©IEEE Transactions on Instrumentation and Measurement, 2023. This is the author's version of the work. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising, promotional purposes, or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the copyright holder. The definite version is accepted for publication in IEEE Transactions on Instrumentation and Measurement, 2023.

# ANR: A New Metric to Quantify the Appliance to Noise Ratio in Load Disaggregation Datasets

João Góis, *Member, IEEE,* and Lucas Pereira, *Member, IEEE*

***Abstract*—Non-intrusive Load Monitoring (NILM), or load disaggregation, aims to decompose aggregate power consumption into appliance components. Factors such as noise power affect algorithm performance, reducing accuracy and increasing complexity. While existing literature often relies on standard machine learning metrics to report noise power proportion in aggregate consumption, these are overall measures that do not specify the ratio between appliance consumption and noise power.**

**This paper proposes a noise metric that assesses the proportion of dataset noise relative to an appliance's consumption in NILM. The proposed metric's sensitivity and applicability are assessed for different data scenarios using a real-world dataset. Additionally, the paper explores a potential association between the proposed metric and disaggregation performance is also inspected. Furthermore, the proposed metric is compared to existing noise metrics to highlight its unique contributions. The results demonstrate that the proposed metric effectively quantifies noise proportion with respect to specific appliances across various data scenarios. It exhibits robustness to suitable granularity variations and complements other noise metrics in interpreting NILM experiments.**

***Index Terms*—Load Disaggregation, Signal Noise, Appliance Consumption, Disaggregation Complexity, Comparability**

## I. Introduction

CONSISTENT reduction in greenhouse gas emissions is necessary for sustainable global energy distribution. Initiatives have been launched over the years to discuss inefficient energy use in buildings, leading to significant investments in smart meters for monitoring overall building consumption and developing novel data-based services [1]. In this regard, Non-Intrusive Load Monitoring (NILM) [2], or load disaggregation, is a technique that uses machine learning and signal processing to estimate appliance consumption using a single meter, typically a smart meter that measures the aggregate signal [3].

In NILM datasets, noise is part of the aggregate signal due to measurement errors or unknown appliance consumption. In previous research, dataset noise was observed to affect algorithms' performance in NILM [4]. The probability of an algorithm correctly detecting an appliance event decreases with measurement noise [5]. For buildings with multiple appliances and only a few submetered, the proportion of noise is significant, increasing the problem complexity [6]. While removing noise entirely improves performance, disaggregation results can be unrealistic and misleading [7]–[9].

Proper noise analysis is crucial for load disaggregation, as it affects load identification and consumption estimation. To ensure comparability, reporting dataset noise is essential in experiments [10]. Two commonly considered noise metrics in NILM are the noise-to-aggregate ratio (NAR) [7], [8] and signal-to-noise ratio (SNR), e.g., [11], [12]. SNR measures the proportion of a signal to underlying noise and is widely used in signal processing. On the other hand, NAR is a NILM-specific metric that quantifies the percentage of noise in the aggregate signal.

The impact of noise on disaggregating specific appliances varies due to differences in appliance power consumption and usage patterns. For instance, noise is more likely to overcome a television than a washing machine because it consumes much less power. General noise metrics like NAR and SNR may not accurately represent the noise impact on different appliances. Therefore, a measure that accounts for the noise proportion with respect to each appliance is needed.

In this regard, the present work proposes the Appliance-to-Noise Ratio (ANR), a non-accumulative metric that averages the appliance consumption and dataset noise ratios at each instant. To our knowledge, this work is the first to extend noise analysis to individual appliances. An extensive analysis to evaluate the applicability of ANR is carried out for different appliances in a real-world dataset. First, the proposed metric's sensitivity is assessed for different sample rates, and then computed against SNR for different data scenarios. Finally, the possible association between noise and the problem complexity is investigated by considering the proposed metric.

The paper is structured as follows: Section II provides background and related work on noise analysis in NILM. Section III defines the proposed metric and examines its general behavior. Section IV describes experiments to evaluate the metric's sensitivity and applicability in different data scenarios and its impact on NILM algorithm performance. Section V presents and discusses the results of the metric assessment. Finally, Section VI summarizes the main conclusions, limitations, and future research directions.

## II. Related Work

In various machine learning fields, noise can significantly increase the complexity of models and learning time, and the performance of learning algorithms is degraded [13]. As previously stated, the dataset noise affects the probability of correct detection of a device in NILM [5]. Since the percentage of dataset noise can vary significantly across datasets, it is essential to report it to reach comparability between NILM problems and thus enable fair benchmarking of NILM approaches [10].

Existing noise metrics in NILM calculate the overall noise-to-aggregate ratio. SNR (also $\text{SNR}_{dB}$, in logarithmic decibel scale) is a well-known signal processing metric used in machine learning. In NILM, SNR reports the magnitude of the aggregate signal concerning underlying noise and the sensitivity of algorithm accuracy and load identification rates [11],

[12]. Alternatively, NAR is a NILM-specific metric reporting the percentage of noise in the aggregate data [8], [10]. $\text{SNR}_{dB}$ is more suitable than NAR to handle wide dynamic ranges in signals [14], i.e., large amplitudes between aggregate and noise.

Existing noise metrics do not consider specific appliance types and their individual characteristics. Noise levels in the dataset can significantly impact the disaggregation performance, especially when the noise proportion is substantial relative to appliance consumption. Still, to the best of our knowledge, the Instrumentation and Measurement literature comprises only a few papers that address noise in the context of NILM [15]–[17]. However, these works only acknowledge the existence of noise without proposing a detailed noise analysis or quantification methods. The proposed measure is based on the definition of $\text{SNR}_{dB}$ to quantify noise in NILM datasets and aims to compare appliance power consumption with noise power.

## III. Metric Proposal

### A. Metric Definition

In terms of power estimation, the NILM problem is conceptually described as providing estimates $(\hat{x}_t^{(1)}, \cdots, \hat{x}_t^{(M)})$ of the real power consumption of M appliances $(x_t^{(1)}, \cdots, x_t^{(M)})$, at time $t$, that compose the aggregate power consumption $y_t$, ending up with:

$$\hat{y}_t = \sum_{i=1}^{M} \hat{x}_t^{(i)} + \eta_t \tag{1}$$

where $\hat{y}_t$ is the estimate of $y_t$ and $\eta_t$ is an error term (or noise) at $t$ that refers to measurement errors from the sensors or unmetered/unknown appliance signals.

