energies

MDPI

Article

Hybrid Transformer–Convolutional Neural Network Approach for Non-Intrusive Load Analysis in Industrial Processes

Gengsheng He [1,*], Yu Huang [2], Ying Zhang [2,3], Yuanzhe Zhu [1], Yuan Leng [1], Nan Shang [1], Jincan Zeng [1] and Zengxin Pu [2]

[1] Energy Development Research Institute, China Southern Power Grid, Guangzhou 510530, China; zhuyuanzhe@csg.cn (Y.Z.); mucunleng@sina.com (Y.L.); shangnan@csg.cn (N.S.); zengjc2@csg.cn (J.Z.)
[2] Electric Power Research Institute of Guizhou Power Grid Co., Ltd., Guizhou 550002, China; huangyu@im.gzsyy.csg (Y.H.); zhangying1007@gzsyy.csg.cn (Y.Z.); puzx0214@gz.csg.cn (Z.P.)
[3] College of Electrical Engineering, Guizhou University, Guizhou 550025, China
* Correspondence: hegs@csg.cn

Abstract: With global efforts intensifying towards achieving carbon neutrality, accurately monitoring and managing energy consumption in industrial sectors has become critical. Non-Intrusive Load Monitoring (NILM) technology presents a cost-effective solution for industrial energy management by decomposing aggregate power data into individual device-level information without extensive hardware requirements. However, existing NILM methods primarily tailored for residential applications struggle to capture complex inter-device correlations and production-dependent load dynamics prevalent in industrial environments, such as cement plants. This paper proposes a novel sequence-to-sequence-based non-intrusive load disaggregation method that integrates Convolutional Neural Networks (CNN) and Transformer architectures, specifically addressing the challenges of multi-device load disaggregation in industrial settings. An innovative time–application attention mechanism was integrated to effectively model long-term temporal dependencies and the collaborative operational relationships between industrial devices. Additionally, global constraints—including consistency, smoothness, and sparsity—were introduced into the loss function to ensure power conservation, reduce noise, and achieve precise zero-power predictions for inactive equipment. The proposed method was validated on real-world power consumption data collected from a cement production facility. Experimental results indicate that the proposed method significantly outperforms traditional NILM approaches with average improvements of 4.98%, 3.70%, and 4.38% in terms of accuracy, recall, and F1-score, respectively. These findings underscore its superior robustness in noisy conditions and under device fault conditions, further affirming its applicability and potential for deployment in industrial settings.

Keywords: Non-Intrusive Load Monitoring (NILM); Transformer; CNN; industrial electricity; cement plant

check for updates

Academic Editor: Tek Tjing Lie

Received: 9 April 2025
Revised: 6 May 2025
Accepted: 8 May 2025
Published: 11 May 2025

Citation: 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. https://doi.org/10.3390/en18102464

Copyright: © 2025 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 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Given the threat of global climate change, governments and enterprises worldwide are actively pursuing carbon peak and carbon neutrality targets. As the world's largest carbon emitter, China has articulated clear dual-carbon goals [1]. Industrial enterprises, as major consumers of energy consumption and sources of emissions, play a crucial role in achieving these objectives. Accurate identification and management of industrial power loads have therefore become increasingly important. The rapid advancement of smart grids

infrastructure and digital technologies has enabled more efficient collection and analysis of electricity consumption, facilitating load identification based on power usage profiles. This capability supports more accurate assessments of energy demand, enabling targeted conservation strategies and emissions reductions. By uncovering potential energy saving opportunities, load identification also informs the design of effective energy management policies, thereby supporting national dual-carbon goals.

Non-Intrusive Load Monitoring (NILM), which disaggregates device-level energy consumption from aggregate power data at a single point, offers a cost-effective and non-invasive solution for load analysis in industrial environments [2]. Compared to intrusive methods that require extensive sensor deployment, NILM significantly reduces hardware cost and implementation complexity, making it highly attractive for large-scale industrial applications.

However, industrial load profiles differ substantially from those in residential and commercial settings. In addition to exhibiting complex temporal dynamics, and inter-device dependencies, industrial load possess distinctive electrical characteristics. These include significant harmonic distortion caused by motor-driven equipment, frequent voltage transients due to high inertia machinery switching, and the prevalence of multiphase power systems. Such factors introduce nonlinearity and noise into power signals, complicating device-level disaggregation. Furthermore, industrial facilities often feature tightly coupled production processes in which multiple devices operate collaboratively, leading to high correlations in load profiles that are difficult to disentangle. For example, in cement plants, equipment such as ball mills, rotary kilns, and crushers exhibit abrupt and interdependent power variations driven by process control logic.

The growing integration of renewable energy sources such as wind and solar into industrial operations introduces significant variability and uncertainty in power demand. As highlighted in [3], the intermittent and weather-dependent nature of renewable generation causes stochastic fluctuations in net electricity consumption, which can obscure the true operating characteristics of individual devices. These fluctuations often appear as transient spikes, irregular oscillations, or baseline shifts in power signals, all of which compromise signal clarity and increase ambiguity in NILM processing. In addition, industrial systems adopting renewable integration commonly implement advanced energy management strategies—such as demand response and real-time power dispatch—that introduce externally driven variability, further complicating the temporal patterns of load behavior. Without effective modeling of these uncertainties, NILM algorithms may experience substantial degradation in both accuracy and robustness.

To tackle these challenges, this paper proposes an advanced Hybrid Transformer–CNN framework designed specifically for industrial NILM applications. The model leverages Convolutional Neural Networks (CNNs) to extract localized temporal features while employing Transformer architectures to capture global temporal dependencies and device interactions. A key innovation is the introduction of a time–application attention mechanism which enhances the model's ability to learn long-range dependencies and interpret collaborative device behaviors under complex operational dynamics. In addition, global constraints—enforcing power consistency, temporal smoothness, and activation sparsity—were incorporated into the loss function to mitigate disaggregation noise, ensure energy conservation, and detect inactive states accurately even under distorted and volatile load conditions. Experimental validation using real-world cement plant data demonstrates the model's superior accuracy and robustness, highlighting its potential applicability to other industrial domains with similarly complex load characteristics.

2. Related Work

Intrusive Load Monitoring requires the installation of individual sensors for each monitored device, necessitating high installation and maintenance costs. To address this issue, Non-Intrusive Load Monitoring (NILM) technology has emerged as an alternative method, requiring only the measurement of aggregated data and using decomposition algorithms to estimate the power consumption of individual devices. NILM technology is not only widely applied in home energy management but also plays a significant role in smart grids, industrial energy management [4], and fault diagnosis [5]. By providing real-time monitoring at the device level, NILM effectively improves energy utilization, helps users optimize their electricity usage behavior, and reduces energy waste. For instance, in smart grids, NILM can facilitate load scheduling [6] and demand response [7], thereby enhancing grid stability and reliability. Additionally, by monitoring load signals to quickly identify signs of equipment failure, NILM provides unique advantages in fault detection and predictive maintenance, offering data support for maintenance decisions [8].

With the continuous advancement of artificial intelligence, modern NILM technologies predominantly rely on machine learning methods. Numerous studies have explored the application of different deep learning algorithms in NILM, such as Transformer [9], Convolutional Neural Networks (CNN) [10], and Long Short-Term Memory networks (LSTM) [11]. These models exhibit powerful capabilities in multi device load disaggregation and real-time monitoring, especially in handling time-series data [12]. The core approach involves extracting features from power data and training machine learning models using these features. Research indicates that CNNs are highly effective at extracting local features from power signals, whereas Recurrent Neural Networks (RNN) and LSTMs excel at processing time-series data and capturing long term dependencies, significantly improving load identification and disaggregation accuracy [13]. Among these, self-attention mechanisms [14] and Trans-former architectures [15,16] offer substantial advantages in processing long time-series data, becoming a focus of current research. However, machine learning approaches are typically data-driven, thus requiring large datasets for learning or training and validation with real-world datasets to ensure the validity and reproducibility of results [17]. Some algorithms relying on high- quality labeled datasets may face challenges such as incomplete labels or data imbalance. Additionally, the majority of current research focuses on residential and small-scale industrial environments [18].

