminative feature hierarchy that cannot be replicated by simply merging independent parallel outputs.

The signal first passes through two 1D convolutional layers. The first uses 64 filters with a kernel size of 5, and the second uses 128 filters with a kernel size of 3. Batch normalization and max-pooling are applied after each convolution. These layers help the model pick up short and sudden variations in the power signal, such as switching events or brief fluctuations in consumption. At the same time, the pooling operations gradually reduce the temporal resolution and make the learned representation more compact.

The feature maps produced by the convolutional block are then given to a Bidirectional LSTM (BiLSTM) layer with 128 units and return_sequences = True. This part of the model reads the sequence in both directions, which means it can use information from both earlier and later time steps. That is helpful in NILM, since many appliances do not operate through isolated events but through patterns that develop over time. A dropout layer with a rate of 0.3 is added after the BiLSTM to improve generalization.

To strengthen the representation further, a Transformer encoder is placed after the recurrent stage. As illustrated in Figure 2, this block includes layer normalization, multi-head self-attention, residual connections, and a feed-forward module implemented with 1D convolutional layers. The attention component uses 4 heads with a key dimension of 64. This allows the network to learn how different parts of the sequence relate to one another, even when they are far apart in time. As a result, the model can focus more effectively on the regions of the signal that carry the most useful information for both appliance recognition and anomaly detection.

The sequence representation obtained from the attention block is then summarized by global average pooling. In this way, a compact feature vector is produced. This vector is passed through a fully connected layer with 128 neurons and ReLU activation, creating a shared latent representation that is used by both output branches.

At the final stage, the network separates into two parallel heads. One head performs appliance classification over four device classes, while the other classifies anomaly categories, namely normal, spike, drop, and noise. Learning these two tasks together allows the model to make use of shared information and encourages it to learn more meaningful features than it would in a single-task setting.

Overall, the proposed architecture brings together three different but complementary components. The CNN layers are useful for local signal patterns, the BiLSTM captures temporal structure, and the Transformer encoder helps model broader contextual relationships. By combining these modules within a single framework, the model is better able to represent the complex and heterogeneous nature of appliance-level electricity consumption signals.

5. Experimental Results

5.1. Experimental Setup

This section presents the experimental setup, training process, and evaluation procedure used for the proposed triple-hybrid multi-task learning framework. All experiments were carried out on Building 1 of the UK-DALE dataset and focused on four household appliances: kettle, microwave, washer dryer, and fridge freezer. The raw power signals were sampled every 6 s and divided into fixed-length windows of 128 time steps using a sliding-window method with a stride of 32. This setup was chosen to preserve short-term signal changes while also retaining temporal continuity across the sequence. Before training, all input signals were normalized with StandardScaler to improve numerical stability and make optimization more reliable. During sample generation, windows with very low activity, defined by a maximum value below 0.1, were generally excluded, although a small portion of them (5%) were kept so that near-idle states would still be represented in the dataset. In addition, a synthetic anomaly injection step was included in the preprocessing pipeline. For each window, anomalies were introduced with a probability of 0.15, producing three perturbation types: spike, drop, and noise. In this way, the model was trained to perform appliance classification and anomaly detection together within the same framework. The dataset was split into training and testing subsets using a stratified strategy based on appliance labels, with 15% reserved for testing. To examine robustness, 5-fold cross-validation was also applied to the training set. Within each fold, 10% of the training data were used as a validation subset for performance monitoring and early stopping. To ensure rigorous validation and address potential overfitting or data leakage concerns, the experimental pipeline incorporates several methodological safeguards. First, the baseline activity filtering protocol (max(x) > 0.1) systematically removes repetitive standby states, ensuring the model evaluates only distinct operational transients. Second, the synthetic anomaly injection framework introduces stochastic shape distortions at random window indices with fixed probability boundaries, rendering individual sequences mathematically unique. The structural independence of these active signatures ensures that random data partitioning does not result in the memorization of contiguous sequence replicas, as validated by the distinct performance variations across the different cross-validation folds. The model was trained with the Adam optimizer and a batch size of 64. Training was limited to 20 epochs, but early stopping with a patience of 3 epochs was used to reduce overfitting and retain the best model weights. The optimization objective consisted of a joint loss formed by two sparse categorical cross-entropy terms, corresponding to appliance classification and anomaly detection. Model performance was assessed using accuracy, precision, recall, and F1-score for both tasks. ROC and precision-recall analyses were also performed to provide a more detailed view of classification behavior across different decision thresholds. All experiments were implemented in TensorFlow/Keras and run on Google Colab using an NVIDIA T4 GPU. The main experimental settings are summarized in Table 1.

Table 1. Experimental configuration.

Category	Setting
Dataset	UK-DALE (Building 1)
Appliances	Kettle, Microwave, Washer Dryer, Fridge Freezer
Sampling& Windowing	6 s sampling, window size = 128, stride = 32
Normalization	StandardScaler
Anomaly Injection	Probability = 0.15 (Spike, Drop, Noise)
Data Split	85% train/15% test
Validation Strategy	5-fold cross-validation, 10% validation split
Training Setup	Batch size = 64, epochs = 20, early stopping (patience = 3)
Optimizer & Loss	Adam, multi-task sparse cross-entropy
Evaluation Metrics	Accuracy, Precision, Recall, F1-score, ROC, PR
Implementation	TensorFlow/Keras, NVIDIA T4 GPU

5.2. Appliance Classification Results

The performance of the proposed model on the appliance classification task was assessed using several evaluation metrics, namely accuracy, precision, recall, and F1-score. The model achieved an overall classification accuracy of 99.48%, indicating that it was able to distinguish appliance categories with a high level of reliability.

The confusion matrix shown in Figure 3 further supports this result by showing that most samples were classified correctly and that misclassification remained very limited. In particular, the model produced almost perfect predictions for the washer dryer and fridge freezer classes. A small amount of confusion was observed between kettle and microwave, which is reasonable given that these two appliances can exhibit somewhat similar consumption characteristics. Even so, the number of misclassified samples between these classes remained low.