The total or partial noise levels can be computed if the aggregate and sub-metered appliance consumption are available in the same units, e.g., active power. The SNR computes the level of a desired signal to the level of noise in the aggregate signal,

$$SNR = \frac{\bar{P}_{agg}}{\bar{P}_{noise}} = \frac{\sum_{t=1}^{T} y_t}{\sum_{t=1}^{T} |y_t - \sum_{m=1}^{M} x_t^{(m)}|} \tag{2}$$

where $\bar{P}_{agg}$ and $\bar{P}_{noise}$ are, respectively, the average aggregate power per record and the average noise per record. Due to the wide dynamic range, SNR is typically converted into the logarithm decibel scale, i.e., $SNR_{dB} = 10 \log_{10} SNR$.

The $SNR$ as defined in Eq. (2) can be written as the sum of two components,

$$\frac{\sum_{t=1}^{T} y_t - x_t^{(k)}}{\sum_{t=1}^{T} |y_t - \sum_{m=1}^{M} x_t^{(m)}|} + \frac{\sum_{t=1}^{T} x_t^{(k)}}{\sum_{t=1}^{T} |y_t - \sum_{m=1}^{M} x_t^{(m)}|} \tag{3}$$

where the first term is the sum of the ratio between the total aggregate signal minus appliance $k$ and total noise, and the second term is the ratio between appliance $k$ total consumption and total noise. The second term of Eq. (3) is of particular interest for this work because it allows direct comparison between the amount of an appliance consumption with the noise, which can significantly impact the disaggregation performance. Henceforth, it represents a ratio between appliance consumption and noise. Notice that the second term in Eq. (3) is an overall quantity, which does not consider the impact of noise on an appliance at a time instant. Furthermore, it could be misleading when an appliance is rarely used or consumes little power. This quantity could be neglected in such a case, which is not a valid assumption.

Thus, to obtain a more informative metric over time, the mean across all instant ratios is calculated and defined as Appliance-to-Noise Ratio (ANR), for an appliance $k$:

$$ANR_k = \frac{1}{T} \sum_{t=1}^{T} \frac{x_t^{(k)}}{|y_t - \sum_{m=1}^{M} x_t^{(m)}|} \tag{4}$$

### B. General Metric Behavior

For Eq. (4), ANR is undefined if individual ratios have zero denominators. Thus, only time instants with nonzero noise power are considered for the metric calculation. There are three possible cases for the appliance-to-noise ratios at each instant:

1) When the appliance consumption is zero and noise power is nonzero, the ratio is zero. The appliance is OFF, and there is noise for specific time instants.
2) When both the appliance consumption and noise power are nonzero, the ratio is greater than zero. The appliance is ON, and there is noise for specific time instants.
   a) If noise is equal to or greater than the appliance consumption, the ratio is between zero and one. The ratio is close to zero for minimal appliance consumption and significant noise power.
   b) If the appliance consumption is greater than noise, the ratio is greater than one. The ratio is notably greater than one if the appliance consumption is much larger than the noise.

   Hence, since the ANR averages across the ratios at each data point, the following scenarios are possible:
   - ANR$=0$, if the appliance is OFF for all time instants, i.e., all ratios are described by case B(1).
   - ANR$\neq 0$, then $0 < ANR < 1$ if the appliance is ON for a short time, i.e., most ratios in case B(1) or noise power overcomes appliance consumption for most time instants (case B(2a)). The latter is expected for low-power appliances, such as televisions.

     Alternatively, $ANR > 1$ if the appliance is ON for a long time (opposite of case B(1)), or appliance consumption overcomes noise power for most time instants. The latter is expected for high-power appliances like washing machines.
3) If there is no dataset noise (the aggregate data matches the ground truth for each instant), the ANR is undefined as reporting dataset noise is irrelevant.

## IV. Metric Assessment Methodology

Three experiments were conducted to analyze the appropriateness of ANR for reporting noise. First, the metric's

sensitivity to different sampling rates is assessed in a real-world dataset to inspect the reliability and representativeness of the results regarding the proportion of noise with respect to the appliance. Second, the ANR is computed hourly, bi-weekly, and monthly, highlighting patterns in noise proportion across different data scenarios and comparing results to SNR. Finally, a possible relationship between the performance of NILM algorithms and noise is assessed using the proposed metric to complement the analysis of the results. The code for reproducing the experiments is available in https://anonymous.4open.science/r/IEEE-TIM-21C7.

### A. Dataset and Selected Appliances

In the following experiments, the REFIT dataset [18] is used for demonstrating ANR computation. It includes aggregate and appliance measurements in the same power units, enabling ANR calculation. The consumption data from 20 UK houses is timestamped at approximately 8-second intervals. The experiment period is one year, and active power is the measuring unit. Various appliance types were considered, such as multi-state appliances like washing machines and dishwashers (type II) [19], [20], as well as permeant appliances like televisions and computers (type IV) [19], [21]. The inclusion of different appliance types ensures more generalizable conclusions.

### B. Experiment 1 - Sensitivity to Sampling Rates

This experiment illustrates ANR computation for different sample rates in order to assess to what extent the ANR varies with granularity and how it can impact the conclusions obtained from the results. The chosen sample rates are: $\approx 1/8$ Hz, $1/60$Hz, and $1/300$ Hz. The $1/8$ Hz scenario corresponds to the original sample rate, while the other two data granularities were obtained through down-sampling. Notably, the $1/300$ Hz rate is a more substantial down-sampling compared to the original rate. The sample rates were selected to track every appliance activation, i.e., any appliance transition from an off-state to an on-state.

The houses were selected such that each appliance was present, ensuring similar conditions for noise and building/consumer characteristics. The experiment focuses on washing machines, dishwashers, televisions, and computers in houses 1, 5, 6, 15, 16, 18, and 20. The variation of ANR results with sample rate is analyzed for each appliance, comparing them with SNR for each house.