NILM has demonstrated substantial effectiveness in residential and small commercial sectors, enhancing energy management, facilitating demand response, and aiding fault diagnostics. However, its adoption in large-scale industrial environments, such as cement plants and steel mills, has been limited due to unique industrial load characteristics. Specifically, industrial loads exhibit complex temporal patterns, frequent interactions between various equipment, and operational states closely linked to production processes. Traditional NILM methodologies typically utilize CNNs and RNN-based architectures, focusing predominantly on independent device behaviors and straightforward temporal dependencies observed in residential scenarios. These methods are thus inadequate for capturing intricate, production-driven dependencies among multiple industrial devices, leading to suboptimal disaggregation accuracy and limited robustness under noisy conditions and operational anomalies [19].

In recent years, research trends have increasingly focused on integrating the Transformer architecture with other model structures to more effectively capture long-term dependencies in time series data. For instance, a study proposed a hybrid model—TransUNet—that combines the Transformer and U-Net architectures for application in Non-Intrusive Load Monitoring (NILM) tasks [20]. This approach converts one-dimensional power time series into a two-dimensional matrix format through a sliding-window technique to meet

the input requirements of the model. The original TransUNet architecture was further simplified to better adapt to low-dimensional electricity consumption data, thereby improving computational efficiency and reducing the risk of overfitting [21].

In terms of model performance optimization, several studies have introduced techniques such as Bayesian optimization for hyperparameter tuning, aiming to enhance the generalization capability and operational efficiency of Transformer-based models [22]. For example, in a hydropower generation forecasting task in Sichuan Province, the model achieved an average relative error of only 9.8% on the test set after optimal hyperparameter configurations were identified through Bayesian optimization, demonstrating the practical potential of Transformers in real-world engineering applications.

However, despite the promising results achieved in residential and small-scale industrial scenarios, the deployment of these advanced methods in large-scale industrial facilities—such as cement plants and steel mills—still faces significant challenges. On one hand, the model architectures are often complex and parameter-heavy, leading to high computational costs that do not align with the real-time monitoring requirements of industrial environments. On the other hand, most existing approaches fail to adequately account for the intricate equipment coupling characteristics and production-driven dynamic behaviors inherent in industrial loads [23]. Therefore, there remains a notable research gap in the development of NILM methodologies tailored to large-scale industrial settings, calling for further investigation into model lightweighting, robustness enhancement, and improved industrial adaptability.

This paper addresses these identified limitations by proposing a Hybrid Transformer-CNN framework tailored explicitly to large industrial settings, effectively modeling both local and global load interactions while maintaining computational efficiency. By integrating an innovative time–application attention mechanism and introducing constraints on consistency, smoothness, and sparsity, the proposed approach addresses the challenges of device interactions, production variability, and real-time application requirements, offering a promising direction for industrial NILM solutions.

3. Materials and Methods

3.1. NILM Method

Non-Intrusive Load Monitoring seeks to determine the energy consumption of individual target devices from aggregated datasets by utilizing various optimization techniques, signal processing methods, and machine learning or deep learning algorithms.

We assumed that aggregated samples from a smart meter were obtained, containing T time points, represented as $P_{\text{agg}} = \{P_{\text{agg1}}, P_{\text{agg2}}, P_{\text{agg3}}, \ldots, P_{\text{aggT}}\}$. The power consumption of device n could be represented as $P_n = \{P_{n1}, P_{n2}, P_{n3}, \ldots, P_{nT}\}$. The aggregated value at time t could then be represented as:

$$P_{\text{agg}}(t) = \sum_{n=1}^{N} s_n(t) \cdot p_n(t) + e(t) \tag{1}$$

where $s_n(t)$ represents the on/off state of device n at time t, $p_n(t)$ represents the power consumption of device n at time t, N is the total number of devices, and $e(t)$ is the error term.

The production process of a cement plant encompasses several critical stages, including raw material processing, calcination, grinding, and finished product packaging. After raw materials undergo crushing, they are adjusted and homogenized by a dual-homogenization speed system to ensure proper mixing ratios. The fine-ground raw meal produced by the raw mill is then fed into the pre-decomposition furnace. Post-decomposition, the material completes high-temperature calcination in the rotary kiln to form clinker. This calcination

process includes combustion and cooling phases. The clinker is further processed through a hot rolling machine before being ground into cement in the cement mill. Finally, the cement products go through packaging and bagging for distribution. In the context of cement production, Non-Intrusive Load Monitoring (NILM) technology is deeply integrated to achieve comprehensive monitoring of the plant's load status and optimize energy consumption. As shown in Figure 1, smart meters installed at the factory entrance collect total load data in real-time and transmit this information to a data center via wireless networks. Based on the overall current and voltage waveforms, NILM algorithms can parse out unique operational characteristics of high-energy-consuming equipment such as the raw mill, decomposition furnace, and cement mill, accurately identifying their start–stop conditions, operation modes, and real-time energy consumption.

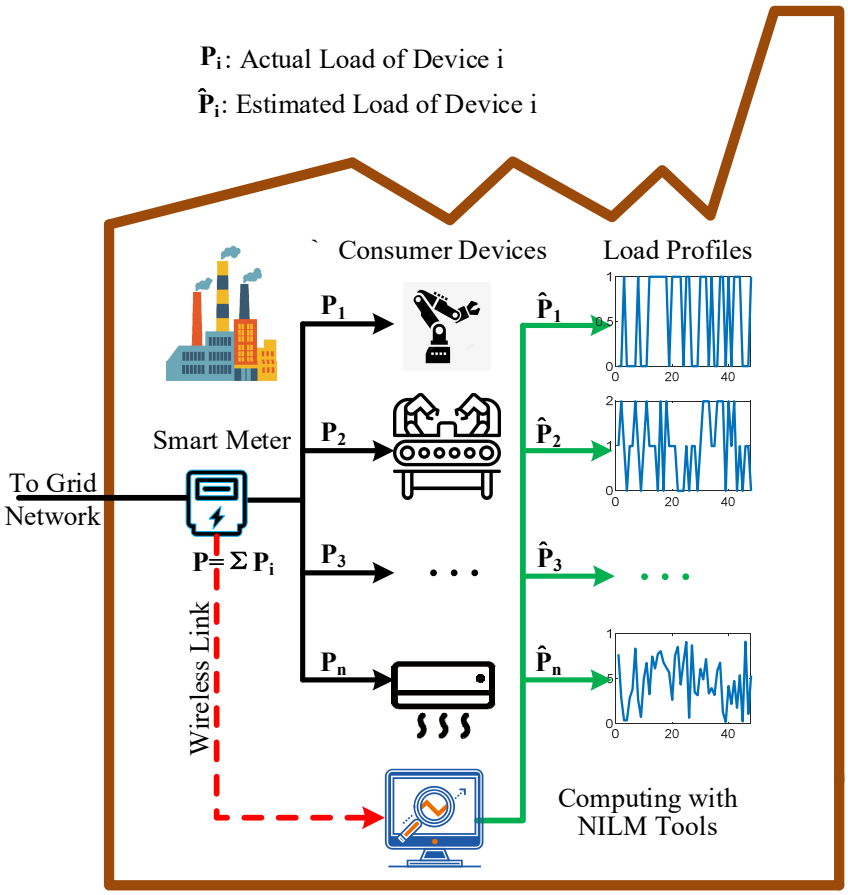


Figure 1. Framework of NILM for devices' state monitoring and load profiles decomposition in industrial facilities.