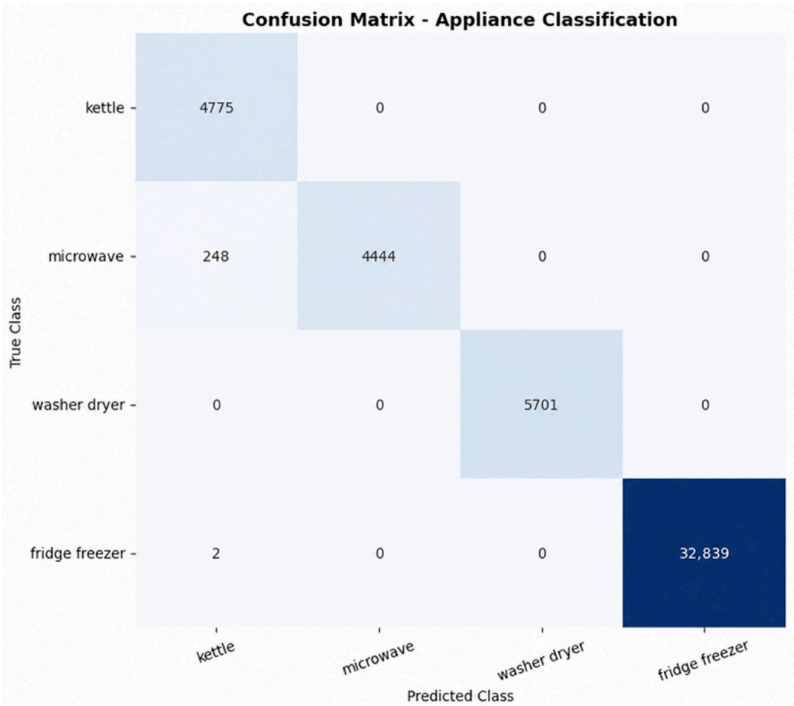


Figure 3. Confusion matrix for appliance classification.

To further assess class separability, the ROC curves in Figure 4 indicate near-ideal classification performance across all appliance categories, with all classes achieving an AUC value close to 1.00. This confirms the robustness of the proposed model under varying decision thresholds.

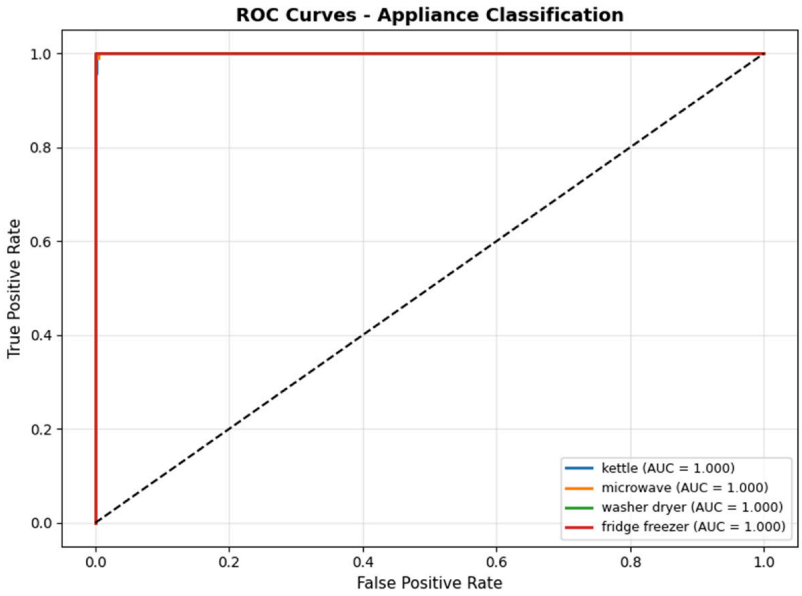


Figure 4. ROC curves for appliance classification.

The class-wise performance results in Figure 5 demonstrate consistently high precision, recall, and F1-score values across all appliance categories. Specifically, the model achieves a precision of 95.02%, 100.00%, 100.00%, and 100.00% for kettle, microwave, washer dryer, and fridge freezer, respectively. The recall values are 100.00%, 94.71%, 100.00%, and 99.99%, while the corresponding F1-scores are 97.44%, 97.28%, 100.00%, and 99.99%. These results indicate that the proposed model maintains a well-balanced performance across all classes, with particularly strong classification capability for high-consumption appliances.

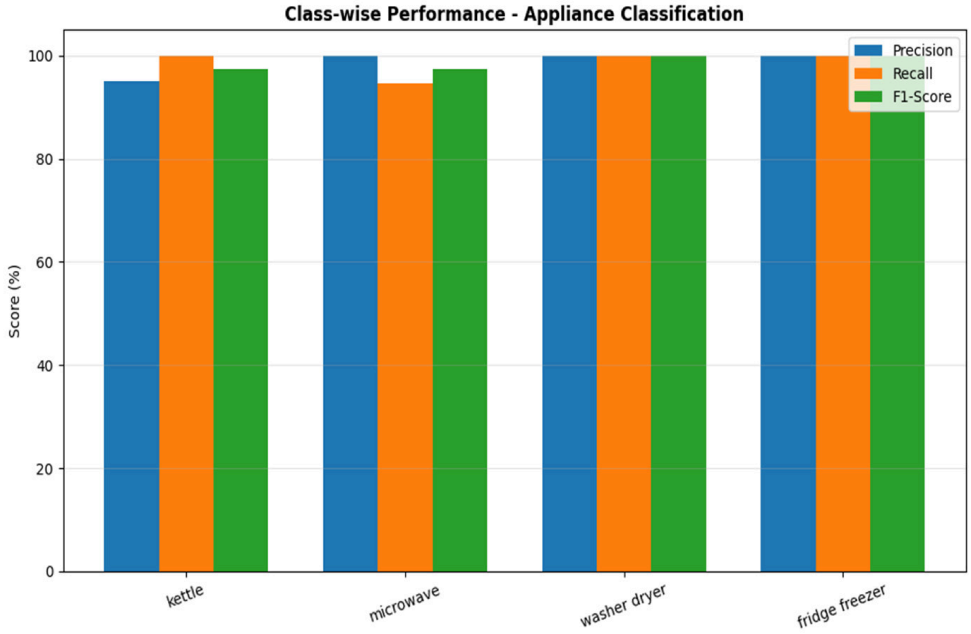


Figure 5. Class-wise performance for appliance classification.

5.3. Anomaly Detection Results

To evaluate anomaly detection performance, accuracy, precision, recall, and F1-score were considered. The proposed model reached an overall accuracy of 98.80%, showing that it could reliably identify both normal and abnormal patterns in appliance-level power signals.

The confusion matrix in Figure 6 shows that most anomaly classes were identified correctly, with only limited misclassification. The normal class, in particular, was detected with very high accuracy. Most of the observed errors occurred between the Noise and Normal classes, and to a lesser extent between the Drop and Normal classes. This outcome is understandable, as the differences between these signal patterns can be relatively subtle.

