# Non-Intrusive Load Monitoring

Chapter: NILM

Submitted on: June 07, 2025

# 1 Introduction

Non-Intrusive Load Monitoring (NILM) is a pivotal technique in energy management, designed to decompose the aggregate energy consumption of a system into the individual contributions of its appliances using data from a single metering point. This approach eliminates the need for multiple sensors, making it cost-effective and scalable for applications in smart grids, energy efficiency, and demand-side management. NILM provides detailed insights into appliance-level energy usage, enabling homeowners and facility managers to optimize consumption, reduce costs, and contribute to sustainability goals. In this chapter, we adapt the TransNILM model, a bidirectional transformer-based architecture, for energy disaggregation using the REDD dataset. This study explores the methodology, implementation, and results of applying this advanced deep learning model to achieve accurate appliance-level energy predictions. We begin by reviewing related works, followed by a detailed methodology, experimental results, and a discussion of findings.

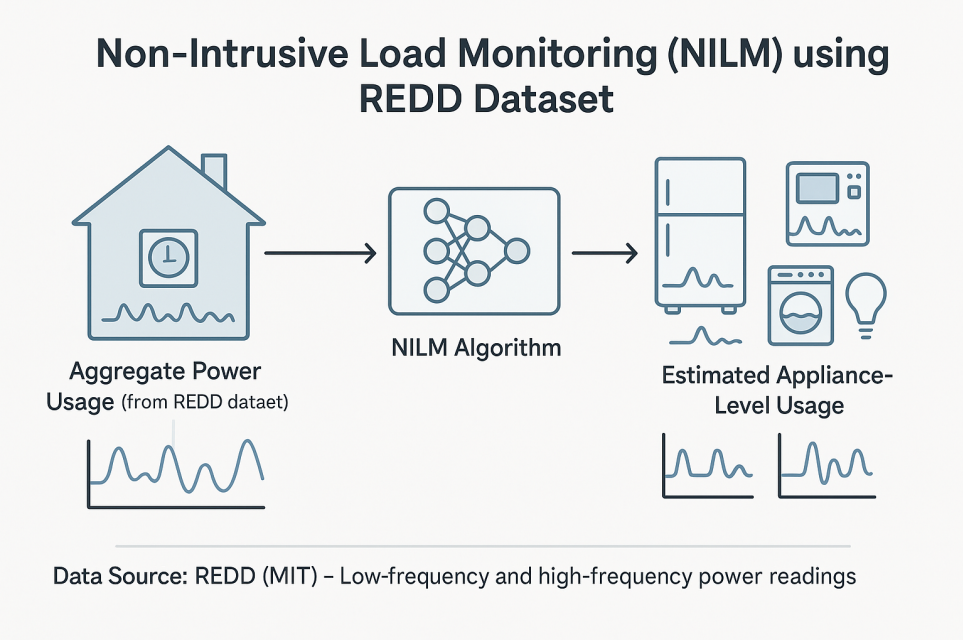


Figure 1: Conceptual diagram of Non-Intrusive Load Monitoring, illustrating the decomposition of aggregate energy into individual appliance contributions

# 2 Related Works

The field of Non-Intrusive Load Monitoring has evolved significantly over the past decades. Early approaches relied on statistical methods, such as Hidden Markov Models, and signal processing techniques to identify appliance signatures from aggregate power data. However, these methods often struggled with overlapping consumption patterns and noise. The advent of deep learning has transformed NILM, with recurrent neural networks (RNNs) like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models capturing temporal dependencies in energy consumption data [1]. Convolutional neural networks (CNNs) have also been effective in extracting spatial features from sequential energy data, improving disaggregation accuracy [2].

More recently, transformer-based models, originally developed for natural language processing, have shown promise in other domains. The Bidirectional Encoder Representations from Transformers (BERT) model, introduced by Devlin et al. (2018), leverages self-attention and bidirectional context to achieve state-of-the-art performance in NLP [3]. Inspired by these advancements, the TransNILM model, a bidirectional transformer approach, has been developed for energy disaggregation, showing promising results on REDD Dataset. This work is influenced by the original BERT4NILM framework proposed by Yue et al. (2020), which demonstrated superior performance in energy disaggregation tasks [4]. This chapter builds upon this work, focusing exclusively on the REDD dataset to evaluate the model's

performance in a specific context. We also consider the role of custom loss functions and masked training in enhancing model accuracy, particularly for challenging appliances with variable usage patterns.