### C. Experiment 2 - Computation of ANR for Different Appliances and Comparison with SNR

This experiment computes ANR for different data scenarios: hourly, bi-weekly, and monthly. Only house 1 from the REFIT dataset is considered for illustration, with a sample rate of $1/60$. At first, ANR is computed hourly for each appliance and is compared to appliance consumption, noise, and SNR. The advantage of using ANR over SNR is inspected. A similar procedure is followed for bi-weekly and monthly scenarios.

### D. Experiment 3 - Assessment of the Relationship between NILM Performance and the Dataset Noise using ANR

This experiment evaluates a potential relationship between ANR and disaggregation performance. For illustration, the performance and ANR are obtained hourly, focusing on house 1. The sample rate is set at $1/60$ Hz. The impact of noise on disaggregation for high-power and low-power appliances is evaluated, because noise has different effects depending on the appliance type. DNN-NILM algorithms like Window Gated Recurrent Units (WGRU) and Attention Recurrent Neural Networks (ARNN) are chosen for the disaggregation experiments, as DNNs typically outperform traditional NILM algorithms [22], [23]. The implementation of WGRU and ARNN follows the architectures in *NILMTK-Contrib* package [24] and [9]. The employed algorithms are trained for a maximum of 150 epochs, with batch size equal to 128.

To avoid underfitting and overfitting, stop patience of 50 epochs and an early stopping criterion are considered, respectively. The sliding window size for WGRU and ARNN should be different for each appliance due to different activation durations and patterns [25], [26]. For the washing machine, dishwasher, television and computer, a window size of 50, 15, 10 and 10 samples was considered, respectively.

The data is divided into training and testing, with $67\%$ and $33\%$ apportioned to each. In case of missing data, the data rows are dropped. The algorithm performance is assessed for each hour of the day. For performance evaluation, a normalized Mean Absolute Error (MAE) [27], [28] is considered. MAE assigns equal importance to all errors, which is crucial when examining performance variations across appliances with different consumption levels. The normalization constant is chosen to be the standard deviation, which enables to weigh MAE based on the variability of appliance consumption values. This approach ensures accurate performance assessment even for appliances with lower or higher power consumption.

In terms of software, the disaggregation experiments were carried out in *Python* 3.6.8, with Keras [29] running on the Tensorflow [30] backend. The cuDNN library (version 8.1.0) has been installed for GPU-accelerated calculations. Hardware-wise, the computer consists of an Intel i7 $-$ 8700k CPU, an NVIDIA 1080TI graphics card, and 64 GB of RAM.

## V. Results and Discussion

### A. Experiment 1

The ANR is expected to be more prominent for the washing machine and dishwasher than for the computer and television. Fig. 1 shows that this assumption holds, considering, for instance, $1/60$ Hz granularity. Furthermore, in Fig. 1, ANR does not vary significantly across the selected sample rates. This is likely because the selected rates capture most appliance activations. If lower rates were used (e.g., 1 sample for every 15 minutes), ANR values could vary because not all appliance activations would be captured. For instance, if a TV or computer is turned on for 10 minutes, it may be missed. Hence, the effectiveness of ANR for reporting noise with respect to appliance consumption also depends on the

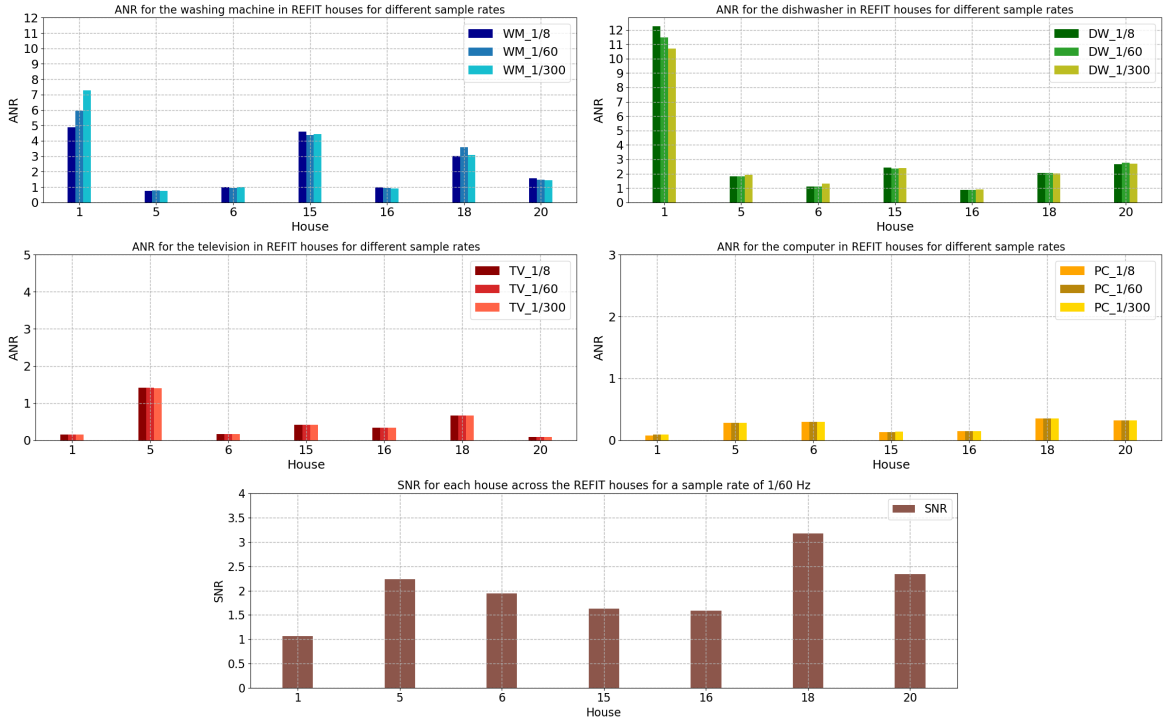


Fig. 1. ANR for washing machine, dishwasher, television and computer for different sample rates across the houses studied in experiment 1. The SNR across the houses for a sample rate of 1/60 Hz is also depicted.

sampling rate. With appropriate rates, ANR provides reliable reports.