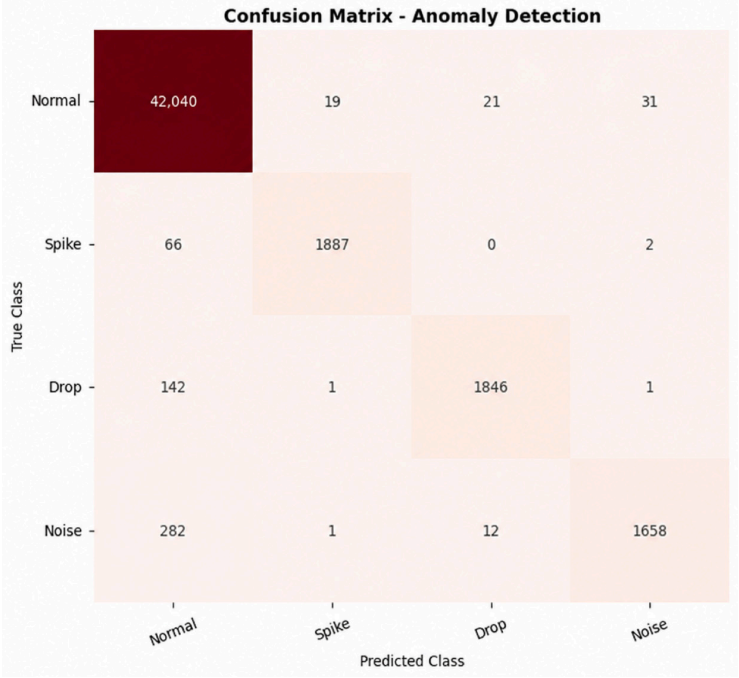


Figure 6. Confusion matrix for anomaly detection.

Figure 7 gives a clearer picture of how well the anomaly classes are separated. The ROC curves remain strong for all classes, with AUC values of 0.993 for Normal, 0.999 for Spike, 0.998 for Drop, and 0.985 for Noise. This suggests that the model can still distinguish between anomaly types effectively even when the decision threshold changes.

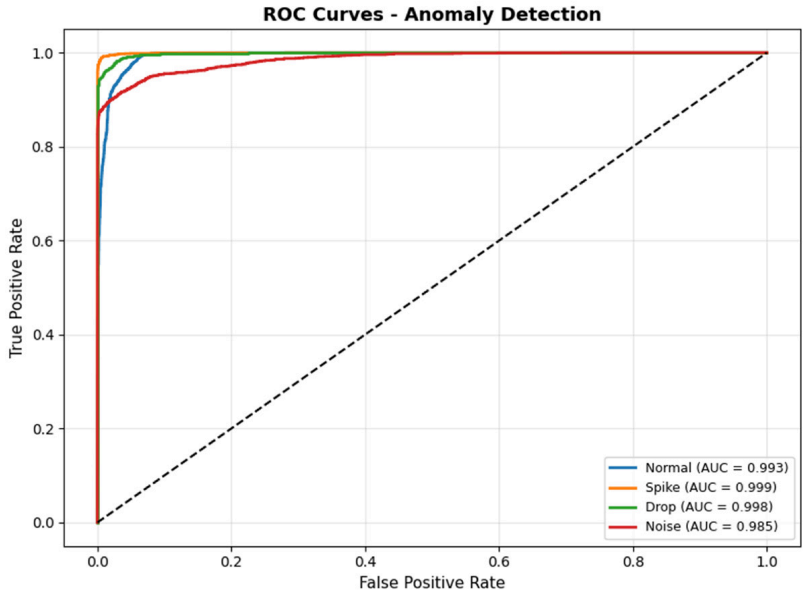


Figure 7. ROC curves for anomaly detection.

A similar pattern can be seen in the class-based results presented in Figure 8. Precision values are 98.85% for Normal, 98.90% for Spike, 98.24% for Drop, and 97.99% for Noise.

The recall values are 99.83%, 96.52%, 92.76%, and 84.90%, respectively, while the F1-scores are 99.34%, 97.70%, 95.43%, and 90.97%.

Overall, the model performs very well for the Normal and Spike classes. The Noise class is more difficult, which is not surprising because noise patterns are usually more irregular and less clearly separated than the other anomaly types. Even with that limitation, the results still show that the model can detect different anomaly classes in a stable and reliable way.



Figure 8. Class-wise performance for anomaly detection.

5.4. Advanced Analysis

To further investigate the effectiveness, robustness, and computational efficiency of the proposed framework, an extensive set of advanced analyses was conducted, including an ablation study, cross-validation, error analysis, normalized confusion matrix analysis, and computational complexity assessment.

5.4.1. Ablation Study

To better understand the role of each architectural component and establish the performance standing of the proposed framework within the contemporary NILM domain, an extensive comparative analysis was carried out. The proposed model was evaluated against three internal ablation variants (CNN, CNN + BiLSTM, and CNN + Transformer) as well as the classical Sequence-to-Point (Seq2Point) CNN architecture, which was adapted for multi-task outputs. The corresponding results are presented in Table 2.

Table 2. Ablation study results showing the contribution of CNN, BiLSTM, and Transformer components in terms of classification performance and computational cost.

Model	App Accuracy (%)	App Macro-F1 (%)	Anom Accuracy (%)	Anom Macro-F1 (%)	Params	Train Time (s)	Inference Time (ms)
CNN	99.72	99.30	97.87	92.29	43.4	172.29	10.31
Seq2Point CNN (Adapted)	98.45	97.20	96.15	90.40	112,480	210.15	4.80
CNN + BiLSTM	93.77	82.51	97.05	89.01	322,952	234.49	7.08
CNN + Transformer	99.41	98.54	97.95	92.07	208,904	149.21	5.42
CNN + BiLSTM + Transformer	99.72	99.29	98.56	94.86	653,064	773.00	7.08

The baseline CNN model already produces strong results, reaching 99.72% accuracy for appliance classification and 97.87% for anomaly detection. This suggests that convolutional layers are highly effective at capturing local patterns in power signals. When compared to the classical Seq2Point CNN benchmark, which achieves an anomaly macro-F1 of 90.40%, the proposed complete CNN + BiLSTM + Transformer model delivers the strongest overall performance. While baseline convolutional models (including Seq2Point) demonstrate acceptable competence in isolating local switching events, they inherently struggle to capture the long-term recursive operating cycles of complex appliances.

The internal ablation variants further highlight this dynamic. When the BiLSTM module is added to form the CNN + BiLSTM configuration, appliance classification performance drops noticeably, with accuracy decreasing to 93.77% and macro-F1 to 82.51%. This result suggests that temporal modeling by itself does not necessarily improve performance unless it is supported by sufficiently strong contextual representation. A different trend is observed when the Transformer module is combined with CNN. The CNN + Transformer model achieves 99.41% classification accuracy and 98.54% macro-F1 for appliance recognition, indicating that the attention mechanism is effective in capturing broader dependencies within electricity consumption signals.

The complete CNN + BiLSTM + Transformer model mitigates these limitations by successfully combining convolutional feature extraction, temporal sequence modeling, and attention-based learning. It achieves superior precision and macro-F1 scores, particularly in the highly challenging multi-class anomaly detection task (reaching 94.86% anomaly Macro-F1).