# 3 Methodology

The TransNILM model, inspired by the bidirectional transformer architecture, is adapted here for non-intrusive load monitoring using the REDD dataset. This section details the dataset description, data preparation, model architecture, and training process, providing a comprehensive overview of the approach.

## 3.1 Dataset Description

The Reference Energy Disaggregation Dataset (REDD) is a publicly available dataset specifically designed for non-intrusive load monitoring research, introduced by Kolter and Johnson (2011) [5]. It contains high-frequency and low-frequency power measurements from six residential households in the United States, capturing both aggregate and appliance-level energy consumption. The dataset includes mains power readings (aggregate consumption) at a 1-second resolution and individual appliance measurements at approximately 3-second intervals, collected via plug-level monitors. Key appliances monitored include high-power devices like washers, dishwashers, and microwaves, as well as lower-power devices like fridges, providing a diverse range of consumption patterns.

The REDD dataset is split into training, validation, and test sets to evaluate model generalization. For this study, we utilize data from houses 2, 3, 4, 5, and 6 for training, house 2 for validation, and house 1 for testing, ensuring the model is assessed on unseen data. The dataset's diversity in household profiles and appliance types makes it suitable for testing the robustness of the TransNILM model. However, challenges include missing data, variable sampling rates, and noise in measurements, which require careful preprocessing to ensure reliable model performance.

## 3.2 Data Preparation

For this study, we preprocess the REDD dataset to ensure compatibility with the TransNILM model. The input sequence length is fixed at 480, reflecting a practical window for capturing temporal patterns in energy consumption, with a window stride of 120 time steps for data sampling. Data is normalized using the mean and standard deviation from the training set to ensure consistency across houses 1, 2, 3, 4, 5, and 6. A masking strategy is applied, where 25% of input elements are randomly replaced with a special token, forcing the model to learn from context and enhancing its ability to identify appliance-specific patterns. Additional preprocessing steps address missing values by interpolation and filter noise to provide robust input for training and evaluation. Power consumption cutoff thresholds are set as follows: 6000 W for aggregate, 400 W for refrigerator, 3500 W for washer/dryer, 1800 W for microwave, and 1200 W for dishwasher, ensuring predictions remain within realistic bounds.

## 3.3 TransNILM Architecture

The TransNILM architecture begins with a convolutional layer (kernel size 5, padding 2) to extract features and increase the hidden size of the one-dimensional input sequence. This is followed by a learned $L^2$ norm pooling operation (kernel size 2, stride 2) to reduce the sequence length by half while preserving critical features. A learnable positional embedding matrix is added to capture the temporal order of the sequence, as formulated in Equation 1:

$$\text{Embedding}(X) = \text{L2Pooling}(\text{Conv}(X)) + E_{\text{pos}} \tag{1}$$

The resulting embedding is fed into a bidirectional transformer with 2 layers and 2 attention heads, each with a maximum hidden size of 256. The multi-head attention mechanism allows the model to focus on different subspaces of the input, computed as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^O \tag{2}$$

where each head applies scaled dot-product attention. A position-wise feed-forward network (PFFN) with GELU activation follows, processing each element independently:

$$\text{PFFN}(X) = \text{GELU}(0, XW_1 + b_1)W_2 + b_2 \tag{3}$$

Residual connections and layer normalization are applied after both the attention and PFFN modules to stabilize training and preserve input features.

The output module consists of a deconvolutional layer (kernel size 4, stride 2, padding 1) to restore the sequence length, followed by a two-layer MLP with Tanh activation to predict energy usage. Predictions are scaled by the maximum device power and clamped to ensure reasonable values. Appliance status is determined by comparing predictions to predefined thresholds: 50 W for refrigerator, 20 W for washer/dryer, 200 W for microwave, and 10 W for dishwasher.