To comprehensively evaluate the performance of NILM algorithms, researchers have adopted the F-measure (*F1-score*) as an evaluation metric [24]. The *F1-score* represents the harmonic mean of precision (*Pre*) and recall (*Rec*), which are metrics derived from information retrieval. *Precision* indicates the proportion of correctly identified states among all identified states, while recall measures the proportion of correctly identified states among all actual states. By integrating these two metrics, the *F1-score* offers a more thorough assessment of an algorithm's accuracy in recognizing the operational states of electrical appliances.

$$\text{F1}-\text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2)$$

where

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

TP: The number of times the algorithm correctly identifies a device as being in the "on" state.

FP: The number of times a device is actually "off" but the algorithm incorrectly identifies it as "on."

FN: The number of times a device is actually "on" but the algorithm incorrectly identifies it as "off".

3.2. Multi-Device Hybrid Load Disaggregation Model Architecture

The multi-device load disaggregation network model proposed in this paper is illustrated in Figure 2. The network architecture comprised an encoder and multiple parallel decoders, enabling simultaneous load disaggregation for several devices.

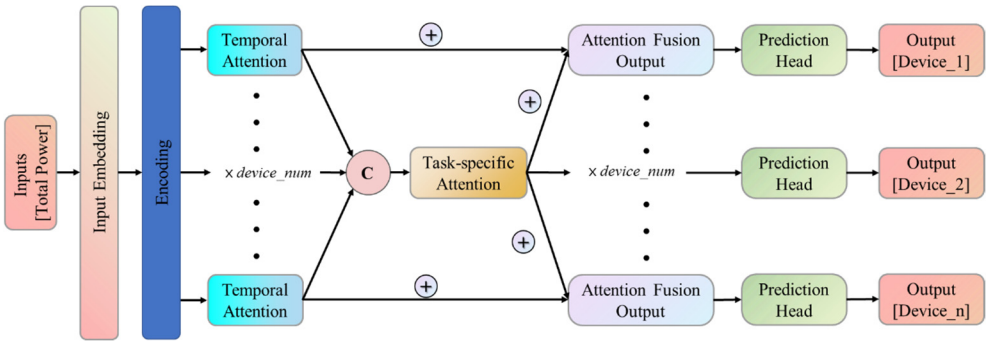


Figure 2. Overall architecture of the multi-device load disaggregation network model.

In the encoder, convolutional layers followed by ReLU activation functions were sequentially stacked to comprehensively extract features from the input data. The extracted features were then distributed to the parallel decoders, each corresponding to a specific device, to obtain device-specific disaggregated power feature representations. Notably, the number of output branches from the encoder aligned with the number of target devices, with identical data features being propagated to each branch.

Within the decoders, a time and application attention module was first employed to derive a fused feature representation that integrated both temporal information and multi-device interactions. Subsequently, the fused features were fed into the final prediction head module, which generated the multi-device disaggregated power outputs. The prediction head module utilized a fully connected network to produce the final disaggregation results.

In the encoder module, the model extracted features from the input load sequence through a Convolutional Neural Network (CNN) composed of multiple convolutional layers. Originally proposed by LeCun et al. for image recognition tasks, CNNs are designed to capture local dependencies and hierarchical patterns through local receptive fields, weight sharing, and layered feature abstraction. In recent years, CNNs have been widely adopted in the field of energy management, particularly in load monitoring tasks, due to their powerful capability to model temporal structures in time-series data. Specifically, 1D convolutional layers are employed to detect localized patterns, abrupt changes, and periodic behaviors in power consumption signals.

Each convolutional layer was followed by a Rectified Linear Unit (ReLU) activation function, defined as ReLU (x) = max (0, x). Compared with traditional activation functions such as Sigmoid and Tanh, ReLU helped mitigate the vanishing gradient problem and introduces nonlinearity into the model, enhancing its ability to capture complex features. This property is especially beneficial for modeling nonlinear transitions common in power signals, such as sudden device activation or shutdown. Moreover, ReLU promotes sparse activation, which improves training efficiency and generalization performance.

The encoder was constructed by stacking multiple convolution-ReLU blocks, with each layer extracting progressively higher-level features. The final output of the encoder was a deep feature representation that was evenly distributed to multiple parallel decoder modules, each of which corresponded to a specific appliance and was responsible for learning its individual load decomposition representation.

Assuming the total load time series $P_{total} \in R^{T\times 1}$ was input into the encoder, which was composed of multiple convolutional layers and ReLU activation functions stacked together, the convolution operation can be expressed as:

$$Z_{encoded} = ReLU(W_1 * P_{total} + b_1) \tag{5}$$

Among these, W_1 is the convolution kernel, $*$ denotes the convolution operation, b_1 is the bias term, and $Z_{encoded} \in R^{T\times F}$ is the feature output by the encoder, with F being the feature dimension. Feature extraction is mainly achieved through convolution operations and the ReLU activation function. These are nonlinear mappings that can relate to the inherent latent features within the data, preventing overfitting and enhancing the generalization ability of the model.

Multiple convolutional layers of the encoder each extracted features at different levels, which were further stacked to form:

$$Z^{(l)}_{encoded} = ReLU\left(W_l * Z^{(l-1)}_{encoded} + b_l\right) \tag{6}$$

Among these, l represents the number of convolutional layers, and $Z^{(l)}_{encoded}$ is the output of the l layer.

The total load was distributed to the parallel decoders of each device based on the features generated by the encoder. Assuming we have N devices, the features output by the encoder would be allocated to the decoders of each device:

$$Z^{i}_{device} = Z^{(l)}_{encoded} for i = 1, 2, \ldots, N \tag{7}$$

Here, $Z^{i}_{device} \in R^{T\times F}$ is the input feature for the i-th device.

However, in NILM tasks, the decomposed power consumption of different devices is interdependent due to the constraint of total power, and the power consumption of each device exhibits temporal correlation. To address these challenges, this paper proposed a temporal application attention module that effectively integrated the interdependencies among devices and captured the temporal correlations of power consumption. As shown in Figure 3, the temporal/application attention module contained a residual multi-head attention mechanism. The residual multi-head attention module in the temporal attention component utilized device features from the corresponding branch decoder to enhance the sensitivity of data feature sequences to specific time steps. This allowed the model to capture both local and global dependencies, improving disaggregation accuracy and robustness.

The process can be described mathematically as:

$$L_{output} = Norm(L_{in} + Dropout(MH(L_{in}))) \tag{8}$$

Multi-head attention is a powerful mechanism for exploring complex relationships between input elements. It learns contextual feature relationships in the input data through self-attention, thereby improving model performance. The self-attention mechanism calculates the attention weights for each position in the input data sequence by multiplying the query (Q) and key (K), then multiplying the result by the value (V) to emphasize the important features. However, self-attention mechanisms can lead to overfocusing on position-specific features due to the reliance on its own feature information. Therefore, multi-head attention utilizes multiple self-attention heads to capture dependencies from different ranges, thereby uncovering the intrinsic features of the data and enhancing the model's learning ability. Specifically, multi-head attention is formulated as:

$$\text{self} - \text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{9}$$

$$\text{head}_i = \text{self}-\text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{10}$$

$$\text{Multi} - \text{Head}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_n)W^O \tag{11}$$

where Q, K, V are the data sequences input into the multi-head attention mechanism, and W is the learnable weight matrix.

In the final disaggregation output prediction head, as shown in Figure 3, the model linearly connects (similar to an artificial neural network) and activates the output of each branch decoder to obtain the final output.