From a computational point of view, the full model is naturally the most demanding configuration, with 653,064 parameters and a training time of 773.00 s. Even so, its inference time remains low at approximately 7.08 ms per sample, which suggests that the model is still practical for real-time use. Overall, these findings confirm that the proposed triple-hybrid architecture offers the most effective balance between predictive power, robustness, and computational feasibility compared to standard baselines.

5.4.2. Cross-Validation Analysis

To examine the robustness and generalization ability of the proposed framework in a more reliable way, a 5-fold cross-validation procedure was applied. The fold-based results are reported in Table 3, and their distribution is further illustrated in Figure 9.

The results in Table 3 show that the model reaches an average appliance classification accuracy of 98.17%, while the standard deviation remains as low as 0.17. Such a small variation indicates that performance is highly consistent from one fold to another and that the learned representations remain effective on unseen data.

Table 3. Five-fold cross-validation performance of the proposed framework, reporting fold-wise results along with mean and standard deviation for appliance classification and anomaly detection.

Fold	Appliance Accuracy (%)	Anomaly Accuracy (%)
1	98.39	99.74
2	98.24	99.78
3	98.16	98.96
4	98.17	99.75
5	97.87	93.47
Mean	98.17	98.34
Std	0.17	2.46

For the anomaly detection task, the model attains a mean accuracy of 98.34%, with a slightly higher standard deviation of 2.46. This increased variability can be attributed to the inherently stochastic and diverse nature of anomaly patterns, which may differ significantly across folds. Notably, a performance drop is observed in Fold 5, as also reflected in Figure 9, where more challenging anomaly instances are present.

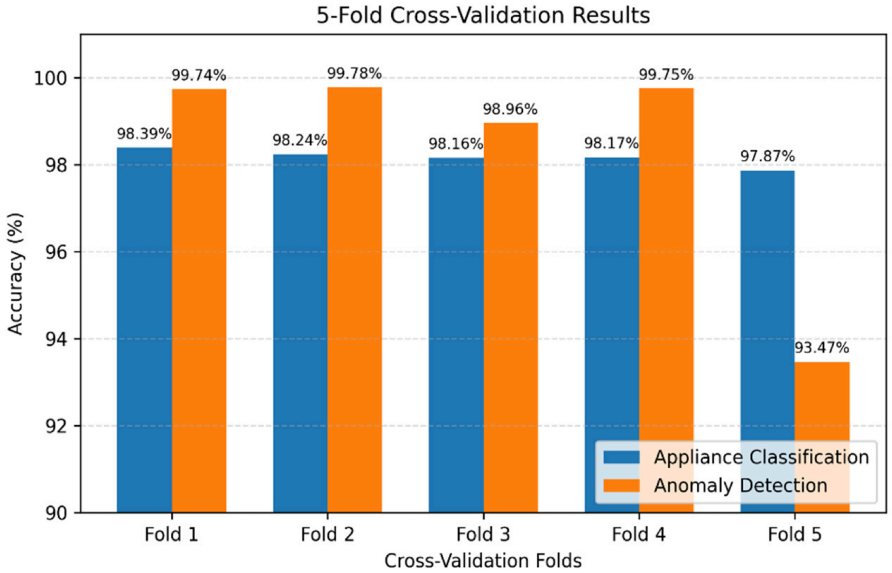


Figure 9. Five-fold cross-validation results for appliance classification and anomaly detection, demonstrating consistent performance across folds.

Despite this variability, the model maintains high accuracy across all folds, demonstrating strong generalization capability and robustness. Overall, these results confirm that the proposed framework is reliable and not sensitive to specific data partitions, which is essential for real-world NILM applications.

Furthermore, to evaluate the cross-dataset generalizability of the proposed framework and verify that the feature abstraction pipeline is not overfitted to a single smart meter infrastructure, an independent 5-fold cross-validation experiment was executed using the REDD dataset (Building 1). Without modifying the core model layout or tuning any hyperparameters optimized for UK-DALE, the framework achieved a highly robust mean appliance classification accuracy of 86.44% across the validation folds. This result statistically validates that the spatial–temporal representations extracted by the cascading architecture successfully capture invariant appliance power signatures across different geographical definitions. Conversely, the mean anomaly detection accuracy stabilized at 35.61%. This localized performance drop is an expected mathematical consequence of severe data starvation in the REDD subset. Due to the limited duration of active power cycles in this specific dataset, the synthetic anomaly injection yielded a heavily constrained absolute sample size per abnormal class during validation splits, leaving the data-hungry multi-head attention blocks with insufficient instances to map complex boundary contexts effectively.

5.4.3. Error Analysis

To gain a clearer understanding of where the proposed framework still struggles, an error analysis was carried out by focusing on the class pairs that were most frequently confused. The corresponding results are summarized in Table 4.

In the appliance classification task, the most noticeable confusion appears between microwave and kettle. In particular, 196 microwave samples are classified as kettle, whereas 34 kettle samples are predicted as microwave. A smaller number of errors is also seen for

washer dryer, which is sometimes confused with microwave, kettle, or fridge freezer. This suggests that the remaining errors are not spread evenly across all appliance classes, but are mostly concentrated in pairs that share partly similar signal behavior, especially short and high-power operating patterns.

A similar observation can be made for anomaly detection. The most frequent confusion occurs between noise and normal, with 521 noise samples being labeled as normal. In addition, 336 drop samples and 90 spike samples are also predicted as normal. This indicates that the model has more difficulty when anomaly patterns are weak, irregular, or not sufficiently distinct from normal operating behavior.

Taken together, these results show that the proposed model performs reliably in most situations. The remaining errors appear to arise mainly from intrinsic similarity between signal patterns and from gradual or ambiguous transitions, rather than from a general weakness in the framework. This is also in line with the normalized confusion matrix, which shows that misclassifications are limited to a relatively small number of challenging cases.

Table 4. Most frequently confused class pairs.

Task	True Class	Predicted Class	Count
Appliance Classification	Microwave	Kettle	196
Appliance Classification	Kettle	Microwave	34
Appliance Classification	Washer Dryer	Microwave	21
Appliance Classification	Washer Dryer	Kettle	5
Appliance Classification	Washer Dryer	Fridge Freezer	5
Anomaly Detection	Noise	Normal	521
Anomaly Detection	Drop	Normal	336
Anomaly Detection	Spike	Normal	90
Anomaly Detection	Noise	Drop	35
Anomaly Detection	Normal	Spike	19

5.4.4. Normalized Confusion Matrix Analysis

To better interpret the class-wise behavior of the proposed model, normalized confusion matrices for both appliance classification and anomaly detection are shown in Figure 10. Compared with standard confusion matrices, these plots make it easier to see the proportion of correct and incorrect predictions within each class and therefore provide a clearer view of how the model behaves across different categories.

