energies

MDPI

Article

Appliance-Specific Noise-Aware Hyperparameter Tuning for Enhancing Non-Intrusive Load Monitoring Systems

João Góis * and Lucas Pereira

Interactive Technologies Institute (SITI/LARSyS), Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisboa, Portugal; lucas.pereira@iti.larsys.pt
* Correspondence: joao.gois@iti.larsys.pt

Abstract

Load disaggregation has emerged as an effective tool for enabling smarter energy management in residential and commercial buildings. By providing appliance-level energy consumption estimation from aggregate data, it supports energy efficiency initiatives, demand-side management, and user awareness. However, several challenges remain in improving the accuracy of energy disaggregation methods. For instance, the amount of noise in energy consumption datasets can heavily impact the accuracy of disaggregation algorithms, especially for low-power consumption appliances. While disaggregation performance depends on hyperparameter tuning, the influence of data characteristics, such as noise, on hyperparameter selection remains underexplored. This work investigates the hypothesis that appliance-specific noise information can guide the selection of algorithm hyperparameters, like the input sequence length, to maximize disaggregation accuracy. The appliance-to-noise ratio metric is used to quantify the noise level relative to each appliance's energy consumption. Then, the selection of the input sequence length hyperparameter is investigated for each case by inspecting disaggregation performance. The results indicate that the noise metric provides valuable guidance for selecting the input sequence length, particularly for user-dependent appliances with more unpredictable usage patterns, such as washing machines and electric kettles.

Keywords: load disaggregation; appliance-to-noise ratio; performance evaluation; hyperparameter tuning; input sequence length

Academic Editor: Steve Burrow

Received: 16 May 2025
Revised: 9 July 2025
Accepted: 17 July 2025
Published: 19 July 2025

Citation: Góis, J.; Pereira, L. Appliance-Specific Noise-Aware Hyperparameter Tuning for Enhancing Load Disaggregation in Smart Energy Systems. *Energies* **2025**, *18*, 3847. https://doi.org/10.3390/en18143847

1. Introduction

Non-Intrusive Load Monitoring (NILM) [1], or load disaggregation, refers to a set of signal processing and machine learning techniques used to estimate the individual energy consumption of appliances by using the total energy consumption in the building collected using a single sensor. In other words, the aggregate signal of the building is disaggregated into individual appliance components. NILM's scalability, cost-effectiveness, and non-intrusive nature make it a practical solution for improving energy management and efficiency [2].

NILM plays a vital role in the broader context of smart energy systems. By offering detailed insights into individual appliance usage, it enables applications such as personalized energy feedback, the detection of faulty appliances, improved demand forecasting, and integration with home automation systems [3]. For utility providers, NILM can support demand-side management programs and reduce the need for expensive submetering in-

frastructure. From a policy standpoint, it can contribute to energy conservation goals and more accurate load profiling [4].

The existence of noise in the aggregate signal affects appliance identification and overall load disaggregation performance. Noise may arise from sensor errors or unmetered appliances that introduce irregular consumption patterns. Formally, noise is quantified as the remaining power in the aggregate readings once the appliances' power has been subtracted [5], as given by Equation (1):

$$\eta_t = y_t - \sum_{i=1}^{M} x_t^{(i)} \quad (1)$$

The noise component poses a considerable challenge for extracting individual appliance signals from the aggregate energy consumption data. According to Dong et al., the probability of accurately distinguishing between different appliance events is reduced when the noise level in the aggregate signal increases [6]. The energy signal becomes more complex as a result of a higher percentage of unknown power [7]. Gupta et al. have corroborated these findings, highlighting the significant impact of noise on the performance of NILM algorithms [8]. The amount of noise can vary significantly across datasets due to several factors, such as the number of available submeters. If a building contains many appliances, and just a few are submetered, there will be a higher percentage of unknown appliance energy consumption [9].

Removing noise from the aggregate signal (denoising) has been attempted to mitigate its effects on disaggregation performance. However, denoised data do not accurately reflect real-world conditions, potentially leading to misleading disaggregation performance evaluations [7,10]. Therefore, efforts have been made to quantify, report, and assess the sensitivity of load disaggregation accuracy relative to noise. The noise-to-aggregate ratio (NAR) metric [5,11] was specifically proposed for NILM, enabling the measurement of the total share of noise in an aggregate signal and assessing the load disaggregation method's sensitivity. Another noise measure used in NILM is the signal-to-noise ratio (SNR), which is widely used in other machine learning disciplines. In this context, the SNR provides the magnitude of the aggregate signal relative to the underlying noise. Meziane et al. used the SNR to analyze the impact of noise on an event-based NILM method [12], while Lu et al. also used it to assess the noise effect on disaggregation accuracy [13].