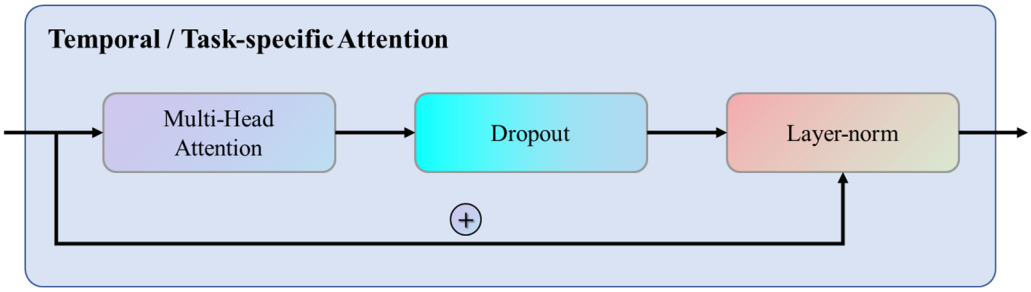


Figure 3. Architecture of the time–application attention module.

In the decoder, the temporal and application attention mechanism were first applied to enhance the feature representation. According to Equation (8), the output of the temporal and application attention module can be expressed as:

$$Z_{attention}^i = \tau_2\left(\tau_1\left(ReLU\left(Z_{device}^i\right)\right) + Z_{device}^i\right) \tag{12}$$

Among these, τ_1 and τ_2 are the linear fully connected operations of the first layer and the second layer, respectively.

The fused features are then fed into the prediction head module, which is a fully connected network used to generate the power disaggregation results for each device. For the i-th device, the output can be expressed as:

$$\hat{P}_{device,i} = W_{fc} \cdot Z_{attention}^i + b_{fc} \tag{13}$$

Among these, $\hat{P}_{device,i} \in R$ is the power disaggregation prediction value for the i-th device, while W_{fc} and b_{fc} are the weights and bias of the fully connected layer, respectively. For all N devices, the final output is an $N \times T$ matrix, where each row corresponds to the power disaggregation result for a device:

$$\hat{P} = \left[\hat{P}_{device,1}, \hat{P}_{device,2}, \cdots \hat{P}_{device,N}\right]^T \tag{14}$$

3.3. Improved Loss Function

Finally, the multi-device load disaggregation model proposed in this paper was built using an end-to-end approach. It took the total aggregated power P_{total} as input and output the disaggregated power values $\{\hat{P}_{device,i}\}$ for each device. The process is mathematically described as follows:

$$P_{total}(t) = \sum_{i=1}^{N} \hat{P}_{device,i}(t) + \hat{e}(t) \tag{15}$$

where $\hat{P}_{device,i}(t)$ represents the power consumption of device i, N is the total number of devices, and $\hat{e}(t)$ is the error term, which accounts for noise or unmodeled factors.

The disaggregation for each device is given by:

$$\hat{P}_{device,i} = f_i(P_{total}, \Theta), i = 1, 2, \ldots, N \tag{16}$$

where $f_i(*)$ is the mapping function of the i-th decoder branch, and Θ represents the model's learnable parameters.

The proposed model considers inter-device dependencies, but Equation (12) treats each device independently, potentially missing global constraints like total power conservation. To address this, we introduced a global loss term to enforce the sum of all $\hat{P}_{device,i}(t)$ to match $P_{total}(t)$ during training. Then, the global consistency loss was denoted as

$$L_{consistency} = \left| P_{total}(t) - \sum_{i=1}^{N} \hat{P}_{device,i}(t) \right| \tag{17}$$

To ensure smoothness and sparsity in the device outputs, we added a regularization loss term. The smoothness loss ensured that the power consumption of each device $\hat{P}_{device,i}(t)$ would change gradually over time unless there were explicit device state transitions. It was formulated as:

$$L_{smooth} = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=2}^{T} \left(\hat{P}_{device,i}(t) - \hat{P}_{device,i}(t-1)\right)^2 \tag{18}$$

The sparsity loss encouraged the model to predict zero power consumption for devices that were likely off, which aligns with real-world NILM scenarios where some devices are off at any given time. This loss was defined as:

$$L_{sparsity} = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \left|\hat{P}_{device,i}(t)\right| \tag{19}$$

Finally, the overall loss function is

$$L = L_{MSE} + \lambda_1 L_{consistency} + \lambda_2 L_{smooth} + \lambda_3 L_{sparsity} \tag{20}$$

where L_{MSE} is the mean squared error (MSE) for disaggregation accuracy, $\lambda_1, \lambda_2, \lambda_3 \in [0, 1]$, with $\lambda_1 + \lambda_2 + \lambda_3 = 1$. These regularization terms were balanced to ensure the model's robustness and accuracy.

As shown in Equation (1), $s_n(t)$ is theoretically necessary to identify whether a device is active, it is often not directly available from the aggregated data. We use a learnable representation that predicts device activation states as part of the disaggregation process.

$$\hat{s}_n(t) = \sigma(g_n(P_{total}, \Theta)), s_n(t) \approx \hat{s}_n(t)\# \quad (21)$$

where σ is the sigmoid function, and $g_n(*)$ predicts activation probabilities.

4. Case Study

4.1. The Cement Plant'S Power Data

This chapter utilizes real electrical consumption data from a cement plant's power data collection project. The data collection period spanned 42 days, with 480 data points collected daily for 12 difference devices listed in Table 1.

Table 1. Cement plant equipment list.

Device	Name	Device	Name
1	Cement mill cycle fan 1	7	High-temperature fan
2	Cement mill cycle fan 2	8	Kiln tail exhaust fan
3	Cement mill main motor 1	9	Raw material cycle fan
4	Cement mill main motor 2	10	Moving roller Motor
5	Kiln head exhaust fan	11	Fixed roller motor
6	Coal mill main motor	12	Limestone crusher main motor

During this period, data acquisition was impacted by noise interference, equipment failures, and shutdowns, leading to missing or erroneous entries. To enhance data quality and ensure suitability for subsequent model testing, we performed data preprocessing, which included imputing missing values, correcting anomalies, and applying smoothing techniques. The resulting preprocessed energy consumption data for the cement plant is presented in Figure 4. Subsequently, the dataset was partitioned, with the first 37 days designated as training data and the final 5 days reserved for testing.

4.2. Model Parameter Configuration and Hyperparameter Selection

The proposed model was implemented using the PyTorch 2.1.0 deep learning framework. PyTorch is an open-source library developed by Facebook's AI Research lab, widely used for building and training neural networks due to its dynamic computation graph, intuitive programming interface, and strong support for GPU acceleration. In this study, PyTorch facilitated rapid prototyping and flexible model design, enabling efficient implementation of the convolutional encoder-decoder architecture and seamless integration of customized layers and loss functions tailored for the Non-Intrusive Load Monitoring (NILM) task. Key hyperparameters and their configurations are detailed below:

(1) Regularization Coefficients:

The regularization coefficients in the multi-constrained loss function were $\lambda_1 = 0.61$, $\lambda_2 = 0.29$, $\lambda_3 = 0.10$. The coefficients were optimized via grid search on the validation set, with search ranges $\lambda_1 \in [0.4, 0.8]$, $\lambda_2 \in [0.1, 0.4]$, and $\lambda_3 \in [0.05, 0.2]$. The combination $\lambda_1 = 0.61$, $\lambda_2 = 0.29$, $\lambda_3 = 0.10$ maximized the F1-score while satisfying $\lambda_1 + \lambda_2 + \lambda_3 = 1$.

(2) Network Architecture Parameters:

Encoder: Four-layer causal convolutions with channel dimensions [64, 128, 256, 512] Kernel width: 7, stride: 2.

Attention Module: Eight-head self-attention mechanism, hidden dimension: 512, dropout rate: 0.1.

Decoder: Three-layer transposed convolutions with channel dimensions [256, 128, 64], kernel width: 5.