As shown in Figure 10a, appliance classification is highly accurate for almost all classes. The model reaches 99% accuracy for kettle, 96% for microwave, and approximately 99–100% for washer dryer and fridge freezer. These results indicate that the proposed framework can distinguish appliance categories with a high level of reliability. The only noticeable confusion appears between kettle and microwave, where about 4% of microwave samples are predicted as kettle. This finding is consistent with the earlier results and is likely related to the similarity of their short-duration power consumption patterns.

The anomaly detection results in Figure 10b show a slightly more challenging picture. The model classifies the normal class perfectly (100%) and also performs strongly for the spike class (95%). However, the accuracy values are lower for the drop (83%) and noise (71%) classes. In particular, 27% of noise samples and 17% of drop samples are classified as normal. This suggests that weak or irregular disturbances are harder to separate from normal operating behavior.

Even with these challenging cases, the overall performance remains strong, and the observed errors are concentrated in only a few class pairs. This indicates that the proposed model is able to learn discriminative patterns for most categories, while the remaining mistakes are mainly associated with signal ambiguity rather than a general limitation of the framework.

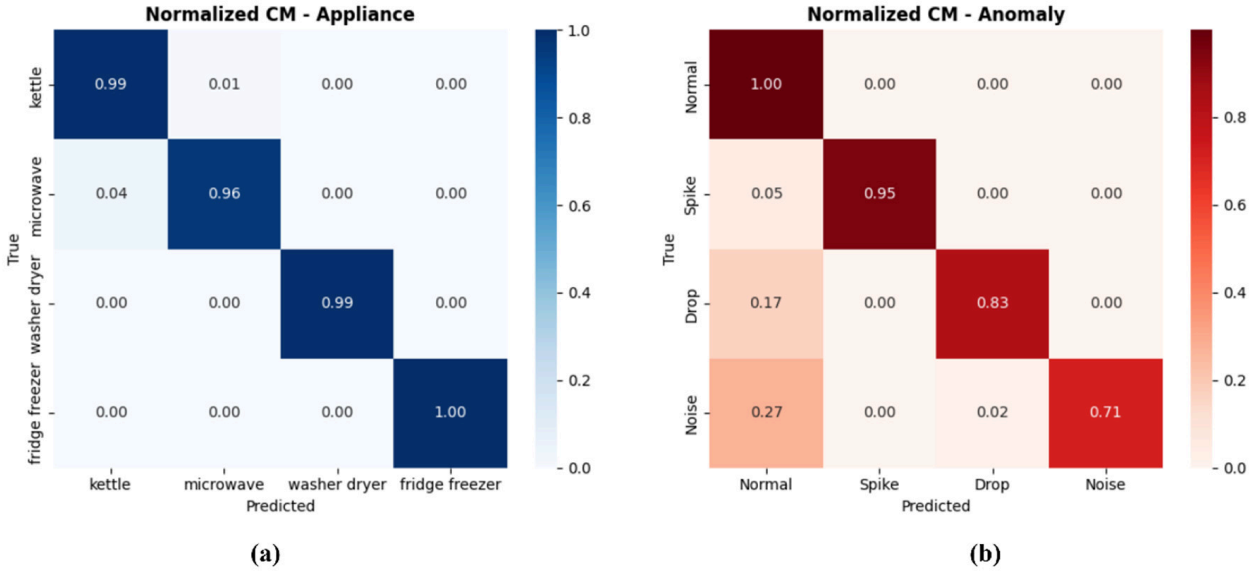


Figure 10. Normalized confusion matrices of the proposed model. (**a**) Appliance classification (**b**) Anomaly detection.

Overall, the normalized confusion matrices confirm that the proposed approach performs reliably in both tasks. The main weakness appears in distinguishing subtle anomaly types, which is a common difficulty in real-world NILM applications.

5.4.5. Computational Complexity Analysis

The computational aspect of the proposed framework was also examined in terms of model complexity and inference cost. Although the detailed numerical values are already given in Table 2, the trade-off between predictive performance and model complexity can be seen more clearly in Figure 11.

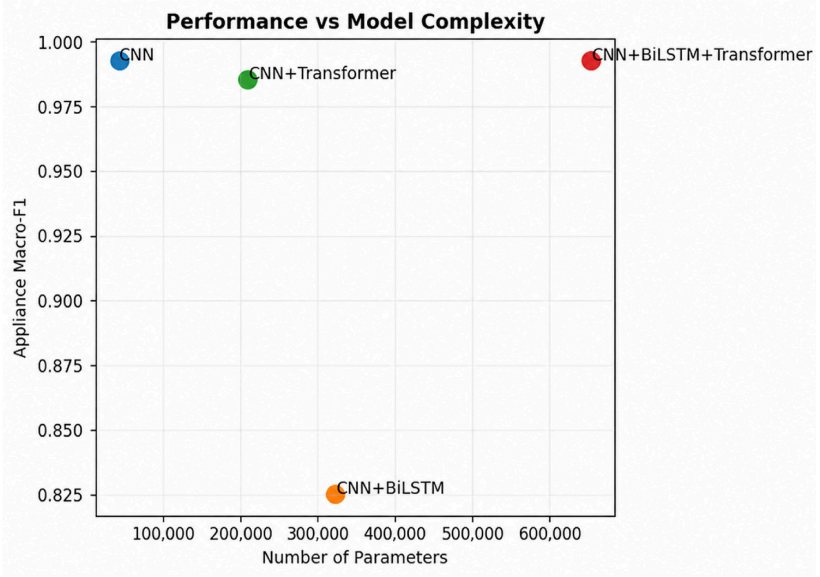


Figure 11. Performance versus model complexity trade-off across different model configurations.

As illustrated in Figure 11, the baseline CNN model has the smallest number of parameters, but it does not deliver the strongest anomaly detection performance. When the BiLSTM module is added, model complexity increases considerably, yet this increase is not matched by a comparable improvement in performance. This suggests that the CNN + BiLSTM configuration offers a less favorable balance between cost and benefit. By contrast, the CNN + Transformer model achieves strong predictive results with a more moderate increase in parameter count, making it a more computationally efficient alternative.

The full CNN + BiLSTM + Transformer model provides the best overall predictive performance, especially for anomaly detection, although this comes with a higher level of complexity. Even so, the inference time reported in Table 2 remains low enough for practical use. This indicates that the added computational burden is acceptable in view of the gains in classification performance and overall robustness.

Taken together, these results suggest that the proposed architecture offers the most effective balance between predictive power and computational feasibility. This makes it a suitable choice for intelligent NILM applications, where both accuracy and efficiency are important.