However, metrics like NAR or SNR do not capture the proportion of noise relative to individual appliances. The disaggregation performance for a specific appliance is likely to be affected by the level of noise, especially when the noise level is a significant fraction of the appliance's power [14]. Noise is more likely to overcome low-power appliances like televisions than high-power appliances like washing machines, although this is not captured by the NAR or SNR. Recently, the appliance-to-noise ratio (ANR) metric has been proposed to quantify noise relative to the energy consumption of individual appliances.

Disaggregation accuracy strongly depends on hyperparameters such as epochs, input sequence length (ISL), and event threshold [4,15]. The event threshold hyperparameter, in particular, interferes directly with the number of events to be counted for a given appliance and should be fine-tuned. The hyperparameter value should be high enough to discard minor voltage as an event and low enough to avoid missing events [16,17].

The ISL is another relevant hyperparameter, which is provided as input to recent deep learning algorithms such as the sequence-to-sequence (S2S) and sequence-to-point (S2P) [18], as well as the more recent bidirectional transformer for NILM (BERT4NILM)[19]. ISL is the length of the sequence of mains readings given as input to the disaggregation model, which is then used for the prediction of a sequence of equal length for the target appliances. The ISL hyperparameter must be carefully selected; otherwise, the input

sequence may be inadequate for effectively training the algorithm and producing accurate appliance-level predictions [3,20]. In this respect, the analysis of data characteristics like usage patterns, appliance power consumption, noise, and sample rate can provide relevant information to improve hyperparameter selection [21–24].

However, the literature concerning the integration of data characteristics in the selection of hyperparameters is still scarce. Furthermore, although the existence of noise in energy signals has been observed to impact load disaggregation accuracy, the potential of noise information for improving hyperparameter selection has not been further researched. Metrics like ANR enable incorporating noise information into this challenge.

Motivated by the need to enhance NILM accuracy under realistic, noise-prone conditions, this work investigates the hypothesis that appliance-specific noise information, given by the ANR, could be utilized for informed selection of algorithm hyperparameters, like ISL, in energy disaggregation. The hypothesis considers that, for a given household appliance, if the ANR is low (either due to high noise or low appliance power consumption), the algorithm performance should increase for higher ISL values, i.e., with high sequence length, the estimation of appliance power consumption is expected to be more accurate. This work attempts to validate this hypothesis to ensure the accurate and reliable prediction of appliance energy consumption. For this purpose, the ANR is computed, and the existence of patterns in algorithm performances across ISL values is assessed for various appliances using real-world data.

The remainder of the paper is structured as follows: Section 2 provides the background and related work; Section 3 describes the dataset, the definition of ANR, and the methods utilized to investigate the initial hypothesis; Section 4 includes results and discussion; and finally, Section 5 establishes the main conclusions, research implications, and future work.

2. Related Work

Noise is a frequent component of energy signals that affects the accuracy of load disaggregation algorithms. Noise metrics like NAR and SNR have been considered for quantifying noise, with SNR being a widely known metric across machine learning disciplines that has been utilized in NILM (e.g., [12,13]) and NAR proposed specifically for NILM [5,7]. At the appliance level, the proportion of noise can be significant, especially for low-power appliances, reducing the disaggregation accuracy. However, these metrics do not capture the relative level of noise compared to the power consumption of individual appliances. Gois and Pereira proposed the ANR metric to measure and report appliance-specific noise information in datasets, enabling a detailed noise analysis [14]. The applicability of ANR was demonstrated across different time resolutions, allowing the identification of patterns in disaggregation accuracy across distinct noise levels.

The disaggregation accuracy is intrinsically related to the configuration of algorithm hyperparameters, which include the number of epochs, batch size, ISL, and event threshold [4,15]. For ISL, in particular, the length of the data sequences used for training the algorithm is essential for obtaining the best predictions of appliance energy consumption [3,20]. According to Reinhardt et al., sensitivity analysis of hyperparameters should be considered to achieve a more reliable algorithm performance [23]. Bousbiat et al. suggested the utilization of optimization techniques for hyperparameter selection, using the functionalities available in the Deep-NILMTK framework [22].