In terms of SNR, houses 18 and 20 have higher SNRs, while houses 1 and 15 have lower ratios (Fig. 1). However, this does not imply that appliance consumption in houses 18 and 20 exceeds the noise significantly. If ANR is considered, the same can only be concluded for some appliances, with the washing machine having the largest ANR in houses 1 and 15, while one of the lowest ANR occurs in house 20. Therefore, while SNR is useful for overall noise reporting, ANR offers a more detailed analysis of noise with respect to each appliance.

## B. Experiment 2

*1) Daily hour consumption:* From Fig. 2, the washing machine and dishwasher appliances have similar ANR patterns throughout the day, with one small peak at night and a larger peak during the day. The ANR pattern is different for the television and computer, as it increases throughout the day and reaches a peak in the evening. The ANR is more significant for the washing machine and dishwasher than the television and computer. The SNR also varies throughout the day, reaching peak values between 12 am and 3 pm. However, the SNR pattern does not resemble the ANR patterns for each appliance. Hence, ANR is more suitable than SNR for analyzing and reporting noise at the appliance level.

By inspecting Fig. 3, the ANR patterns are reflected in those appliances' consumption. If one picks, for instance, the dishwasher consumption at 9 am, it is smaller as compared to 5 am and 10 am. Given the similarity in the total noise power at these times, the ANR value is naturally lower at 9 am as compared to 5 am and 10 am. Also, at 8 pm, the total dishwasher consumption is smaller than that at 5 am and 10 am, which is supported by a lower total consumption at 8 pm. If, instead, one examines each hour of the day separately, it can also be seen that a larger power consumption leads to a larger ANR, and vice-versa. For instance, at 22 pm, the total consumption for the washing machine and dishwasher is much smaller (near zero) in comparison with the television and computer, which is a tendency that holds in terms of ANR.

Hence, the ANR is predicted satisfactorily by inspecting the total appliance's consumption and noise.

*2) Bi-weekly consumption:* Fig. 4 shows a considerable variation of the ANR when computed bi-weekly for all appliances. Again, ANR is greater for the washing machine and dishwasher than for the television and computer. From Fig. 5, the total consumption varies across the biweekly periods for each appliance. For instance, for the computer, the total consumption increases across biweekly periods, while the opposite seems to happen for the television. The fluctuation in the total noise power accounts for the alterations in the ANR patterns.

For the SNR, there is significant variation throughout the bi-weekly periods; the SNR decreases until week 19 and then increases. Total noise power is also not conclusive concerning SNR, which depends on the aggregate consumption (Fig. 5). Therefore, SNR does not reflect the patterns obtained with ANR for each appliance. Again, ANR is shown to be more suitable than SNR for analyzing and reporting noise at the appliance level.

*3) Monthly consumption:* Fig. 6 shows that ANR varies throughout the months, with the washing machine and dish-

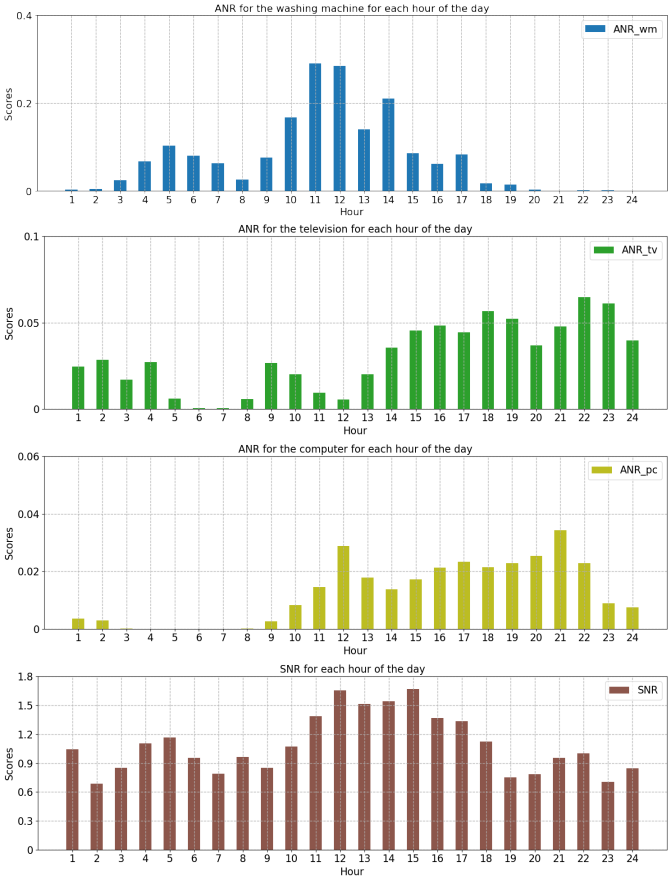


Fig. 2. ANR for the washing machine, dishwasher, television, and computer, and SNR in house 1 for each hour of the day in experiment 2.