6. Discussion

The findings of this study show that the proposed multi-task triple-hybrid architecture is highly effective for jointly performing appliance classification and anomaly detection in NILM. By integrating 1D-CNN, BiLSTM, and Transformer Attention within a single framework, the model benefits from three complementary strengths: local transient feature extraction, bidirectional temporal modeling, and global contextual representation learning. These advantages are reflected in the experimental results, where the model achieved 99.48% accuracy for appliance classification and 98.80% accuracy for anomaly detection, together with consistently strong precision, recall, and F1-score values across classes.

The appliance classification results suggest that the proposed framework is particularly successful for appliances with distinct and stable consumption patterns, such as washer dryers and fridge freezers, for which almost perfect predictions were obtained. The remaining classification errors were concentrated mainly between kettle and microwave. This is consistent with a well-known challenge in NILM, namely the difficulty of separating appliances that produce partially overlapping, short-duration, high-power signatures. Recent review studies have similarly noted that NILM performance tends to decline when appliances exhibit similar electrical behavior or when training and evaluation settings vary substantially across datasets [1]. In this context, the strong results obtained here indicate that the proposed feature extraction pipeline is sufficiently discriminative for most classes, while the remaining errors are more likely related to intrinsic signal similarity than to a general limitation of the model.

The anomaly detection branch also produced strong results, although its error pattern was somewhat more challenging than that of appliance classification. In particular, the noise and drop classes were more often confused with the normal class. From an application perspective, this is a meaningful finding, since low-amplitude or irregular distortions do not always produce clearly separable signal patterns. Even so, the model preserved high overall anomaly detection accuracy together with strong AUC values, indicating that the multi-task framework was able to learn both appliance-related and anomaly-related features effectively. Prior work in multi-task NILM has shown that learning related tasks within a shared architecture can improve representation learning and reduce the deployment burden associated with separate models. For example, Dash and Sahoo [2] proposed a multi-task NILM framework that jointly addressed appliance state detection and energy disaggregation within a single deep model, emphasizing both feasibility and efficiency

on real-world datasets. Compared with that line of work, the present study extends the multi-task perspective toward the joint handling of appliance classification and anomaly detection, which is especially relevant for practical NILM systems that must remain aware of signal quality.

The ablation study offers additional insight into why the proposed architecture performs well. Although the CNN baseline already achieved strong performance, the results show clearly that the full CNN + BiLSTM + Transformer configuration provides the best overall balance, particularly in terms of anomaly-related macro-F1. This observation is in line with recent NILM research, where convolutional and attention-based structures are increasingly combined to exploit both local and global signal information. For instance, Azad et al. reported strong transformer-based NILM performance on UK-DALE, REDD, and REFIT by combining temporal pooling, residual connections, and attention mechanisms, while GTCN improved generalization on UK-DALE by integrating CNN layers with a gate-transformer structure [3]. Likewise, Huzzat et al. [4] showed that a GRU-BERT hybrid can improve NILM performance by combining temporal modeling with transformer-style contextual learning, especially for transient and complex appliance behaviors. The present results are consistent with these studies, but they also suggest that a triple-hybrid structure may offer additional advantages when the problem is extended beyond disaggregation alone to include anomaly-aware analysis. Beyond the primary dataset, testing the model on the REDD dataset provided strong evidence of its cross-dataset generalizability. Even without tuning the parameters for this new environment, the model maintained a highly robust 86.44% accuracy in appliance classification, proving that the CNN-BiLSTM-Transformer pipeline captures universal power signatures rather than memorizing local data. Furthermore, the comparison with the standard Seq2Point benchmark highlighted exactly why the triple-hybrid structure is necessary. A concise comparison with representative studies is provided in Table 5.

The cross-validation results further support the reliability of the proposed framework. The very low standard deviation observed in appliance classification indicates stable performance across different data splits, whereas the slightly higher variation in anomaly detection appears to be related mainly to the stochastic and heterogeneous nature of anomaly patterns. This is a reasonable outcome and does not weaken the overall robustness of the framework. In fact, recent NILM research has increasingly emphasized the importance of robustness, generalization, and sound evaluation methodology, particularly when models are tested under different datasets, unseen scenarios, or limited-data conditions [5].

From a broader perspective, the present study occupies a distinct position within the NILM literature. Many recent studies have focused on load disaggregation, unknown appliance detection, or transformer architectures designed to handle non-stationary behavior. For example, de Diego-Otón et al. [6] proposed a framework for classifying known appliances and detecting unseen ones under limited-data conditions using learned latent representations, while NILMFormer addressed subsequence non-stationarity in transformer-based load disaggregation and was presented as suitable for deployment-scale monitoring. Although these contributions are important, they are not directly equivalent to the problem addressed here. The main novelty of the present work lies in jointly addressing appliance classification and anomaly detection within a unified architecture, while also providing a comprehensive experimental evaluation through confusion analysis, cross-validation, ablation studies, multi-task assessment, and computational complexity analysis.

The computational analysis also offers a practically important result. Although the full architecture is more demanding than lightweight baselines in terms of parameter count and training cost, its inference time remains low enough for real-world use. This matters because NILM systems designed for smart energy monitoring should not only be accurate,

but also computationally feasible for deployment. The performance–complexity trade-off observed in this study suggests that the additional architectural depth is justified by the gains in predictive accuracy and robustness.

Table 5. Comparison of representative recent NILM studies with the proposed framework.

Study	Architecture/Focus	Main Task	Dataset/Scope	Relevance to This Study
Azad et al. (2024) [29]	Transformer-based attention, temporal pooling, residual connections	Appliance-level NILM disaggregation	UK-DALE, REDD, REFIT; seen and unseen cases	Strong attention-based NILM baseline; focuses on disaggregation rather than joint appliance classification and anomaly detection.
de Diego-Otón et al. (2025) [16]	Learned latent features for known and unseen appliance identification	Appliance classification + anomaly/unknown appliance awareness	WHITED; limited-data setting	Closest study in spirit, but focuses on known vs. unseen appliance recognition rather than signal anomaly classification.
Huzzat et al. (2025) [10]	GRU-BERT hybrid recurrent–attention model	Load disaggregation	NILM setting with hybrid recurrent-transformer modeling	Shows the benefit of combining temporal modeling and attention; the present work extends this idea with a CNN–BiLSTM–Transformer triple-hybrid design.
Dash and Sahoo (2024) [30]	Attention-powered multi-task deep learning framework	Multi-task NILM (state detection + disaggregation)	Two real-world datasets	Important multi-task NILM reference; unlike that work, the present study jointly performs appliance classification and anomaly detection.
Zai et al. (2023) [31]	Gate-Transformer + CNN	Energy disaggregation	UK-DALE	Demonstrates that convolution + transformer combinations improve generalization; the present study further integrates BiLSTM and anomaly-aware learning.
Petralia et al. (2025) [32]	Transformer architecture for non-stationary NILM	Energy disaggregation	Four real-world datasets; deployment-oriented evaluation	Addresses non-stationarity at scale; the present study differs by focusing on classification-oriented NILM and anomaly-aware joint learning.
Proposed study	1D-CNN + BiLSTM + Transformer Attention (triple-hybrid), dual-output	Appliance classification + anomaly detection	UK-DALE, Building 1; extensive performance and robustness analysis	Provides a unified, anomaly-aware NILM classification framework with strong predictive performance and practical computational feasibility.