Studies recommend leveraging usage patterns, noise, and sample rate when selecting algorithm hyperparameters (such as ISL) for models like S2S, S2P, and BERT4NILM [21,23,24]. Bousbiat et al. also suggest the consideration of power consumption variation between high-power and low-power appliances for ISL selection [22]. This implies that ISL selection in

algorithms such as S2S, S2P, and BERT4NILM likely varies across appliance types, depending on user behavior and power consumption characteristics.

While previous studies have explored the impact of noise on disaggregation accuracy, there has been limited investigation into how such data characteristics can inform hyperparameter selection. In particular, most prior works approached hyperparameter tuning using default values, trial-and-error, or brute-force search methods [22,23], which are often computationally intensive and not tailored to specific data conditions. This paper addresses this gap by introducing a data-driven approach that incorporates appliance-specific noise information into hyperparameter selection. By leveraging the ANR metric, the proposed method enables a more informed and adaptive selection of the ISL hyperparameter, aiming to enhance disaggregation performance across a variety of appliances.

3. Materials and Methods

This section introduces the definition of the ANR metric, the dataset, the NILM algorithm, the performance metric, and the experiments conducted for investigating the initial hypothesis.

3.1. Appliance-to-Noise Ratio

The ANR [14] metric has been proposed for assessing the proportion of noise relative to an individual appliance's electricity consumption. For a specific appliance k, the metric's formula is given by Equation (2):

$$ANR_k = \frac{1}{T}\sum_{t=1}^{T}\frac{x_t^{(k)}}{|y_t - \sum_{m=1}^{M} x_t^{(m)}|} \tag{2}$$

where $x_t^{(k)}$ is the actual power of appliance k at time t, and y_t is the total power consumption at time t. The denominator provides the amount of noise at time t. This metric averages the instant ratios of appliance electricity consumption and noise. Each instant ratio contributes to the final score, summarizing the impact of noise on an appliance over time. Only instant ratios with non-zero noise power (denominators greater than zero) are considered for the metric calculation; otherwise, the ANR would be undefined.

3.2. Dataset

The REFIT [25] dataset includes active and apparent power measurements at the aggregate and ground-truth levels at the time resolution of 8 s. The data was collected continuously over 2 years (2013–2015) from 20 houses located in Loughborough, the United Kingdom, while household occupants were conducting their usual domestic routines. The households are originally numbered from 1–13 and 15–21, which is the naming convention also used in this work.

The houses under study were equipped with plug meters that collected the energy consumption data from the 9 appliances with the highest demand. The main targets included cold appliances (like refrigerators and freezers), cooking appliances (like microwaves and electric kettles), information and communication technology (like computers and screens), and utility room appliances (washing machines and dishwashers). The television and washing machine were the most submetered appliances across the houses.

The REFIT dataset was chosen due to the combination of properties rarely found in other publicly available datasets, such as long-term measurements, relatively high sample rate, and appliance-level and aggregate data across multiple households. Furthermore, existing load disaggregation toolkits contain functionality that facilitates loading and analysis for this specific dataset, making it more convenient for this work.

3.3. Algorithm and Performance Metric

Despite the rise of attention mechanisms in NILM with convolutional and recurrent neural networks, the recent development of natural language processing brought the BERT model [26], originally proposed for text classification tasks, into the field of NILM. The proposed architecture for NILM, known as BERT4NILM [19], contains an embedding module, transformer layers, and a multilayer perceptron output layer, operating with a self-attention mechanism. The network takes sequential data with a fixed length and predicts appliance energy usage with an output of the same shape, while the appliance states are additionally computed by comparing the on-thresholds defined by the researcher. The bidirectional structure of BERT captures time dependencies and patterns in the data, which are suitable for improving load disaggregation accuracy. This innovative approach rivals other recent deep learning NILM algorithms, showing the potential of BERT4NILM.