(3) Training Parameters:

Optimizer: Adam with initial learning rate 3e-4 and cosine annealing scheduling.
Batch Size: 256 sequences/batch, each containing 480-time steps (1-day sampling).
Training Epochs: 300 epochs with early stopping (patience = 20 epochs).
Input Normalization: Robust Scaler (median-centered, scaled by interquartile range).

(4) Parameter Sensitivity Analysis:

As illustrated in Figure 5, the regularization coefficients exhibited distinct impacts: λ_1 primarily governed total power conservation errors. When $\lambda_1 > 0.5$, the average root mean square error (RMSE) for all 12 devices dropped to approximately 5.

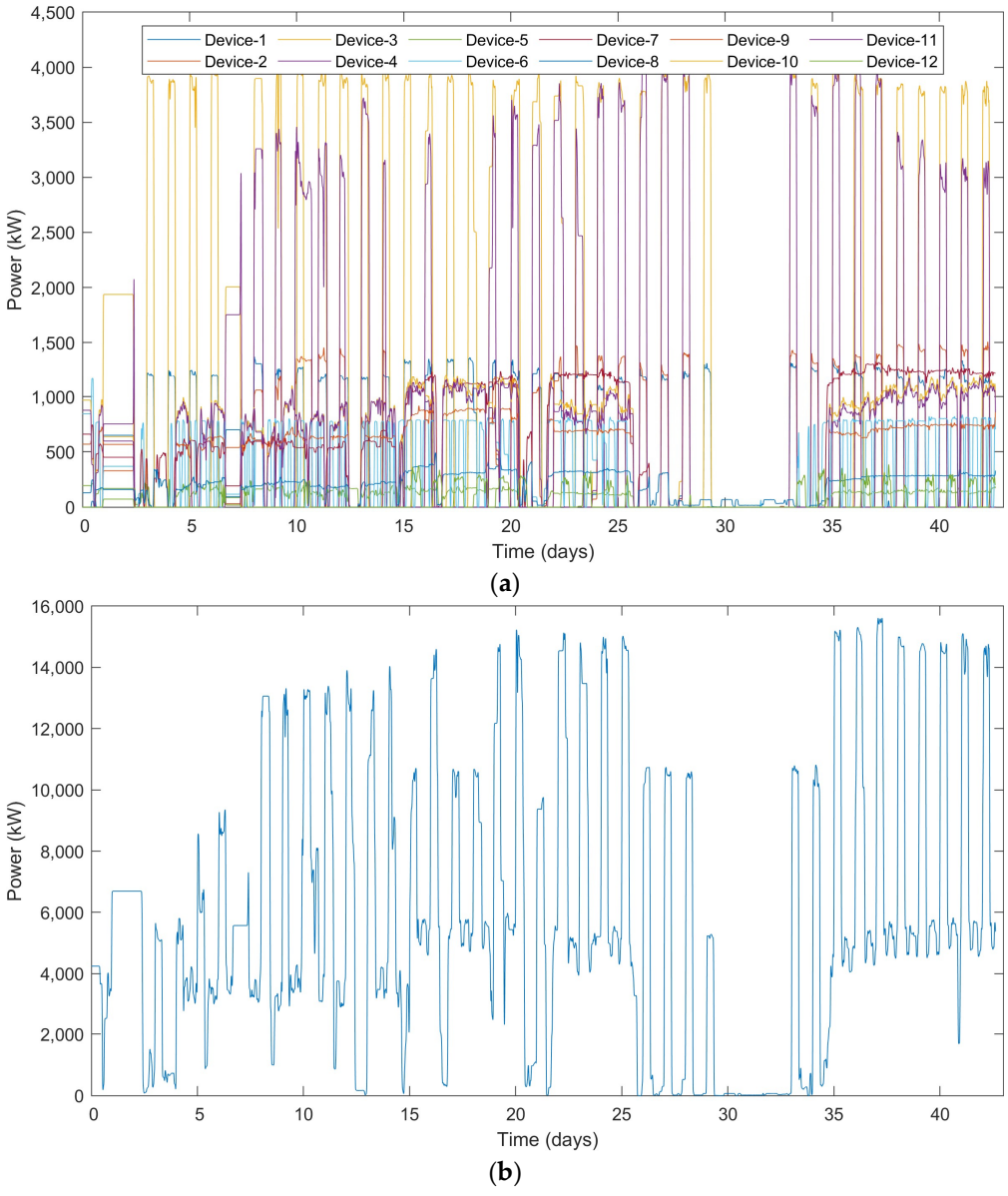


Figure 4. Preprocessed energy consumption data of the cement plant over 42 days: (**a**) real-time energy consumption profiles for each of the 12 devices; (**b**) aggregated total energy consumption across all 12 devices.

λ_2 optimally suppressed power fluctuations within the range [0.2, 0.4]. This configuration achieved a 37.6% reduction in fluctuations, quantified by differences between adjacent time steps.

λ_3 balanced sparsity and detection accuracy at a value of 0.1. This setting reduced false positives to 4.2% while maintaining sensitivity to low-power device signals.

Hence, in Cement Production Scenarios: (1) High λ_1 (0.61): mitigates industrial measurement noise through strict power conservation, enhancing physical plausibility. (2) Moderate λ_2 (0.29): balances transient responses during equipment start-up/shutdown with steady-state operation requirements. (3) Low λ_3 (0.10): preserves weak standby power signals while maintaining sparsity, critical for industrial equipment idle states.

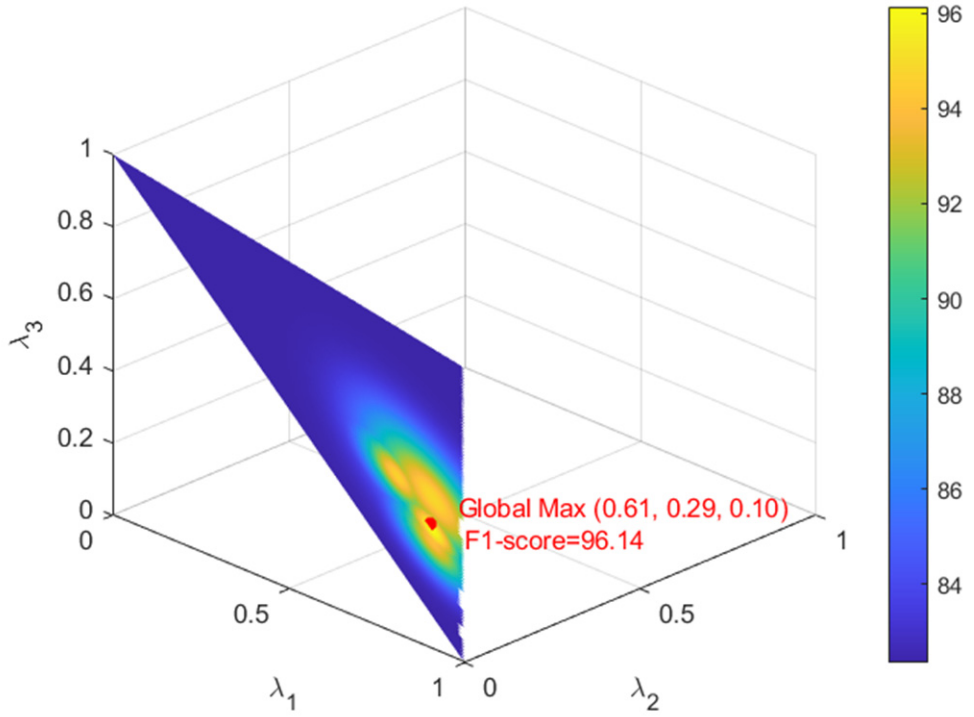


Figure 5. Hypothetical parameter sensitivity surface (global max point at $\lambda_1 = 0.61$, $\lambda_2 = 0.29$, $\lambda_3 = 0.10$, F1-score = 96.14).