Overall, the proposed framework compares favorably with recent NILM studies by combining strong predictive performance, anomaly-aware analysis, and multi-task efficiency within a single model. At the same time, the remaining errors—particularly those involving similar appliance signatures or subtle anomaly patterns—point to realistic directions for future work, including richer input representations, cross-household validation, and more adaptive anomaly modeling.

Furthermore, while the proposed hybrid model demonstrates robust classification capabilities in a window-based deployment framework, incorporating complementary methodological perspectives could further advance its real-world scalability. Integrating front-end event-based NILM algorithms represents a valuable strategy to optimize computational efficiency, as the deep learning architecture would only be triggered during detected transient states rather than processing continuous streaming inputs. Additionally, implementing Explainable AI (XAI) frameworks—such as SHAP (SHapley Additive exPlanations) or Integrated Gradients—could offer significant insights into model transparency by mapping out the exact feature attributions driving anomaly classifications. Finally, transitioning toward uncertainty-aware deep learning methods, such as Monte Carlo Dropout or Bayesian Neural Networks, would enable the network to quantify its own prediction confidence. This uncertainty quantification is essential for industrial smart grid applications, ensuring that false positives are minimized and highly ambiguous signal perturbations are handled with measurable statistical reliability.

7. Conclusions

This study proposed a multi-task triple-hybrid NILM framework that jointly performs appliance classification and anomaly detection from power signals. The proposed architecture combines 1D-CNN, BiLSTM, and Transformer Attention modules in a unified pipeline, allowing the model to capture local transient characteristics, bidirectional temporal dependencies, and globally discriminative feature representations. The dual-output design makes it possible to simultaneously classify appliances and identify signal anomalies, which increases the practical relevance of the framework for intelligent energy monitoring systems.

Tests on the main UK-DALE dataset showed the proposed model is highly effective, reaching 99.48% accuracy for appliance classification and 98.80% for anomaly detection. A secondary cross-dataset evaluation on the REDD dataset confirmed the architecture's strong ability to generalize, producing an 86.44% appliance recognition accuracy in an entirely different smart grid environment. The comparative benchmark analysis also proved that the triple-hybrid structure easily outperforms classical baselines like the Seq2Point model, especially when dealing with complex anomaly patterns. Furthermore, additional analyses, including the ablation study, cross-validation, error analysis, and computational complexity assessment, demonstrated that the proposed full architecture provides the most favorable balance between accuracy, robustness, and efficiency.

The main strengths of the proposed study can be summarized as follows. First, it introduces a unified framework that jointly addresses two practically relevant NILM tasks instead of handling them separately. Second, it integrates convolutional, recurrent, and attention-based mechanisms in a complementary manner, resulting in strong discriminative performance. Third, it offers a detailed evaluation protocol that goes beyond overall accuracy and includes robustness and complexity analyses, thereby strengthening the practical credibility of the model.

Despite these strengths, several limitations should also be acknowledged. The anomaly classes used in this study were generated synthetically, which, although useful for controlled evaluation, may not fully reflect the diversity of real-world anomaly patterns. Furthermore, while the cross-dataset evaluation on the REDD dataset provided strong evidence of the model's adaptability, testing across a wider variety of households and diverse smart grid infrastructures would further solidify these generalization claims. Finally, while the proposed architecture remains efficient at inference time, its training cost is higher than that of simpler baselines.

Future work may therefore focus on evaluating the model across a broader spectrum of international datasets, incorporating real anomaly annotations when available, and exploring adaptive or lightweight variants of the proposed architecture for large-scale real-time deployment. Overall, the findings of this study show that the proposed multi-task triple-hybrid framework is a reliable and effective solution for anomaly-aware NILM applications.

Author Contributions: Conceptualization, M.G. and Y.D.; methodology, M.G., Y.D. and M.Z.; software, M.G.; validation, Y.D. and M.Z.; formal analysis, M.G. and Y.D.; investigation, M.G. and M.Z.; writing—review and editing, M.G., Y.D. and M.Z.; visualization, M.G.; supervision, Y.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been accomplished with financial support by the European Regional Development Fund within the operational program "Bulgarian national recovery and resilience plan", procedure for direct provision of grants "Establishing of a network of research higher education institutions in Bulgaria", and under Project BG-RRP-2.004-0005 "Improving the research capacity and quality to achieve international recognition and resilience of TU-Sofia (IDEAS)".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset analyzed in this study are publicly available. The UK-DALE dataset is available at https://data.ceda.ac.uk/edc/d1/b3baa0aa-05ac-4ec8-b608-da5751758698 (accessed on 2 February 2026) [33].

Acknowledgments: This article was derived from the doctoral thesis of Mihriban Gunay, carried out at the Graduate School of Natural and Applied Sciences, Fırat University, Türkiye, in the Department of Electrical and Electronics Engineering. Mihriban Gunay's doctoral education was financially supported by the Scientific and Technological Research Council of Türkiye (TUBITAK) under the BIDEB 2211—Domestic Graduate Scholarship Programme. The authors gratefully acknowledge this support. During the preparation of this manuscript, the authors used GenIA tools such as Google Gemini (Gemini 3.5 Pro) and ChatGPT (GPT-5) to improve the clarity and fluency of certain sentences. These tools were used solely for linguistic refinement and improving expression clarity. The authors have carefully reviewed and edited the outputs from these tools and take full responsibility for the entire content of this publication.

