An Evaluation of NILM Approaches on Industrial Energy-Consumption Data

Florian Kalinke
florian.kalinke@kit.edu
Karlsruhe Institute of Technology
Germany

Pawel Bielski
pawel.bielski@kit.edu
Karlsruhe Institute of Technology
Germany

Snigdha Singh
snigdha.singh@kit.edu
Karlsruhe Institute of Technology
Germany

Edouard Fouché
edouard.fouche@kit.edu
Karlsruhe Institute of Technology
Germany

Klemens Böhm
klemens.boehm@kit.edu
Karlsruhe Institute of Technology
Germany

ABSTRACT

Load disaggregation methods infer the energy consumption of individual appliances from their aggregated consumption. This facilitates energy savings and efficient energy management. However, most existing work on load disaggregation has only considered household settings. This may be due to companies preferring to not share their data, rendering such data hardly available.

This article makes three contributions: First, we compare data describing the energy consumption of two facilities and of households. Second, we study the performance of seven prominent load disaggregation algorithms on industrial data and compare it to the one on household data. Our results indicate a performance gap on individual appliances. Third, we publish a tool converting an industrial data set to a standard format for load disaggregation, to facilitate further research and benchmarking in the field.

CCS CONCEPTS

• **Computing methodologies → Machine learning**; • **Hardware → Power and energy**; **Energy metering**.

KEYWORDS

Load Disaggregation, Non-intrusive Load Monitoring (NILM), Industrial Data, Energy Disaggregation, Algorithm Evaluation

ACM Reference Format:

Florian Kalinke, Pawel Bielski, Snigdha Singh, Edouard Fouché, and Klemens Böhm. 2021. An Evaluation of NILM Approaches on Industrial Energy-Consumption Data. In *The Twelfth ACM International Conference on Future Energy Systems (e-Energy '21), June 28-July 2, 2021, Virtual Event, Italy*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3447555.3464863

1 INTRODUCTION

Motivation. Better energy management is expected to reduce the global energy consumption [8]. With the industrial sector being responsible for 45.7% of the consumption in Germany [6], the potential for savings in this sector is large. Load disaggregation, or Non-intrusive Load Monitoring (NILM), often is a first step towards more effective energy management, as proposed by [12]: First, NILM determines the energy consumption of individual appliances based on the aggregated consumption of the household or of the facility. This fine-grained information then helps to find sources of high consumption and to subsequently act upon them. Examples of such subsequent actions are peak shaving or job scheduling [2].

However, most existing disaggregation methods [16, 19, 24] have only been tested against household benchmarks, i.e., data describing the consumption of households. One reason may be a lack of data on energy consumption in industry that is openly available. While researchers are conscious of the importance of industry data [1, 25], we are aware of only one such data set for load disaggregation, together with a disaggregation method for this setting [20]. Next, data should also be in a format usable with NILMTK [3] (NILM Toolkit), the framework commonly employed for respective benchmarks. [1] is the only industrial data set available with this characteristic.

The core hypothesis we study in this article is that (a) industrial data differs from household data, due to different behavior of the machines and the appliances measured, and, in consequence, (b) NILM algorithms that work well on household data do not necessarily do so on industrial data sets. We also study to which extent patterns in energy-consumption data vary across industries.

Contributions. This article makes the following contributions: (1) We identify and describe specific differences between industrial data sets. One example is the presence of temporal dependencies originating from industrial processes. (2) By means of extensive experiments, we show that most disaggregation algorithms, which work well on household data, do not necessarily do so on industrial data. Our results indicate that building robust and versatile load disaggregation algorithms calls for benchmark data that is sufficiently diverse. (3) We publish a data converter for a data set collected within an electronics factory in Germany, the HIPE data set [7]. This converter is compatible with the NILMTK format, extending the set of available industrial benchmarks. This also helps reproducing our results.