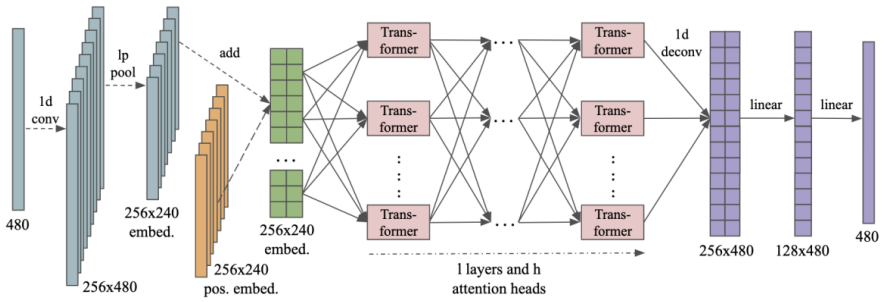


Figure 2: Architecture of the TransNILM model, showing the flow from input sequence to energy prediction

### 3.4 Training Process

The model is initialized with truncated normal distributions for weights and biases. Training is conducted using the Adam optimizer (learning rate $10^{-4}$, betas 0.9 and 0.999, zero weight decay) to minimize a custom loss function as described in [4]. The loss function integrates energy consumption prediction and appliance status classification for each appliance $k$. For a sequence of length $T$, the loss for appliance $k$ at each time step $t$ combines an energy prediction error and a status classification penalty, defined as:

$$\ell_k(t) = |\hat{p}_{k,t} - p_{k,t}| + \lambda \cdot \log\left(1 + \exp\left(-z_{k,t} \cdot \frac{s_{k,t}}{\tau}\right)\right) \tag{4}$$

where $\hat{p}_{k,t}$ is the predicted energy consumption, $p_{k,t}$ is the ground truth energy consumption, $z_{k,t}$ is the predicted status logit, $s_{k,t} \in \{0, 1\}$ is the ground truth status, $\tau$ is the softmax temperature, and $\lambda$ is a weighting factor balancing regression and classification. The log term approximates the Binary Cross-Entropy loss with temperature-scaled softmax, equivalent to the original formulation after normalization. The total loss for appliance $k$ is averaged over the sequence:

$$\mathcal{L}_k = \frac{1}{T} \sum_{t=1}^{T} \ell_k(t) \tag{5}$$

The overall loss across all appliances includes L2 regularization to prevent overfitting:

$$\mathcal{L} = \sum_k \left(\mathcal{L}_k + c_0^k \|\mathbf{w}\|^2\right) \tag{6}$$

where $\mathbf{w}$ denotes the model parameters, and $c_0^k$ is the regularization coefficient, set to $1 \times 10^{-6}$ for refrigerator, 0.001 for washer/dryer, 1.0 for microwave, and 1.0 for dishwasher. A dropout rate of 0.1 is applied to prevent overfitting. The batch size is set to 128, and the learning rate schedule is disabled. Training runs for 100 epochs for the refrigerator and 20 epochs for the washer/dryer, microwave, and dishwasher, reflecting appliance-specific convergence rates. Minimum on and off durations (in time steps) are enforced: 10 and 2 for refrigerator, 300 and 26 for washer/dryer, 2 and 5 for microwave, and 300 and 300 for dishwasher, ensuring realistic state transitions. Validation and test data are evaluated without masking to assess real-world performance, leveraging the REDD dataset's diversity for generalization across households and appliance types.

# 4 Results

The performance of TransNILM on the REDD dataset is evaluated using four standard metrics: accuracy, F1 score, mean relative error (MRE), and mean absolute error (MAE). These metrics assess the model's ability to predict both energy consumption and appliance status accurately, calculated using confusion matrix components (true negatives, false positives, false negatives, true positives) for classification and relative/absolute error formulas for regression. The model is compared against baseline architectures, including LSTM (LSTM+), and a seq2seq CNN model, all modified for consistent input length (480) and hidden size (256). Results are presented in Table 1.