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. Parson, O.; Ghosh, S.; Weal, M.; Rogers, A. Non-intrusive load monitoring using prior models of general appliance types. *Proc. AAAI Conf. Artif. Intell.* **2012**, *26*, 356–362. [CrossRef]
3. Medeiros, A.P.; Canha, L.N.; Bertineti, D.P.; de Azevedo, R.M. Event classification in non-intrusive load monitoring using convolutional neural network. In *Proceedings of the 2019 IEEE PES Innovative Smart Grid Technologies Conference-Latin America (ISGT Latin America)*; IEEE: Piscataway, NJ, USA, 2019; pp. 1–6.
4. De Diego-Oton, L.; Fuentes-Jimenez, D.; Hernandez, A.; Nieto, R. Recurrent LSTM architecture for appliance identification in non-intrusive load monitoring. In *Proceedings of the 2021 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*; IEEE: Piscataway, NJ, USA, 2021; pp. 1–6. [CrossRef]
5. Sykiotis, S.; Kaselimi, M.; Doulamis, A.; Doulamis, N. ELECTRIcity: An efficient transformer for non-intrusive load monitoring. *Sensors* **2022**, *22*, 2926. [CrossRef] [PubMed]
6. Nie, Z.; Yang, Y.; Xu, Q. An ensemble-policy non-intrusive load monitoring technique based entirely on deep feature-guided attention mechanism. *Energy Build.* **2022**, *273*, 112356. [CrossRef]
7. Zhang, C.; Zhong, M.; Wang, Z.; Goddard, N.; Sutton, C. Sequence-to-point learning with neural networks for non-intrusive load monitoring. *Proc. AAAI Conf. Artif. Intell.* **2018**, *32*, 2604–2611. [CrossRef]
8. Rashid, H.; Singh, P.; Stankovic, V.; Stankovic, L. Can non-intrusive load monitoring be used for identifying an appliance's anomalous behaviour? *Appl. Energy* **2019**, *238*, 796–805. [CrossRef]
9. Kelly, J.; Knottenbelt, W. Neural NILM: Deep neural networks applied to energy disaggregation. In *BuildSys 2015—Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*; Association for Computing Machinery, Inc.: New York, NY, USA, 2015; pp. 55–64. [CrossRef]
10. Huzzat, A.; Khwaja, A.S.; Alnoman, A.A.; Adhikari, B.; Anpalagan, A.; Woungang, I. GRU-BERT for NILM: A hybrid deep learning architecture for load disaggregation. *AI* **2025**, *6*, 238. [CrossRef]
11. He, G.; Huang, Y.; Zhang, Y.; Zhu, Y.; Leng, Y.; Shang, N.; Zeng, J.; Pu, Z. Hybrid transformer-convolutional neural network approach for non-intrusive load analysis in industrial processes. *Energies* **2025**, *18*, 2464. [CrossRef]
12. Ouzine, J.; Marzouq, M.; Dosse Bennani, S.; Lahrech, K.; El Fadili, H. New parallel hybrid PHCNN-GRU deep learning model for multi-output NILM disaggregation. *Energy Effic.* **2025**, *18*, 56. [CrossRef]
13. Shin, C.; Joo, S.; Yim, J.; Lee, H.; Moon, T.; Rhee, W. Subtask gated networks for non-intrusive load monitoring. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 1150–1157. [CrossRef]
14. Çimen, H.; Çetinkaya, N.; Vasquez, J.C.; Guerrero, J.M. A microgrid energy management system based on non-intrusive load monitoring via multitask learning. *IEEE Trans. Smart Grid* **2021**, *12*, 977–987. [CrossRef]
15. Dash, S.; Sahoo, N.C. Attention-based multitask probabilistic network for nonintrusive appliance load monitoring. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 2513412. [CrossRef]
16. de Diego-Otón, L.; Hernández, Á.; Fuentes, D.; Nieto, R.; Navarro, V.M. Architectural strategies for enhanced NILM classification and anomaly detection: Addressing limited data scenarios. *Expert Syst. Appl.* **2025**, *282*, 127756. [CrossRef]
17. Saha, D.; Bhattacharjee, A.; Chowdhury, D.; Hossain, E.; Islam, M.M. Comprehensive NILM framework: Device type classification and device activity status monitoring using capsule network. *IEEE Access* **2020**, *8*, 179995–180009. [CrossRef]

18. Hu, L.; Wei, J.; Yin, L. Convolutional Gated Power Prediction Combined with Multiscale Multilabel Classification for Nonintrusive Load Monitoring. *IEEE Trans. Instrum. Meas.* 2026, *in press*. [CrossRef]
19. Cheng, Y.; Zhong, Y. Non-intrusive load monitoring based on a combination of transformer and CNN. In *Proceedings of the 2024 4th International Conference on Energy, Power and Electrical Engineering (EPEE 2024)*; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2024; pp. 417–421. [CrossRef]
20. McLaughlin, S.; Holbert, B.; Fawaz, A.; Berthier, R.; Zonouz, S. A multi-sensor energy theft detection framework for advanced metering infrastructures. *IEEE J. Sel. Areas Commun.* **2013**, *31*, 1319–1330. [CrossRef]
21. Caruana, R. Multitask learning. *Mach. Learn.* **1997**, *28*, 41–75. [CrossRef]
22. Ismail Fawaz, H.; Forestier, G.; Weber, J.; Idoumghar, L.; Muller, P.A. Deep learning for time series classification: A review. *Data Min. Knowl. Discov.* **2019**, *33*, 917–963. [CrossRef]
23. Kiranyaz, S.; Avci, O.; Abdeljaber, O.; Ince, T.; Gabbouj, M.; Inman, D.J. 1D convolutional neural networks and applications: A survey. *Mech. Syst. Signal Process.* **2021**, *151*, 107398. [CrossRef]
24. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
25. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef]
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30*; Curran Associates, Inc.: Red Hook, NY, USA, 2017. [CrossRef]
27. 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]
28. Kolter, J.Z.; Johnson, M.J. REDD: A public data set for energy disaggregation research. In *Proceedings of the Workshop on Data Mining Applications in Sustainability (SIGKDD)*; Citeseer: San Diego, CA, USA, 2011; pp. 59–62. Available online: https://www.researchgate.net/publication/266597071_REDD_A_Public_Data_Set_for_Energy_Disaggregation_Research (accessed on 11 February 2026).
29. Irani Azad, M.; Rajabi, R.; Estebsari, A. Nonintrusive load monitoring (NILM) using a deep learning model with a transformer-based attention mechanism and temporal pooling. *Electronics* **2024**, *13*, 407. [CrossRef]
30. 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]
31. Zai, Z.; Zhao, S.; Zhang, Z.; Li, H.; Sun, N. Non-intrusive load monitoring based on the combination of gate-transformer and CNN. *Electronics* **2023**, *12*, 2824. [CrossRef]
32. Petralia, A.; Charpentier, P.; Kadhi, Y.; Palpanas, T. NILMFormer: Non-intrusive load monitoring that accounts for non-stationarity. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*; ACM: New York, NY, USA, 2025; pp. 4761–4772. [CrossRef]
33. UKERC EDC: Data. Available online: https://ukerc.rl.ac.uk/cgi-bin/dataDiscover.pl?Action=detail&dataid=b3baa0aa-05ac-4ec8-b608-da5751758698 (accessed on 2 February 2026).

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.