The selected metric for evaluating the algorithm's accuracy is the Match Rate (MR), which is a normalized metric that evaluates the overlapping rate of the true and estimated energy, varying between 0 (weak match) and 1 (strong match). MR can be calculated as per Equation (3):

$$MR(\hat{y}, y) = \frac{\sum_{t=1}^{T} \min\{\hat{y}_t, y_t\}}{\sum_{t=1}^{T} \max\{\hat{y}_t, y_t\}} \tag{3}$$

where y_t is the actual power consumption for appliance k at time t and $\hat{y}_t$ is the estimated power consumption for appliance k at time t. The MR was selected for this analysis due to the suitability of this metric for evaluating the performance of NILM algorithms observed in previous works. The metric's normalization constant corresponds to the total appliance energy consumption (or more, in cases of overestimation), aligning with recommendations by Garcia et al. to better quantify estimation errors [27]. In addition, MR has been observed to adequately provide information on disaggregation performance for diverse estimation scenarios, like overestimation and underestimation, being easily interpreted compared to other metrics [28].

3.4. Experiment Setup

This experiment investigates the hypothesis that the information provided by the ANR metric can guide the selection of the ISL hyperparameter to improve disaggregation performance.

This analysis focuses on household data from the REFIT dataset, considering different appliances (washing machine, electric kettle, fridge-freezer, and fridge) to generalize the hypothesis testing. The washing machines, fridges, and fridge-freezers are multistate (type II) appliances, while electric kettles are on/off (type I) appliances [29,30]. The washing machines have three modes of operation: in use, standby, and switched off. Also, the washing machines are user-dependent, meaning the operation patterns vary for different user behaviors. The fridges and fridge-freezers, however, are user-independent, as their operation is continuous and cyclic throughout the day between zero and a set of power levels under thermostatic control. The electric kettles have two modes of operation (on/off) and are user-dependent.

The experiments are conducted between 20 June 2014 and 20 October 2014 (4 months). The chosen data resolution is one sample per minute, balancing data granularity and computational cost. For each appliance, the BERT4NILM is trained in 75% of the sample (three months) and test in 25% of the sample data (one month). The time series rolling k-cross validation is utilized, as it is superior to traditional k-fold cross-validation for sequential energy data. The number of folds k is set at 4, creating folds of approximately 1 month. The MR metric evaluates the accuracy of BERT4NILM estimations for each case.

The BERT4NILM algorithm is computed for different values of ISL for each appliance: 15, 30, 60, 120, and 240. The range from 15 min to 240 min sequences was selected to balance short and long input lengths while capturing meaningful performance trends through intermediate values (30, 60, and 120). ISL values below 15 would result in sequences that are too short, while those above 240 would produce excessively long input windows.

To test the initial hypothesis, the ANR was calculated for each appliance across the households. Then, for fixed ANR values, it was assessed for each case whether patterns arise in the MR scores across different ISL values. For an appliance in a given household, if the ANR was low (high noise and/or low appliance power consumption), it was expected that the BERT4NILM performance (shown through the MR score) would increase for higher ISL values. The assessment of this hypothesis was crucial for ensuring that better performance was obtained for each case.

3.5. Hardware and Software

Hardware-wise, the computer comprised an Intel i7-8700k CPU, an NVIDIA 1080TI graphics card, and 64 GB of RAM. Software-wise, the experiments conducted throughout this work were carried out using Python 3.8.16, with Keras [31] running on the Tensorflow [32] backend. Python's package Deep-NILMTK [22] was considered for its comprehensive support of functionalities for running experiments, such as data analysis and preprocessing tools, experiment management and templating, implementation of recent deep learning algorithms like BERT4NILM, and availability for changing the configurations of algorithm hyperparameters easily. Deep-NILMTK enhances functionalities over earlier toolkits like NILMTK [15,33]. To reduce the computation time of BERT4NILM, the cuDNN library (version 8.1.0) was utilized to accelerate the GPU calculations.

4. Results and Discussion

Table 1 shows the ANR scores for the washing machine across different houses and the respective MR scores for different ISL values. MR scores are generally lower for shorter ISL values (15, 30, and 60), but tend to rise as ISL increases, which was particularly verified for households 3, 4, 5, 7, 10, 13, and 15.

ANR scores are generally low across households, with exceptions in households 5, 17, and 21. According to the definition of ANR, low appliance power consumption or sufficiently high noise levels may explain low ANR values for the washing machine across the houses. For 56% of the house's washing machines with low ANRs, the MR scores increase significantly as ISL increased. As such, there is evidence that selecting a higher ISL value in the context of a low ANR can help improve the MR scores for a washing machine. This is in line with the fact that washing machines are user-dependent appliances, with possibly unpredictable usage patterns; hence, a higher ISL value may be required to increase the disaggregation performance.