4.3. Results

We conducted comparative experiments using several advanced NILM models, including GRU+ [25], LSTM+ [26], seq2seq CNN [27], and TCN+ [28], to evaluate the improvements achieved by the proposed model. GRU+ and LSTM+ refer to NILM models based on gated recurrent unit and long short-term memory networks, respectively, which are capable of modeling temporal dependencies in power consumption sequences. The seq2seq CNN model employs a sequence-to-sequence convolutional architecture to capture complex temporal features and has also demonstrated effectiveness in load disaggregation tasks. Temporal Convolutional Networks (TCN) models leverage stacks of dilated, causal 1D convolutional layers to efficiently capture long-range temporal dependencies in aggregate power signals—enabling fast, parallelizable inference and improved appliance-level load disaggregation accuracy in NILM tasks. These models were configured with the same input length and trained under similar settings until convergence. The analysis results of all models on the cement plant dataset are presented in Table 2.

In Table 2, the best performance results are marked in bold, and Figure 6 illustrates the comparison of F1-scores across different models on various devices. As shown in Table 1 and Figure 5, our model outperformed GRU+, LSTM+, CNN, and TCN+ baselines. It achieves average improvements of 4.98% in precision, 3.70% in recall, and 4.38% in F1-score. These results demonstrate its superior performance across multiple devices. Particularly, for the high-temperature fan and cement mill main motor 1, the spatiotemporal dual-stream feature extraction architecture integrating Transformer with CNN proposed in this paper captured local power fluctuations during the start–stop phases of cement plant equipment through CNN, while employing a hierarchical temporal attention mechanism to model dynamic collaboration relationships among devices. This design allowed the model to achieve the maximum increase in precision for the high-temperature fan and in recall and F1-score for the cement mill main motor 1, with improvements of 7.54%, 8.75%, and 6.52% respectively, validating its accurate modeling of collaborative operational characteristics of the equipment.

Table 2. Experimental results (the best results are marked in bold).

Device	Model	Precision	Recall	F1-Score	Device	Model	Precision	Recall	F1-Score
1	GRU+	88.57%	97.69%	92.91%	7	GRU+	88.73%	92.15%	90.41%
	LSTM+	82.24%	87.80%	84.93%		LSTM+	92.44%	93.58%	93.01%
	CNN	76.44%	89.44%	82.43%		CNN	87.62%	88.98%	88.29%
	TCN+	88.57%	97.69%	92.91%		TCN+	88.75%	92.15%	90.42%
	Ours	**93.93%**	**99.38%**	**96.58%**		Ours	**99.98%**	**95.92%**	**97.92%**
2	GRU+	98.49%	91.02%	94.61%	8	GRU+	90.25%	88.68%	89.46%
	LSTM+	85.77%	92.79%	89.14%		LSTM+	97.44%	97.87%	97.66%
	CNN	96.69%	94.32%	95.49%		CNN	98.55%	**98.19%**	98.37%
	TCN+	92.24%	**97.80%**	94.93%		TCN+	94.86%	92.07%	93.44%
	Ours	**99.19%**	97.71%	**98.44%**		Ours	**99.97%**	96.93%	**98.44%**
3	GRU+	98.08%	86.67%	92.03%	9	GRU+	97.16%	97.25%	97.20%
	LSTM+	94.84%	89.11%	91.88%		LSTM+	94.73%	94.33%	94.53%
	CNN	98.86%	82.69%	90.06%		CNN	97.66%	97.60%	97.63%
	TCN+	96.69%	94.32%	95.49%		TCN+	95.58%	98.37%	96.95%
	Ours	**99.26%**	**97.86%**	**98.55%**		Ours	**99.75%**	**99.19%**	**99.47%**
4	GRU+	98.99%	90.52%	94.57%	10	GRU+	95.38%	96.14%	95.76%
	LSTM+	88.08%	93.73%	90.82%		LSTM+	95.19%	93.06%	94.11%
	CNN	99.54%	88.81%	93.87%		CNN	97.62%	97.09%	**97.36%**
	TCN+	93.58%	88.68%	91.07%		TCN+	91.10%	**99.07%**	94.91%
	Ours	**93.07%**	**99.21%**	**96.04%**		Ours	**99.99%**	94.00%	96.90%
5	GRU+	**94.18%**	92.32%	93.24%	11	GRU+	90.36%	92.19%	91.27%
	LSTM+	91.05%	92.91%	91.97%		LSTM+	96.16%	**95.69%**	95.92%
	CNN	91.01%	90.64%	90.82%		CNN	95.59%	95.23%	95.41%
	TCN+	94.84%	89.11%	91.88%		TCN+	**99.99%**	84.20%	91.42%
	Ours	90.79%	**96.95%**	**93.77%**		Ours	99.96%	94.05%	**96.91%**
6	GRU+	82.97%	86.62%	84.76%	12	GRU+	86.75%	85.51%	86.13%
	LSTM+	86.75%	85.51%	86.13%		LSTM+	88.82%	83.34%	86.00%
	CNN	88.82%	83.34%	86.00%		CNN	79.40%	**92.54%**	85.47%
	TCN+	88.86%	82.69%	85.06%		TCN+	82.65%	**90.70%**	86.11%
	Ours	**89.00%**	**88.69%**	**88.84%**		Ours	**99.64%**	85.08%	**91.78%**

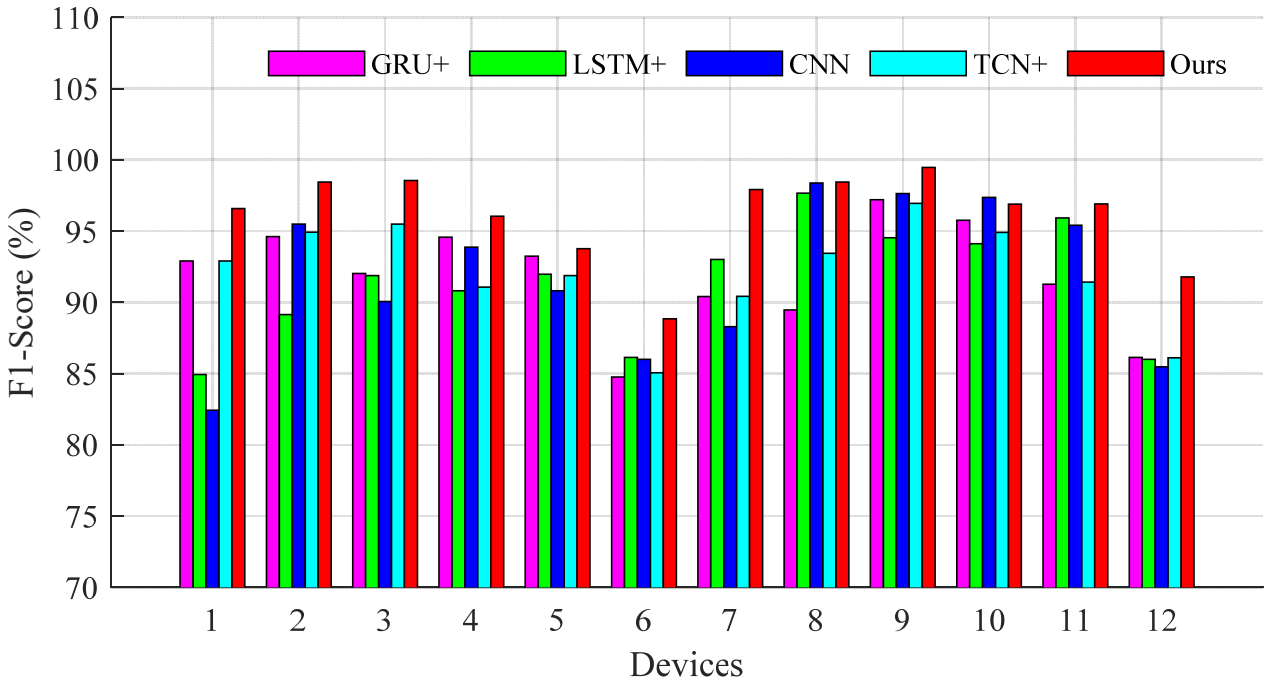


Figure 6. Bar chart comparing F1-scores of different models across various devices.