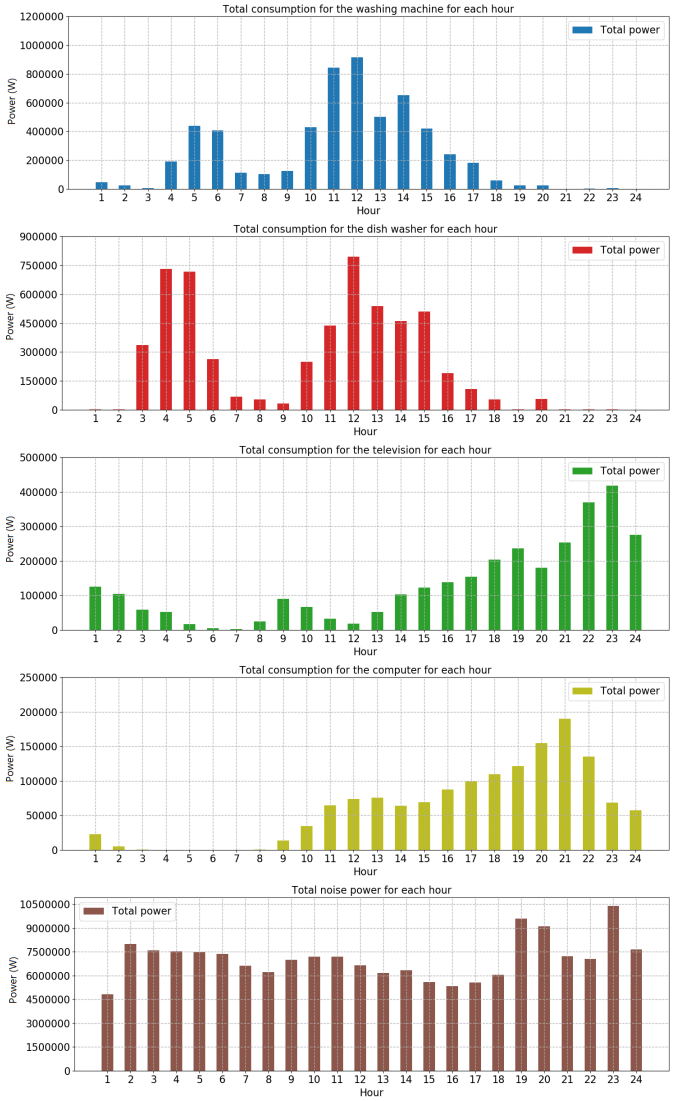


Fig. 3. Total power consumption of the washing machine, dishwasher, television and computer, and total noise power in house 1 for each hour of the day in experiment 2.

washer showing greater ANR than the television and computer. The total consumption for the appliances varies across the months and resembles what was obtained for biweekly periods. Again, the alterations in ANR are explained by a joint analysis total appliance consumption and noise power across the months (Fig. 7). Furthermore, the ANR results for each appliance are not reflected in SNR. Once again, ANR is a more appropriate metric than SNR for analyzing and reporting the percentage of noise with respect to each appliance.

### C. Experiment 3

The WGRU algorithm generally achieved the best performance for each appliance in the experiments. The performances for the selected appliances in house 1 for each hour of the day are shown in Fig. 8. For the washing machine, the error increases at night (9 pm to 12 pm), in which the ANR is small (Fig. 2). Although the washing machine is minimally utilized for those hours, the algorithm predicts appliance power and the disaggregation performance is worse. Furthermore, the error decreases in the morning and afternoon (1 am to 8 pm), in which the ANR is larger than zero. Hence, for the washing machine, the performance improves with the ANR, and vice-versa. For the dishwasher, the conclusion is similar, as the error increases at night and early morning (6 pm to 2 am), in which the ANR is small. In the remaining hours, the error is small and the ANR is large. For the television, the ANR slightly increases throughout the day and the error decreases in the same proportion. The ANR increases throughout the day because the television is ON, in contrast with nighttime, in which the television has minimal usage. During the night, even with small noise, the ANR for the television tends to be small. For the computer, the error is clearly larger between 4 am and 7 am, in which the ANR is very small (Fig. 2). In the remaining hours, the ANR increases and the performance improves.

By inspecting the distribution of power consumption for each appliance and the disaggregation performances, the WGRU algorithm predicts appliance power for hours of minimal usage, for instance, for the washing machine and dishwasher, where the disaggregation performances are clearly worse at night compared to daytime. This is captured by the normalized MAE and impacts the association observed between ANR and the disaggregation performance. Hence, in

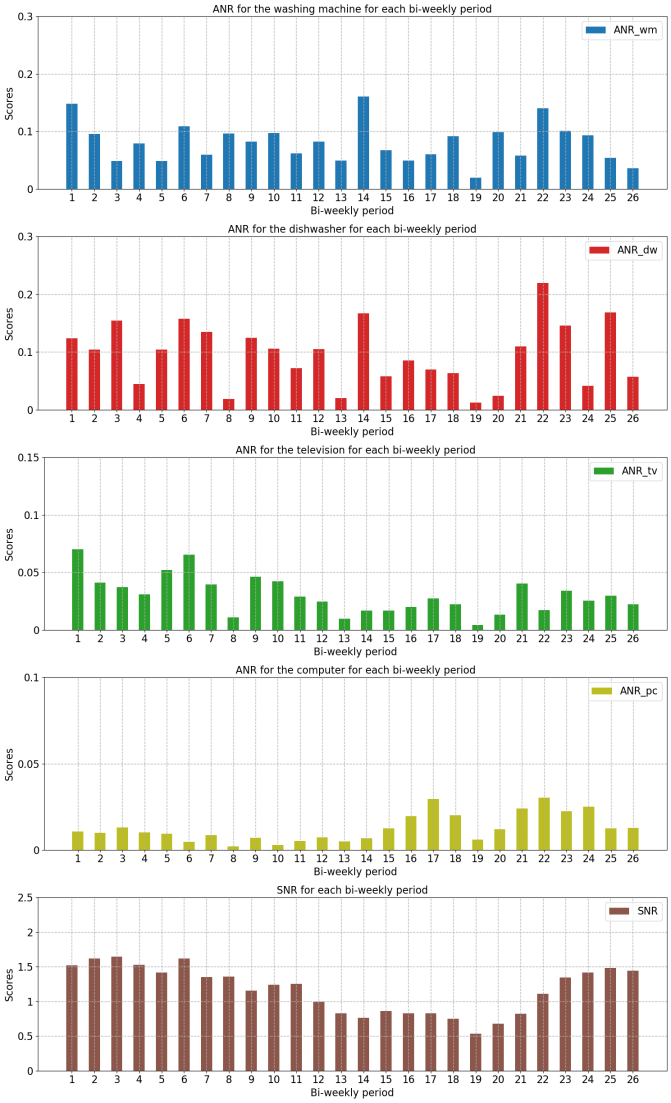


Fig. 4. ANR for the washing machine, dishwasher, television, and computer, and SNR for house 1 for each bi-weekly period in experiment 2.