Next, we investigated whether such a relationship between the ANR and the selection of ISL held for the fridge-freezer, which is a user-independent appliance. From Table 2, the ANRs for the fridge-freezer seemed to vary significantly across the households, with higher ANR values in households 5 and 18, and much lower ANR values in households 10, 11, and 15. In contrast to the washing machine, for most households, the MR scores did not vary significantly with the ISL in general, independently of the ANR values. This suggests that the noise levels do not seem to impact the selection of ISL for this appliance. This may be attributed to the fridge-freezer's user-independent, cyclic operation; thus, the energy signal has a well-defined pattern that can be detected with a relatively good performance, independently of the choice of ISL.

Table 1. ANR and MR scores for the washing machine considering different ISL values across the households. A darker color represents a higher ANR, as well as a higher MR score across ISL values.

WM Households	ANR	ISL				
		15	30	60	120	240
1	0.08	0	0	0.002	0	0.06
2	0.07	0	0	0	0.005	0.06
3	0.08	0.03	0.004	0.002	0.08	0.12
4	0.06	0.002	0.005	0.07	0.17	0.13
5	0.21	0	0.007	0	0.03	0.08
6	0.01	0	0	0	0.003	0.002
7	0.08	0	0	0.07	0.09	0.17
8	0.09	0	0.002	0.001	0.002	0.02
9	0.05	0	0	0	0.009	0.02
10	0.09	0	0.02	0.006	0.11	0.14
11	0.06	0	0	0	0	0.003
13	0.06	0	0	0.003	0.03	0.18
15	0.08	0.002	0.003	0.04	0.19	0.21
16	0.05	0	0.001	0.01	0.03	0.07
17	0.13	0	0	0	0.002	0.04
18	0.02	0	0	0	0.001	0
19	0.08	0	0	0	0	0.002
20	0.08	0	0	0	0.004	0.03
21	0.17	0	0	0	0	0

Table 2. ANR and MR scores for the fridge-freezer considering different ISL values across the households. A darker color represents a higher ANR, as well as a higher MR score across ISL values.

FFZ Households	ANR	ISL				
		15	30	60	120	240
2	0.47	0.31	0.29	0.31	0.29	0.27
3	0.45	0.59	0.62	0.59	0.62	0.58
4	0.31	0.19	0.18	0.23	0.23	0.3
5	1.81	0.44	0.43	0.45	0.39	0.46
9	0.32	0.32	0.34	0.35	0.34	0.33
10	0.18	0.15	0.12	0.14	0.17	0.11
11	0.17	0.25	0.21	0.23	0.23	0.22
12	0.5	0.22	0.17	0.17	0.22	0.21
15	0.24	0.26	0.27	0.22	0.22	0.23
16	0.38	0.36	0.34	0.4	0.4	0.39
17	0.65	0.27	0.22	0.26	0.22	0.19
18	1.6	0.4	0.41	0.42	0.41	0.43
19	0.83	0.41	0.38	0.42	0.4	0.38
21	0.61	0.35	0.36	0.37	0.36	0.36

The link between the ANR and ISL selection appears to be influenced by appliance usage patterns. Next, an analysis was conducted for the electric kettle and the fridge to check if the initial hypothesis held. For the electric kettle, Table 3 indicates that the MR scores tend to increase when the ISL increases, particularly for households 3, 4, 7, 8, and 20. The ANRs for the electric kettle were generally low across the households, except for households 5, 9, 11, and 17. For 90% of the households with low ANRs, MR scores significantly increased with ISL. This supports the hypothesis that a longer ISL improves MR scores for electric kettles in low-ANR conditions. Similarly to the washing machine, electric kettles are also user-dependent, with an unpredictable usage pattern, and the initial hypothesis seems to be verified.

Table 3. ANR and MR scores for the electric kettle considering different ISL values across the households. A darker color represents a higher ANR, as well as a higher MR score across ISL values.