Additionally, the multi-constraint joint optimization strategy, through power conservation consistency constraints and device sparsity regularization terms, maintained an advantage in F1-score even when there was a 7.46% decrease in recall for the limestone crusher main motor, indicating strong robustness against noise interference. Combined with cross-device comparisons of F1-scores in Figure 5, the proposed model demonstrated an average increase of 4.38% in F1-score under high-noise scenarios in cement plants. The output smoothness constraint effectively alleviated abnormal power jitter, improving the accuracy of zero-power identification for inactive devices by 4.98%, confirming the universal improvement of multi-constraint strategies for industrial-grade load disaggregation.

The prediction results of the Ours model are shown in Figure 7, where the blue line represents the actual power curve, and the red dashed line indicates the predicted values. For periodic devices (such as the cement mill recirculation fan and the cement mill main motor), the model accurately captured the periodic start-stop patterns through the hierarchical temporal attention mechanism of the spatiotemporal dual-stream architecture, leading to a high degree of alignment between the predicted curve and the actual values. For non-periodic and continuously variable power devices (such as the kiln head exhaust fan, high-temperature fan, and kiln tail exhaust fan), the model effectively tracked the continuously changing load demands by combining local power gradient features extracted by CNN with power conservation consistency constraints. However, for variable-period devices (such as the coal mill main motor, raw meal recirculation fan, roller press moving/fixing rollers, and limestone crusher main motor), the model faced significant challenges. These devices have strong randomness in their operation cycles and weak time correlation (as seen in the intermittent power surges of the coal mill main motor shown in Figure 6), making accurate predictions more difficult.

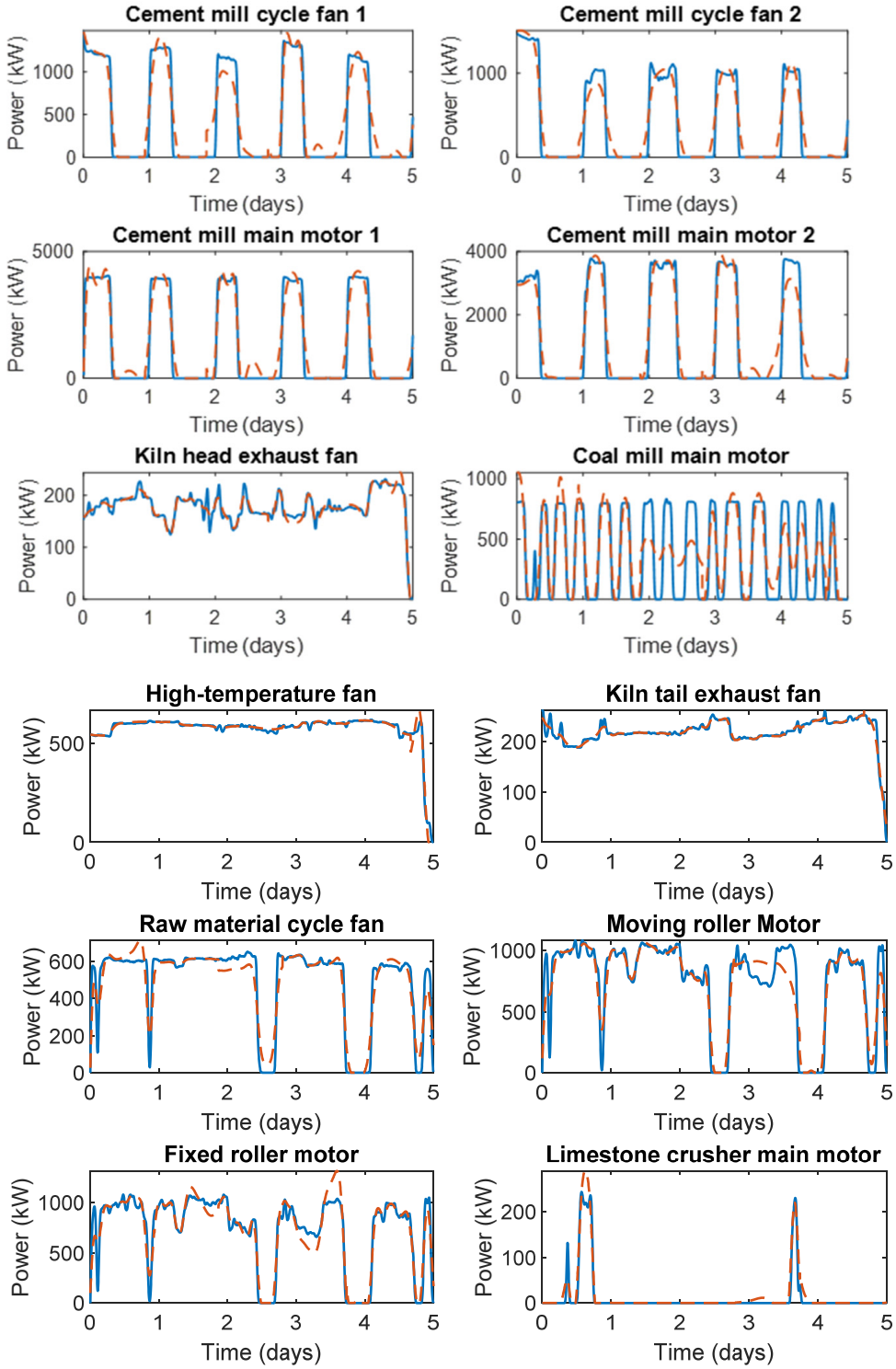


Figure 7. Five-day prediction results of the proposed model on cement plant data. The blue line represents the actual values, and the red dashed line indicates the predicted values.

5. Conclusions

In this study, a Non-Intrusive Load Monitoring (NILM) method tailored for high-noise industrial environments such as cement plants was proposed. By integrating Transformer with CNN in a hybrid architecture, the model aimed to overcome the limitations of traditional models. Innovatively, it established a spatiotemporal dual-stream feature extraction mechanism that leveraged the self-attention layers of Transformer to analyze dynamic collaboration relationships between devices and capture long-term power dependencies

across devices. Meanwhile, it utilized deep convolution kernels of CNNs to extract local spatiotemporal patterns, enhancing the modeling capability of short-term load features. Experimental results demonstrated that the proposed model achieved average improvements of 4.98%, 3.70%, and 4.38% on precision, recall, and F1-score respectively, compared to benchmark models like GRU+, LSTM+, CNN, and TCN+. This method not only effectively decomposed individual device energy consumption but also exhibited strong robustness in addressing key challenges in industrial NILM applications, such as noise interference and equipment load coupling.

Despite these encouraging results, this work is currently limited to data collected from a single cement plant, which may constrain the generalizability of the model to other industrial sectors. Furthermore, although public industrial NILM datasets are scarce and vary significantly in sampling protocols and device taxonomies, future work will focus on extending this method to additional facilities, such as steel mills and chemical plants, to validate its performance in cross-industry scenarios. We also plan to integrate new industrial datasets as they become available and explore transfer learning strategies to improve generalization. These efforts aim to enhance the scalability and practical applicability of NILM technology in broader industrial contexts.