Table 1: Model Performances on REDD Dataset

| Device | Model | Accuracy | F1 | MRE | MAE |
|---|---|---|---|---|---|
| Fridge | LSTM+ | 0.789 | 0.709 | 0.841 | 44.82 |
| | CNN | 0.796 | 0.689 | 0.822 | 35.69 |
| | TransNILM | 0.841 | 0.756 | 0.806 | 32.35 |
| Washer | LSTM+ | 0.989 | 0.125 | 0.020 | 35.73 |
| | CNN | 0.970 | 0.274 | 0.042 | 36.12 |
| | TransNILM | 0.991 | 0.559 | 0.022 | 34.96 |
| Microwave | LSTM+ | 0.989 | 0.604 | 0.042 | 35.99 |
| | CNN | 0.986 | 0.378 | 0.060 | 18.59 |
| | TransNILM | 0.989 | 0.476 | 0.057 | 17.58 |
| Dishwasher | LSTM+ | 0.956 | 0.421 | 0.056 | 25.25 |
| | CNN | 0.953 | 0.298 | 0.053 | 25.29 |
| | TransNILM | 0.969 | 0.523 | 0.039 | 20.49 |

TransNILM demonstrates superior performance across most appliances, achieving higher accuracy and F1 scores, and lower MRE and MAE values compared to baseline models. For the fridge, TransNILM achieves an accuracy of 0.841 and an MAE of 32.35, outperforming LSTM+ (0.789, 44.82) and CNN (0.796, 35.69). Similar trends are observed for the washer and dishwasher, where TransNILM excels in capturing both energy usage and on/off states. The bidirectional transformer architecture, combined with masked training, enables the model to learn contextual patterns effectively. However, performance on the microwave is less pronounced, likely due to its infrequent usage and the masking strategy's impact on learning on-states.

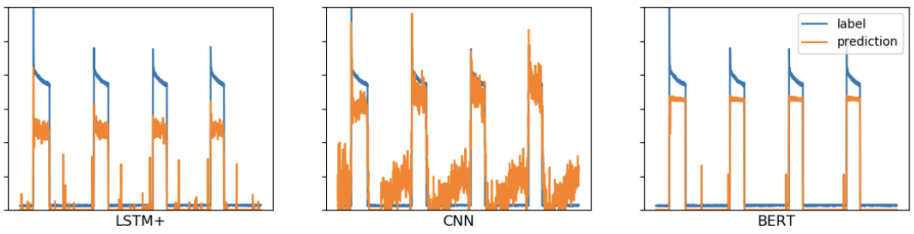


Figure 3: Sample output of refrigerator on REDD dataset by different models, highlighting TransNILM's stability

# 5 Discussion

The results highlight TransNILM's strengths in energy disaggregation, particularly its ability to model complex temporal and contextual relationships in the REDD dataset. The bidirectional transformer architecture captures dependencies across the sequence, while the custom loss function balances energy prediction and status classification. However, challenges remain for appliances with sporadic usage, such as the microwave, where low F1 scores suggest limitations in the masking strategy. The model's

computational complexity also poses challenges for real-time applications. Future work could address these by exploring lightweight transformer variants and adaptive masking techniques.

# 6 Conclusion

This chapter adapted the TransNILM model for non-intrusive load monitoring using the REDD dataset, demonstrating its effectiveness in energy disaggregation. The bidirectional transformer architecture, combined with a custom loss function and masked training, outperforms baseline models across key metrics, including accuracy, F1 score, MRE, and MAE. The model excels in predicting consumption for frequently used appliances like the fridge and dishwasher, though improvements are needed for sporadic appliances. Future work could focus on developing lightweight transformer models to reduce computational demands, refining masking strategies, and exploring multi-staged appliance modeling to enhance performance on unbalanced datasets.

# 7 References

# References


[1] Kelly, J., & Knottenbelt, W. (2015). Neural NILM: Deep Neural Networks Applied to Energy Disaggregation. In *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments* (BuildSys '15) (pp. 55–64). ACM.

[2] Zhang, C., Zhong, M., Wang, Z., Goddard, N., & Sutton, C. (2018). Sequence-to-Point Learning with Neural Networks for Non-Intrusive Load Monitoring. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence* (AAAI '18) (pp. 2604–2611). AAAI Press.

[3] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

[4] Yue, K., Dou, Z., Wang, R., & Li, J. (2020). BERT4NILM: A Bidirectional Transformer Model for Non-Intrusive Load Monitoring. *arXiv preprint arXiv:2009.10431*.

[5] Kolter, J. Z., & Johnson, M. J. (2011). REDD: A Public Data Set for Energy Disaggregation Research. In *Proceedings of the SustKDD Workshop on Data Mining Applications in Sustainability* (SustKDD '11).