KT Households	ANR	ISL				
		15	30	60	120	240
2	0.09	0.01	0.02	0.17	0.18	0.24
3	0.01	0.01	0.01	0.02	0.12	0.13
4	0.01	0.01	0.11	0.13	0.28	0.2
5	0.17	0.02	0.02	0.03	0.18	0.23
6	0.11	0.04	0.23	0.2	0.3	0.27
7	0.01	0.01	0.01	0.01	0.1	0.07
8	0.01	0.01	0.01	0.17	0.17	0.22
9	0.15	0.01	0.11	0.15	0.14	0.18
11	0.2	0.01	0.01	0.02	0.18	0.17
12	0.05	0.03	0.03	0.02	0.03	0.005
13	0.05	0.03	0.03	0.03	0.16	0.14
17	0.39	0.02	0.03	0.06	0.22	0.18
19	0.11	0.02	0.21	0.19	0.29	0.19
20	0.01	0.16	0.03	0.3	0.24	0.32

Regarding the fridge, according to Table 4, the ANRs seem to vary significantly across the households, with higher values for households 18 and 20 in contrast to households 4, 8, and 11. As with the fridge-freezer, MR scores show minimal variation with changes in ISL across most households, regardless of the corresponding ANR values. Hence, the noise levels do not seem to influence the choice of the ISL for the fridge.

Therefore, there is evidence that the initial hypothesis is validated for user-dependent appliances, with more unpredictable usage patterns. For user-independent appliances, like fridge-freezers and fridges, the hypothesis does not seem to hold, which may be related to the fact that these appliances have recurrent operating patterns. Nevertheless, to solidify the conclusions obtained in this work, it would be relevant to investigate the initial hypothesis for other user-dependent and user-independent appliances. The use of different sources of data would also be important for generalization.

Table 4. ANR and MR scores for the fridge considering different ISL values across the households. A darker color represents a higher ANR, as well as a higher MR score across ISL values.

FRI Households	ANR	ISL				
		15	30	60	120	240
1	0.15	0.14	0.15	0.16	0.16	0.15
4	0.1	0.14	0.14	0.13	0.14	0.14
7	0.13	0.13	0.14	0.14	0.15	0.13
8	0.09	0.12	0.14	0.13	0.13	0.15
11	0	0	0.003	0.04	0	0
18	0.38	0.14	0.16	0.13	0.14	0.12
20	0.38	0.32	0.34	0.34	0.32	0.33

The code for reproducing the experiments shown in Tables 1–4 is available in https://anonymous.4open.science/r/Hyperparameter-Tuning-in-Load-Disaggregation-using-Appliance-specific-Noise-Information-ADEF (accessed on 15 May 2025).

5. Conclusions

This work investigates the hypothesis that noise information can be used for guiding the selection of algorithm hyperparameters for maximizing load disaggregation accuracy. The recently proposed ANR metric was used to quantify appliance-specific noise. Consider-

ing the ANR scores, the choice of the ISL hyperparameter was investigated across different household appliances. The results show that, for specific usage patterns of appliances, the initial hypothesis is verified. For user-dependent appliances with less predictable usage patterns, such as washing machines and electric kettles, a high ANR indicates that a longer ISL may enhance disaggregation accuracy. However, for appliances with continuous cyclic operation, like fridges and fridge-freezers, the initial hypothesis does not seem to be verified, as the ANR does not impact the choice of the ISL.

This study advances out understanding of hyperparameter selection in NILM, showing that noise information can help maximize disaggregation accuracy. This contribution is particularly relevant for the deployment of NILM systems in realistic settings, where noise is present and can hinder model performance. By adapting hyperparameter tuning to data-specific characteristics, such as appliance-level noise, it is possible to improve the robustness and applicability of NILM.

Although the objectives of this work have been achieved, some limitations could be addressed in future work. Testing the initial hypothesis on additional appliances would improve generalizability. In addition, while the REFIT dataset includes both aggregate and appliance-level energy data from a diverse set of households, being representative of different types of users, further generalization of the conclusions would require that the hypothesis be tested in datasets from different geographies.

Future work could explore additional hypotheses, such as assessing whether the analysis of noise may help select other algorithm hyperparameters. The analysis of usage patterns and anomaly detection techniques from other machine learning disciplines could also be considered for studying hyperparameter selection, along with noise. Furthermore, it would be relevant to investigate whether certain NILM algorithms are inherently more robust to noise, independent of hyperparameter tuning or appliance characteristics. Future work could involve a comparative analysis of algorithmic sensitivity to noise, improving our understanding of model robustness under real-world conditions. Although prior work has shown that usage patterns are relevant for ISL selection, conducting more experiments would be important to support this hypothesis.