2 RELATED WORK

[21] is a recent overview of NILM techniques. Batra et al. [3] propose the NILMTK framework to help with the standardization and reproducibility of load disaggregation experiments. The framework

e-energy '21, June 28–July 2, 2021, Virtual Event, Italy

includes common methods, such as [11, 13]. Further algorithms are available from `nilmtk-contrib`, e.g., [16, 19, 24]. NILMTK also includes benchmark data sets that are publicly available. However, others have commented that the fact that there only are 14 household data sets available is limiting [22].

[14] are first to address load disaggregation in the industrial setting, but they have not released their data. [1] describes the first industrial data set that is available in the NILMTK format. The data is collected in a poultry feed factory. The authors also published an algorithm tailored to their use case [20]. [7] describes HIPE, an energy-consumption data set from an electronics factory in Germany. We use this data as a use case and as our benchmark for load disaggregation in the industrial setting.

3 DISAGGREGATION METHODS

Starting with [12, 13], many algorithms that disaggregate energy data exist. They take a sum of loads $P(t)$ from appliances $i = 1, \dots, n$ at each point in time t as input. The goal is to find the individual loads of each appliance, which are assumed to be additive [25]:

$$P(t) = p_1(t) + p_2(t) + \cdots + p_n(t).$$

Depending on the type of appliance, the behavior of $p_i(t)$ differs. [17] group appliances into three types: Type I appliances, which only have an "on" and an "off" state, e.g., light bulbs. Type II appliances, which have several but fixed states, e.g., washing machines. Type III appliances, whose demand is continuous. Type III are often called Variable Frequency Drive (VFD) or Continuously Variable Devices (CVD). They are the hardest to disaggregate [9, 21].

In industry, Type III appliances/machines are ubiquitous: Think of a workshop with drilling and milling machines. Their power demands depend on, say, the engine speeds and are not fixed.

We now review the methods studied in this article. We use implementations from the NILMTK[1] framework or `nilmtk-contrib`[2] [4]. We exclude AFHMM [23] because of its long runtime.

Combinatorial Optimization (CO) [12, 13] This is the first algorithm for load disaggregation. The algorithm identifies the power demands of each appliance and then chooses a subset of all appliances at each point in time so that the sum of their demands approximates the actual demand.

Factorial Hidden Markov Model (FHMM) [11] FHMM models each appliance as a Hidden Markov Model. Each state corresponds to a particular power demand and each transition models a change in power demand. The algorithm then fits these models to the observed power demand.

Seq2Seq (S2S) [16] The approach trains one Deep Neural Network per appliance and learns its behavior based on sequences of the input data.

Recurrent Neural Network (RNN) [16] Instead of relying only on input sequences like S2S, this network features memory elements to base its output on previous inputs.

Denoising Autoencoder (DAE) [16, 19] This approach uses an Autoencoder to learn how to filter the "true" signal of an appliance from the aggregate signal by removing the "noise" from other appliances running simultaneously.

[1] https://github.com/nilmtk/nilmtk
[2] https://github.com/nilmtk/nilmtk-contrib

Seq2Point (S2P) [24] S2P is similar to S2S but outputs a single value for a given appliance using a Convolutional Neural Network.

WindowGRU (WGRU) [19] WGRU uses a Recurrent Neural Networks on a sliding window. Instead of focusing on a single point in time, WGRU uses a longer sequence of past observations, the "window", to predict the consumption of an appliance.

Mean Mean prediction provides the baseline by always predicting the mean demand of an appliance.

By design, CO and FHMM cannot deal with Type III appliances, i.e., most machines in an industrial setting.

4 DATA SET

4.1 Description

The HIPE data set used here covers three months of energy consumption in an electronics production site at KIT in Karlsruhe, Germany, consisting of 10 machines. It is low-frequency, with measurements taken every 5 seconds, containing active, reactive and apparent power. Figure 1 shows the demand curves of one week, with the curves of IMDELD [1] and REDD [18] for comparison.

The data has been collected from 2017-10-01 to 2018-01-01. The machines did not run at the end of December. So we limit our experiments to the period from 2017-10-01 to 2017-12-21.