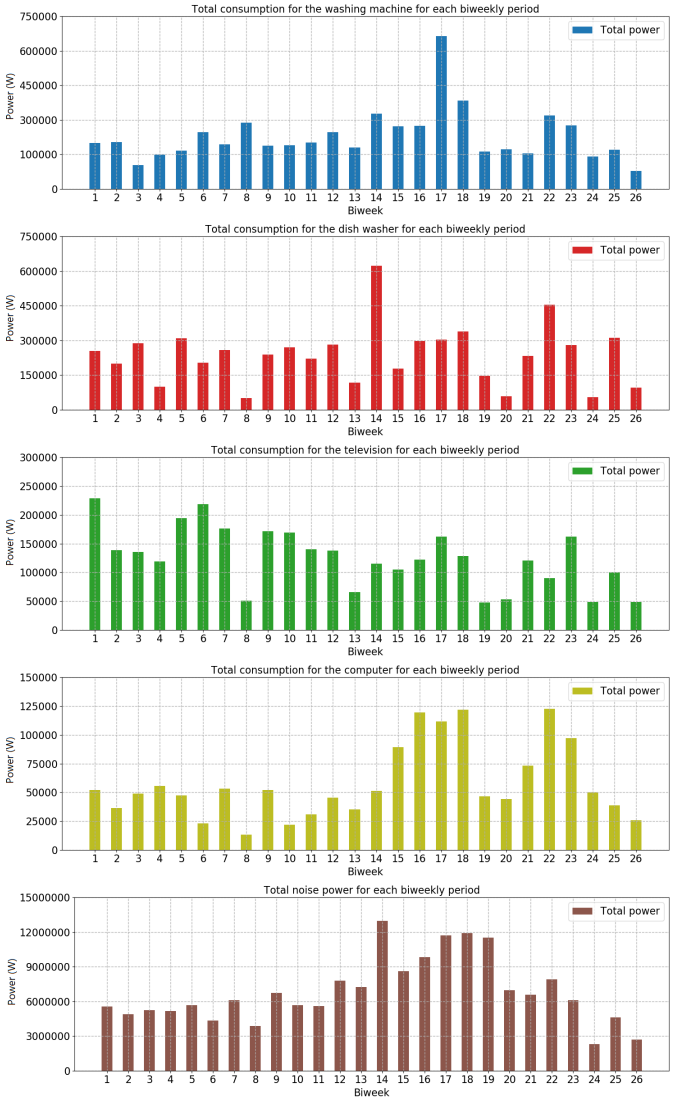


Fig. 5. Total power consumption of the washing machine, dishwasher, television and computer, and total noise power in house 1 for each bi-weekly period in experiment 2.

the particular case of house 1, it is safe to say that there is a tendency for better performances when ANR increases and vice-versa.

Considering Fig. 2, the SNR varies across the hours of the day, with peak values between 11 am and 5 pm. However, the SNR is insufficient to explain any variation in performance for each appliance throughout the day. On the other hand, although the algorithm performance for each appliance tends to improve when ANR increases, it can be hypothesized that noise alone is not an indicator of the complexity of a given dataset concerning a particular appliance. Nevertheless, it should be stressed that this experiment considered only one household. As such, to obtain more definitive answers regarding the relationship between ANR and performance additional experiments should be performed with a higher number of datasets and appliances.

## VI. Conclusion

### A. Research Implications and Potential Applications

In this paper, a metric that reports the noise proportion in NILM datasets with respect to individual appliances was presented to complement existing noise metrics, such as SNR.

Three experiments were conducted to apply the proposed metric in different scenarios. The first experiment examined various sample rates, emphasizing the need for covering appliance activations for accurate reports. The second experiment explored different data scenarios (hourly, bi-weekly, and monthly), demonstrating the metric's versatility beyond using the whole dataset. In contrast to SNR, ANR provided specific appliance-level information for each scenario. The third experiment assessed the relationship between noise and disaggregation performance using ANR, revealing valuable insights at the appliance level for analyzing disaggregation performance. If the analysis is extended to more buildings, the proposed metric can help define NILM performance bench-

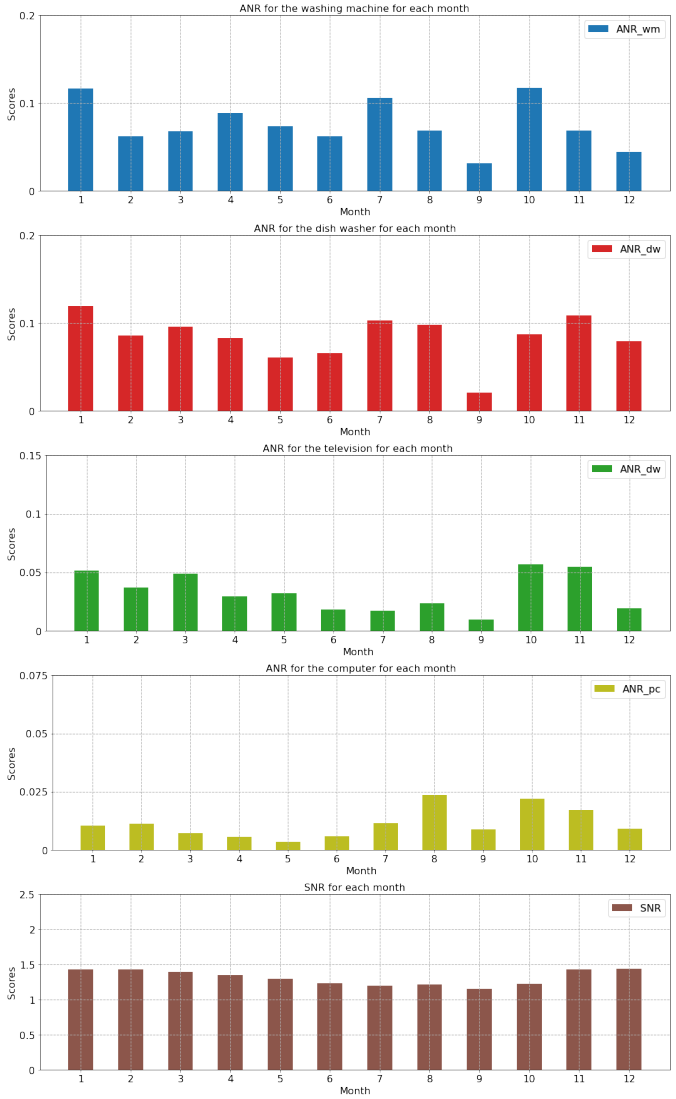


Fig. 6. ANR for the washing machine, dishwasher, television, and computer, and SNR for house 1 for each monthly period in experiment 2.