Author Contributions: Conceptualization, J.G. and L.P.; methodology, J.G.; software, J.G.; validation, J.G. and L.P.; formal analysis, J.G.; investigation, J.G.; resources, J.G.; data curation, J.G.; writing—original draft preparation, J.G. and L.P.; writing—review and editing, J.G. and L.P.; visualization, J.G.; supervision, L.P.; project administration, J.G. and L.P.; funding acquisition, J.G. and L.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Portuguese Foundation for Science and Technology (FCT) under projects 10.54499/LA/P/0083/2020; 10.54499/UIDP/50009/2020 & 10.54499/UIDB/50009/2020, and grants DFA/BD/8075/2020 (J.G.) and CEECIND/01179/2017 (L.P.).

Data Availability Statement: The data utilized is publicly available in https://pureportal.strath.ac.uk/en/datasets/refit-electrical-load-measurements-cleaned (accessed on 15 May 2025).

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

Abbreviations

The following abbreviations are used in this manuscript:

ANR	Appliance-to-Noise Ratio
BERT	Bidirectional Transformer Model
ISL	Input Sequence Length
MR	Match Rate

NAR	Noise-to-Aggregate Ratio
NILM	Non-intrusive Load Monitoring
BERT4NILM	Bidirectional Transformer for NILM
SNR	Signal-to-Noise Ratio
S2P	Sequence-to-Point
S2S	Sequence-to-Sequence

References

1. Hart, G.W. Nonintrusive appliance load monitoring. *Proc. IEEE* **1992**, *80*, 1870–1891.
2. Kaselimi, M.; Protopapadakis, E.; Voulodimos, A.; Doulamis, N.; Doulamis, A. Towards Trustworthy Energy Disaggregation: A Review of Challenges, Methods, and Perspectives for Non-Intrusive Load Monitoring. *Sensors* **2022**, *22*, 5872.
3. Francou, Y. Contribution of Non-Intrusive Load Monitoring to Home Energy Management Systems. Ph.D. Thesis, Université de la Réunion, Saint-Denis, France, 2023.
4. Gopinath, R.; Kumar, M. DeepEdge-NILM: A case study of non-intrusive load monitoring edge device in commercial building. *Energy Build.* **2023**, *294*, 113226.
5. Klemenjak, C.; Makonin, S.; Elmenreich, W. Towards comparability in non-intrusive load monitoring: On data and performance evaluation. In Proceedings of the 2020 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), Washington, DC, USA, 17–20 February 2020; pp. 1–5.
6. Dong, R.; Ratliff, L.; Ohlsson, H.; Sastry, S.S. Fundamental limits of nonintrusive load monitoring. In Proceedings of the 3rd International Conference on High Confidence Networked Systems, Berlin, Germany, 15–17 April 2014; pp. 11–18.
7. Klemenjak, C.; Makonin, S.; Elmenreich, W. Investigating the performance gap between testing on real and denoised aggregates in non-intrusive load monitoring. *Energy Inform.* **2021**, *4*, 3.
8. Gupta, S.; Gupta, A. Dealing with noise problem in machine learning data-sets: A systematic review. *Procedia Comput. Sci.* **2019**, *161*, 466–474.
9. Egarter, D.; Pöchacker, M.; Elmenreich, W. Complexity of power draws for load disaggregation. *arXiv* **2015**, arXiv:1501.02954.
10. Gomes, E.; Pereira, L. PB-NILM: Pinball guided deep non-intrusive load monitoring. *IEEE Access* **2020**, *8*, 48386–48398.
11. Makonin, S.; Popowich, F. Nonintrusive load monitoring (NILM) performance evaluation. *Energy Effic.* **2015**, *8*, 809–814.
12. Meziane, M.N.; Ravier, P.; Lamarque, G.; Le Bunetel, J.C.; Raingeaud, Y. High accuracy event detection for non-intrusive load monitoring. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2452–2456.
13. Lu, C.; Ma, L.; Xu, T.; Ding, G.; Wu, C.; Jiang, X. Non-intrusive load monitoring method based on improved differential evolution algorithm. In Proceedings of the 2019 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), Qiqihar, China, 28–29 April 2019; pp. 279–283.
14. Góis, J.; Pereira, L. ANR: A New Metric to Quantify the Appliance to Noise Ratio in Load Disaggregation Datasets. *IEEE Trans. Instrum. Meas.* **2023,** 72, 9004508.
15. Batra, N.; Kukunuri, R.; Pandey, A.; Malakar, R.; Kumar, R.; Krystalakos, O.; Zhong, M.; Meira, P.; Parson, O. Towards reproducible state-of-the-art energy disaggregation. In Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, New York, NY, USA, 10–14 November 2019; pp. 193–202.
16. Precioso, D.; Gómez-Ullate, D. NILM as a regression versus classification problem: The importance of thresholding. *arXiv* **2020**, arXiv:2010.16050.
17. Pereira, L.; Nunes, N. A comparison of performance metrics for event classification in non-intrusive load monitoring. In Proceedings of the 2017 IEEE International Conference on Smart Grid Communications (SmartGridComm), Dresden, Germany, 23–27 October 2017; pp. 159–164.
18. Zhang, C.; Zhong, M.; Wang, Z.; Goddard, N.; Sutton, C. Sequence-to-point learning with neural networks for non-intrusive load monitoring. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
19. Yue, Z.; Witzig, C.R.; Jorde, D.; Jacobsen, H.A. Bert4nilm: A bidirectional transformer model for non-intrusive load monitoring. In Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring, Virtual, 18 November 2020; pp. 89–93.
20. Mari, S.; Bucci, G.; Ciancetta, F.; Fiorucci, E.; Fioravanti, A. An embedded deep learning nilm system: A year-long field study in real houses. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 2531215.
21. Hidalgo Cruzalegui, C.M. Impact of sampling rates for state-of-the-art NILM: A case study with sequence-to-point algorithms and the UK-DALE dataset. Master's Thesis, Universitat Politècnica de Catalunya, Barcelona, Spain, 2023.
22. Bousbiat, H.; Faustine, A.; Klemenjak, C.; Pereira, L.; Elmenreich, W. Unlocking the Full Potential of Neural NILM: On Automation, Hyperparameters & Modular Pipelines. *IEEE Trans. Ind. Inform.* **2022**, *19*, 7002–7010.