The owners of the plant had informed us that machine usages depend on each other. Figure 2 shows this. For example, one vacuum pump is active only when the primary one does not suffice. As another example, the data shows a dependence between the soldering oven and the screen printer (which is expected). In comparison, such strong dependencies are less common in household data.

Another characteristic of industrial machines is the continuous nature of their energy demand (Type III) [14]. Next, [5] states that more appliances tend to run in parallel in commercial buildings than in households. Like [5, 22], we hypothesize that these points will render disaggregation more difficult. The HIPE data set allows to study these important issues: It contains three Type I (Vacuum Pump I/II, Screen Printer), three Type II (Pick and Place Unit, Chip Press, Washing Machine), and four Type III machines (High Temperature Oven, Chip Saw, Soldering Oven, Vacuum Oven). Figure 2 shows that up to eight out of ten machines may run in parallel.

We publish our data set converter, which renders the HIPE data set [7] directly usable with NILMTK, together with the source used to produce the figures on GitHub[3].

4.2 Comparison

In the following, we compare the HIPE data to REDD, a household data set, and to IMDELD, an industrial data set.

Appliance Activity The usage of machines in both industrial data sets only occurs during working hours. This is common for commercial buildings [5]. With REDD in turn, appliances may run at any time.

Cyclic Patterns Patterns which repeat themselves, like the fridge in REDD, do not occur in HIPE. In IMDELD however, machine usage repeats throughout all working days. While [5] claims that these so-called temporal dependencies occur mainly in household data, IMDELD is a counterexample.

[3] https://github.com/flopska/hipe-nilm

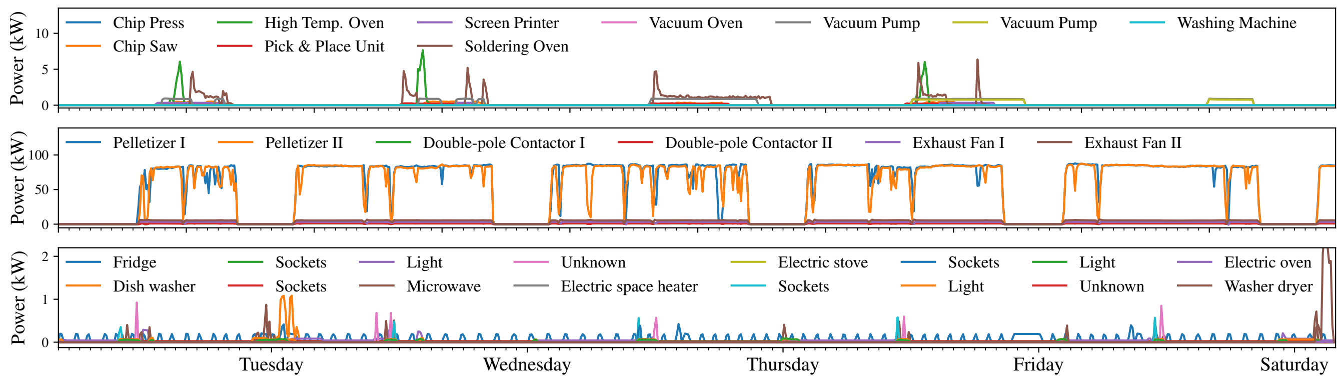

Figure 1: Active power demand of all monitored machines in the HIPE (top), IMDELD (middle), and REDD (bottom) data sets.

Appliance Types HIPE contains Type I, II, and III appliances, while REDD only contains Type II appliances, and IMDELD only contains Type I appliances.

Data with only Type I appliances does not match what others expect from an industrial setting [14], so providing and studying another data set clearly has its benefits.

Total Demand The total energy demand is larger in both industrial data sets than it is in REDD.

Parallel Usage All data sets feature several appliances running in parallel. [22] argues that this is important for the proper evaluation of load disaggregation algorithms.