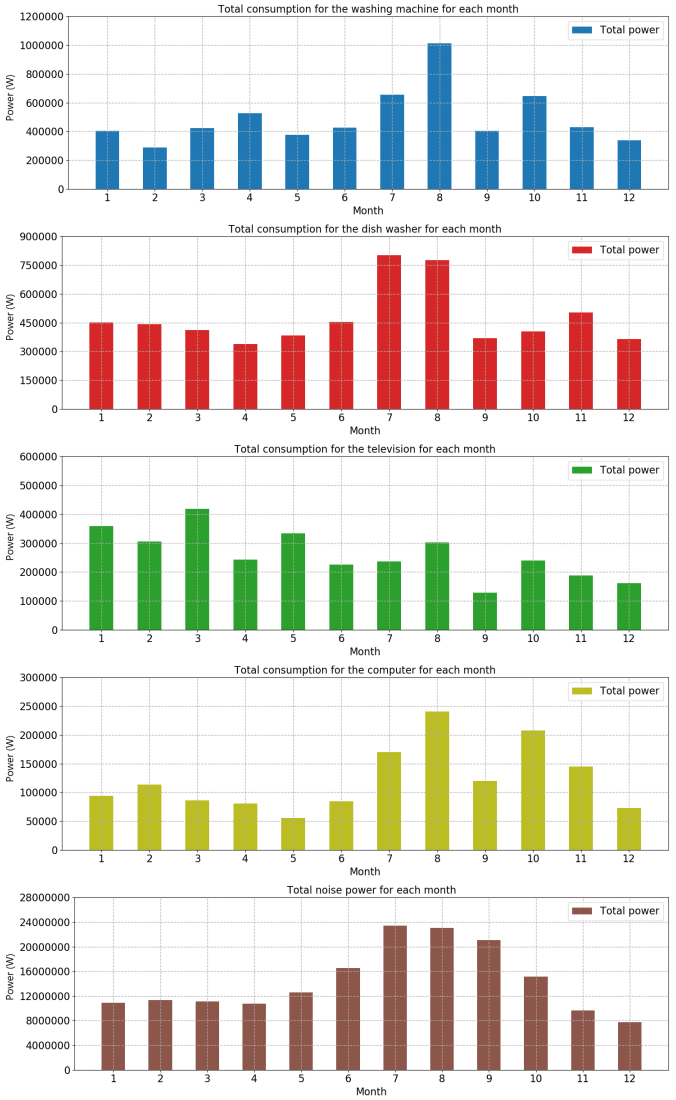


Fig. 7. Total power consumption of the washing machine, dishwasher, television and computer, and total noise power in house 1 for each monthly period in experiment 2.

marks based on the possible relationship between noise and disaggregation performance.

This measure can assist researchers in establishing noise thresholds for satisfactory disaggregation performance in specific appliances. Hence, usage pattern identification and fault detection can improve, with a direct impact on the reliability and accuracy of NILM. Furthermore, ANR can aid in the improvement or development of NILM algorithms. Specifically, the analysis of ANR could allow for more precise selection of hyperparameters in the algorithms based on the appliance. Finally, the proposed metric could also be extended to other machine learning problems, such as Blind Source Separation (BSS) [31]. Since NILM is a BSS problem with a single channel, ANR could be generalized for more channels. The proposed ANR measure is an essential addition to help deploy NILM tools in a real-world setting.

### *B. Limitations and Future Work*

Although the purpose of this work is achieved, some limitations and future work suggestions are identified. In the metric assessment, ANR was computed using a benchmark dataset. Further investigation with other datasets would yield more generalizable results. Also, since many datasets only include aggregate consumption data from buildings, analyzing the estimation of ANR from external information, such as usage patterns, could be valuable.

Regarding the third experiment, including more algorithms would enhance comparisons and generalizability. In this experiment, only houses that contained all four appliances were considered; therefore, the number of dwellings considered was small and from a specific geographical area. Large-scale empirical studies are recommended to better understand the impact of noise on disaggregation.

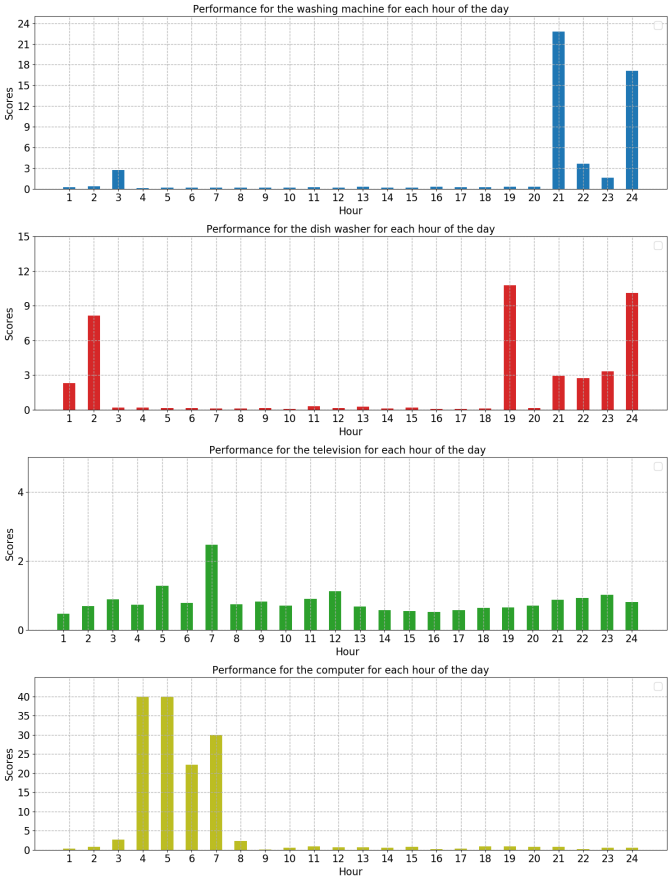


Fig. 8. Disaggregation performance - Normalized MAE - of the WGRU algorithm for the washing machine, dishwasher, television, and computer appliances for each hour of the day in house 1 in experiment 3.

## REFERENCES


[1] J. Batalla-Bejerano, E. Trujillo-Baute, and M. Villa-Arrieta, "Smart meters and consumer behaviour: Insights from the empirical literature," *Energy Policy*, vol. 144, p. 111610, Sep. 2020.