Author Contributions: Conceptualization, G.H. and N.S.; methodology, Y.Z. (Ying Zhang) and Y.Z. (Yuanzhe Zhu); software, J.Z.; validation, Z.P.; formal analysis, Y.L.; data curation, Y.H.; writing—original draft preparation, G.H.; writing—review and editing, Y.Z. (Yuanzhe Zhu). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the China Southern Power Grid Corporation Technology Project (066600KK52222044, GZKJXM20222165).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Conflicts of Interest: Authors Gengsheng He, Yuanzhe Zhu, Yuan Leng, Nan Shang and Jincan Zeng were employed by the company Energy Development Research Institute, China Southern Power Grid. Authors Yu Huang, Ying Zhang and Zengxin Pu were employed by the company Electric Power Research Institute of Guizhou Power Grid Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Zhao, X.; Ma, X.; Chen, B.; Shang, Y.; Song, M. Challenges toward carbon neutrality in China: Strategies and countermeasures. *Resour. Conserv. Recycl.* **2022**, *176*, 105959. [CrossRef]
2. Devlin, M.; Hayes, B.P. Non-Intrusive Load Monitoring using Electricity Smart Meter Data: A Deep Learning Approach. In Proceedings of the 2019 IEEE Power & Energy Society General Meeting (PESGM), Atlanta, GA, USA, 4–8 August 2019. [CrossRef]
3. Shayan, M.E.; Petrollese, M.; Rouhani, S.H.; Mobayen, S.; Zhilenkov, A.; Su, C.L. An innovative two-stage machine learning-based adaptive robust unit commitment strategy for addressing uncertainty in renewable energy systems. *Int. J. Electr. Power Energy Syst.* **2024**, *160*, 110087. [CrossRef]
4. Liu, Q.; Kamoto, K.M.; Liu, X.; Sun, M.; Linge, N. Low-Complexity Non-Intrusive Load Monitoring Using Unsupervised Learning and Generalized Appliance Models. *IEEE Trans. Consum. Electron.* **2019**, *65*, 28–37. [CrossRef]
5. Lindahl, P.A.; Green, D.H.; Bredariol, G.; Aboulian, A.; Donnal, J.S.; Leeb, S.B. Shipboard Fault Detection Through Nonintrusive Load Monitoring: A Case Study. *IEEE Sens. J.* **2018**, *18*, 8986–8995. [CrossRef]
6. Lin, Y.H.; Tsai, M.S. An Advanced Home Energy Management System Facilitated by Nonintrusive Load Monitoring with Automated Multiobjective Power Scheduling. *IEEE Trans. Smart Grid* **2015**, *6*, 1839–1851. [CrossRef]
7. Rehman, A.U.; Lie, T.T.; Vallès, B.; Rahman, T.S. Non-invasive load-shed authentication model for demand response applications assisted by event-based non-intrusive load monitoring. *Energy AI* **2021**, *3*, 100055. [CrossRef]
8. Schirmer, P.A.; Mporas, I. Non-Intrusive Load Monitoring: A Review. *IEEE Trans. Smart Grid* **2023**, *14*, 769–784. [CrossRef]
9. Varanasi, L.N.S.; Karri, S.P.K. STNILM: Switch Transformer based Non-Intrusive Load Monitoring for short and long duration appliances. *Sustain. Energy Grids Netw.* **2024**, *37*, 101246. [CrossRef]

10. Zhou, X.; Feng, J.; Li, Y. Non-intrusive load decomposition based on CNN–LSTM hybrid deep learning model. *Energy Rep.* **2021**, *7*, 5762–5771. [CrossRef]
11. Hwang, H.; Kang, S. Nonintrusive Load Monitoring Using an LSTM With Feedback Structure. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–11. [CrossRef]
12. Xia, M.; Liu, W.A.; Wang, K.; Song, W.; Chen, C.; Li, Y. Non-intrusive load disaggregation based on composite deep long short-term memory network. *Expert Syst. Appl.* **2020**, *160*, 113669. [CrossRef]
13. Dash, S.; Sahoo, N.C. Electric energy disaggregation via non-intrusive load monitoring: A state-of-the-art systematic review. *Electr. Power Syst. Res.* **2022**, *213*, 108673. [CrossRef]
14. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. Advances in Neural Information Processing Systems. *arXiv* **2017**, arXiv:1706.03762. [CrossRef]
15. Xiong, J.; Hong, T.; Zhao, D.; Zhang, Y. MATNilm: Multi-Appliance-Task Non-Intrusive Load Monitoring with Limited Labeled Data. *IEEE Trans. Ind. Inform.* **2024**, *20*, 3177–3187. [CrossRef]
16. Wang, Y.; Huo, C.; Xu, F.; Zheng, L.; Hao, L. Ultra-short-term distributed photovoltaic power probabilistic forecasting method based on federated learning and joint probability distribution modeling. *Energies* **2025**, *18*, 197. [CrossRef]
17. Iqbal, H.K.; Malik, F.H.; Muhammad, A.; Qureshi, M.A.; Abbasi, M.N.; Chishti, A.R. A critical review of state-of-the-art non-intrusive load monitoring datasets. *Electr. Power Syst. Res.* **2021**, *192*, 106921. [CrossRef]
18. Ruano, A.; Hernandez, A.; Ureña, J.; Ruano, M.; Garcia, J. NILM Techniques for Intelligent Home Energy Management and Ambient Assisted Living: A Review. *Energies* **2019**, *12*, 2203. [CrossRef]
19. Angelos, G.F.; Timplalexis, C.; Krinidis, S.; Ioannidis, D.; Tzovaras, D. NILM applications: Literature review of learning approaches, recent developments and challenges. *Energy Build.* **2022**, *261*, 111951. [CrossRef]
20. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
21. Chen, J.; Mei, J.; Li, X.; Lu, Y.; Yu, Q.; Wei, Q.; Luo, X.; Xie, Y.; Adeli, E.; Wang, Y.; et al. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Med. Image Anal.* **2024**, *97*, 103280. [CrossRef]
22. Xiao, Y.; Shao, H.; Wang, J.; Yan, S.; Liu, B. Bayesian variational transformer: A generalizable model for rotating machinery fault diagnosis. *Mech. Syst. Signal Process.* **2024**, *207*, 110936. [CrossRef]
23. Arbel, Y.; Beck, Y. Advances in non-intrusive load monitoring for the industrial domain: Challenges, insights, and path forward. Renewable and Sustainable Energy. *Reviews* **2025**, *210*, 115136.
24. Wang, S.; Xiang, X.; Zhang, J.; Liang, Z.; Li, S.; Zhong, P.; Zeng, J.; Wang, C. A multi-task spatiotemporal graph neural network for transient stability and state prediction in power systems. *Energies* **2025**, *18*, 1531. [CrossRef]
25. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv* **2014**, arXiv:1406.1078. [CrossRef]
26. Rafiq, H.; Zhang, H.X.; Li, H.M.; Ochani, M. Regularized LSTM Based Deep Learning Model: First Step towards Real-Time Non-Intrusive Load Monitoring. In Proceedings of the 6th IEEE International Conference on Smart Energy Grid Engineering, Uoit, Oshawa, ON, Canada, 12–15 August 2018. [CrossRef]
27. Zhang, C.; Zhong, M.; Wang, Z.; Goddard, N.; Goddard, N. Sequence-to-point learning with neural networks for nonintrusive load monitoring. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32. [CrossRef]
28. Qian, Y.; Yang, Q.; Li, D.; An, D.; Zhou, S. An Improved Temporal Convolutional Network for Non-intrusive Load Monitoring. In Proceeding of the 33rd Chinese Control and Decision Conference (CCDC), Kunming, China, 22–24 May 2021; pp. 2557–2562. [CrossRef]

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.