23. Reinhardt, A.; Bouchur, M. On the impact of the sequence length on sequence-to-sequence and sequence-to-point learning for nilm. In Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring, Virtual, 18 November 2020; pp. 75–78.
24. Yang, A.; Li, W.; Yang, X. Short-term electricity load forecasting based on feature selection and Least Squares Support Vector Machines. *Knowl.-Based Syst.* **2019**, *163*, 159–173.
25. Murray, D.; Stankovic, L.; Stankovic, V. An electrical load measurements dataset of United Kingdom households from a two-year longitudinal study. *Sci. Data* **2017**, *4*, 160122. https://doi.org/10/f9k7k9.
26. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
27. Garcia-Perez, D.; Pérez-López, D.; Diaz-Blanco, I.; González-Muñiz, A.; Domínguez-González, M.; Vega, A.A.C. Fully-convolutional denoising auto-encoders for NILM in large non-residential buildings. *IEEE Trans. Smart Grid* **2020**, *12*, 2722–2731.
28. Mayhorn, E.T.; Sullivan, G.P.; Petersen, J.M.; Butner, R.S.; Johnson, E.M. *Load Disaggregation Technologies: Real World and Laboratory Performance*; Tech. Rep. PNNL-SA-116560; Pacific Northwest National Laboratory (PNNL): Richland, WA, USA, 2016.
29. Zoha, A.; Gluhak, A.; Imran, M.A.; Rajasegarar, S. Non-intrusive load monitoring approaches for disaggregated energy sensing: A survey. *Sensors* **2012**, *12*, 16838–16866.
30. Kong, W.; Dong, Z.Y.; Wang, B.; Zhao, J.; Huang, J. A practical solution for non-intrusive type II load monitoring based on deep learning and post-processing. *IEEE Trans. Smart Grid* **2019**, *11*, 148–160.
31. Chollet, F. Keras: The Python Deep Learning Library, Astrophysics Source Code Library, [1806.022]. June 2018. http://ascl.net/1806.022 (accessed on 15 January 2025).
32. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. Tensorflow: A system for large-scale machine learning. In Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), Savannah, GA, USA, 2–4 November 2016; pp. 265–283.
33. Batra, N.; Kelly, J.; Parson, O.; Dutta, H.; Knottenbelt, W.; Rogers, A.; Singh, A.; Srivastava, M. NILMTK: An open source toolkit for non-intrusive load monitoring. In Proceedings of the 5th International Conference on Future Energy Systems, Cambridge, UK, 11–13 June 2014; pp. 265–276.