[2] G. W. Hart, "Nonintrusive appliance load monitoring," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870–1891, 1992.

[3] R. Gopinath, M. Kumar, C. P. C. Joshua, and K. Srinivas, "Energy management using non-intrusive load monitoring techniques-state-of-the-art and future research directions," *Sustainable Cities and Society*, p. 102411, 2020.

[4] S. Gupta and A. Gupta, "Dealing with noise problem in machine learning data-sets: A systematic review," *Procedia Computer Science*, vol. 161, pp. 466–474, 2019.

[5] R. Dong, L. Ratliff, H. Ohlsson, and S. S. Sastry, "Fundamental limits of nonintrusive load monitoring," in *Proceedings of the 3rd international conference on High confidence networked systems*, 2014, pp. 11–18.

[6] D. Egarter, M. Pöchacker, and W. Elmenreich, "Complexity of power draws for load disaggregation," *arXiv preprint arXiv:1501.02954*, 2015.

[7] C. Klemenjak, S. Makonin, and W. Elmenreich, "Investigating the performance gap between testing on real and denoised aggregates in non-intrusive load monitoring," *Energy Informatics*, vol. 4, no. 1, pp. 1–15, 2021.

[8] S. Makonin and F. Popowich, "Nonintrusive load monitoring (nilm) performance evaluation," *Energy Efficiency*, vol. 8, no. 4, pp. 809–814, 2015.

[9] E. Gomes and L. Pereira, "Pb-nilm: Pinball guided deep non-intrusive load monitoring," *IEEE Access*, vol. 8, pp. 48 386–48 398, 2020.

[10] C. Klemenjak, S. Makonin, and W. Elmenreich, "Towards comparability in non-intrusive load monitoring: On data and performance evaluation," in *2020 IEEE power & energy society innovative smart grid technologies conference (ISGT)*. IEEE, 2020, pp. 1–5.

[11] M. N. Meziane, P. Ravier, G. Lamarque, J.-C. Le Bunetel, and Y. Raingeaud, "High accuracy event detection for non-intrusive load monitoring," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2452–2456.

[12] C. Lu, L. Ma, T. Xu, G. Ding, C. Wu, and X. Jiang, "Non-intrusive load monitoring method based on improved differential evolution algorithm," in *2019 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*. IEEE, 2019, pp. 279–283.

[13] A. T. Saseendran, L. Setia, V. Chhabria, D. Chakraborty, and A. Barman Roy, "Impact of noise in dataset on machine learning algorithms," *In Machine Learning Research*, pp. 0–8, 2019.

[14] V. I. Slyusar, "A method of investigation of the linear dynamic range of reception channels in a digital antenna array," *Radioelectronics and Communications Systems*, vol. 47, no. 9, pp. 29–38, 2004.

[15] D. Egarter, V. P. Bhuvana, and W. Elmenreich, "Paldi: Online load disaggregation via particle filtering," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 2, pp. 467–477, 2014.

[16] R. Feng, W. Yuan, L. Ge, and S. Ji, "Nonintrusive load disaggregation for residential users based on alternating optimization and downsampling," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2021.

[17] S. Chen, B. Zhao, M. Zhong, W. Luan, and Y. Yu, "Non-intrusive load monitoring based on self-supervised learning," *IEEE Transactions on Instrumentation and Measurement*, 2023.

[18] D. Murray, L. Stankovic, and V. Stankovic, "An electrical load measurements dataset of united kingdom households from a two-year longitudinal study," *Scientific data*, vol. 4, no. 1, pp. 1–12, 2017.

[19] A. Zoha, A. Gluhak, M. A. Imran, and S. Rajasegarar, "Non-intrusive load monitoring approaches for disaggregated energy sensing: A survey," *Sensors*, vol. 12, no. 12, pp. 16 838–16 866, 2012.

[20] W. Kong, Z. Y. Dong, B. Wang, J. Zhao, and J. Huang, "A practical solution for non-intrusive type ii load monitoring based on deep learning and post-processing," *IEEE Transactions on Smart Grid*, vol. 11, no. 1, pp. 148–160, 2019.

[21] M. Coleman, N. Brown, A. Wright, and S. K. Firth, "Information, communication and entertainment appliance use—insights from a uk household study," *Energy and Buildings*, vol. 54, pp. 61–72, 2012.

[22] C. Zhang, M. Zhong, Z. Wang, N. Goddard, and C. Sutton, "Sequence-to-point learning with neural networks for non-intrusive load monitoring," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[23] C. Nalmpantis and D. Vrakas, "Machine learning approaches for non-intrusive load monitoring: from qualitative to quantitative comparation," *Artificial Intelligence Review*, vol. 52, no. 1, pp. 217–243, 2019.

[24] N. Batra, R. Kukunuri, A. Pandey, R. Malakar, R. Kumar, O. Krystalakos, M. Zhong, P. Meira, and O. Parson, "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*, 2019, pp. 193–202.

[25] O. Krystalakos, C. Nalmpantis, and D. Vrakas, "Sliding window approach for online energy disaggregation using artificial neural networks," in *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, 2018, pp. 1–6.

[26] D. Garcia-Perez, D. Pérez-López, I. Diaz-Blanco, A. González-Muñiz, M. Domínguez-González, and A. A. C. Vega, "Fully-convolutional denoising auto-encoders for nilm in large non-residential buildings," *IEEE Transactions on Smart Grid*, vol. 12, no. 3, pp. 2722–2731, 2020.

[27] P. Huber, A. Calatroni, A. Rumsch, and A. Paice, "Review on deep neural networks applied to low-frequency nilm," *Energies*, vol. 14, no. 9, p. 2390, 2021.

[28] L. Pereira and N. Nunes, "Performance evaluation in non-intrusive load monitoring: Datasets, metrics, and tools—a review," *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, vol. 8, no. 6, p. e1265, 2018.

[29] F. Chollet and others, "Keras: The Python Deep Learning library," p. ascl:1806.022, Jun. 2018.

[30] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 2016, pp. 265–283.

[31] M. Pal, R. Roy, J. Basu, and M. S. Bepari, "Blind source separation: A review and analysis," in *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*. IEEE, 2013, pp. 1–5.