We conclude that important characteristics of the HIPE data are different from the ones of household data as well as from the one openly available industrial data set. We see this as a motivation for the investigation that follows.

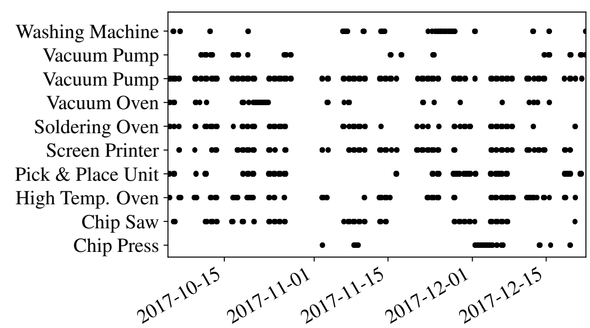

Figure 2: Times when the individual machines were turned on, i.e., their power consumption exceeded 10 Watts.

5 EXPERIMENTS

5.1 Setup & Metrics

To produce our results we use NILMTK `0.4.0.dev1+git`. We split the HIPE data into a training set (2017-10-04 to 2017-12-04) and a test set (2017-12-04 to 2017-12-21), corresponding to a total of 2 months and 17 days and a 80/20 split. For simplicity, all results use only the active power part of the observations.

For comparison, we also run the algorithms on the REDD [18] and the IMDELD [1] data sets. Here, we use the same sampling rate (6) and the time frames 2011-04-18 to 2011-05-16 / 2011-05-16 to 2011-05-24 and 2017-10-30 to 2018-03-02 / 2018-03-02 to 2018-04-03 (training/test) respectively. This also corresponds to a 80/20 split.

For FHMM, following [10], we set the number of states to 3. For the algorithms based on neural networks, we set the number of epochs to 10 and maximize the batch size depending on RAM usage. We set the other parameters to their default values. We run each algorithm 10 times, to estimate its average performance on a single appliance together with its standard deviation.

We evaluate the algorithms using two different metrics, highlighting the different strengths and weaknesses of the algorithms. Common metrics, such as MAE (Mean Absolute Error) or RMSE (Root Mean Squared Error), do not allow for a comparison across appliances and data sets if the power consumptions are different [15]. F1 is not usable with Type III appliances.

Accordingly, we use normalized errors instead. y is the ground truth, $\hat{y}$ the prediction, and superscript i refers to the appliance for which the metric is computed. The metrics are as follows:

Normalized Disaggregation Error (NDE) This metric captures the quality of the disaggregation at every time point and normalizes it by total demand [10]

$$NDE^{(i)} = \frac{\sum_{t=1}^{T}(\hat{y}_t^{(i)} - y_t^{(i)})^2}{\sum_{t=1}^{T}(y_t^{(i)})^2}.$$

Normalized Signal Aggregated Error (SAE) The SAE measures the absolute error of the energy estimation [10, 24]

$$SAE^{(i)} = \frac{|\sum_{t=1}^{T}\hat{y}_t^{(i)} - \sum_{t=1}^{T}y_t^{(i)}|}{\sum_{t=1}^{T}y_t^{(i)}}.$$

5.2 Results

With our experiments, we measure the performance of load disaggregation algorithms across various appliances and data sets.

Table 1 contains aggregates of these values. The aggregate for HIPE (I+II+III) is the average over all appliances in this data set.

At a coarse level, S2S performs best against each data set with respect to SAE. In terms of NDE, S2P performs best on the REDD data.

Table 1: Average performance of load disaggregation algorithms per appliance type. The best values are in bold print.

Data	Type	CO	Mean	FHMM	WGRU	RNN	DAE	S2P	S2S
					SAE				
HIPE	I	6.18	**0.14**	6.01	13.74	0.36	0.27	0.33	0.36
	II	114.41	0.50	113.69	17.43	0.44	0.81	0.70	**0.38**
	III	134.39	0.40	88.80	2.02	0.66	**0.26**	0.31	0.32
	I+II+III	99.24	0.38	78.70	9.76	0.52	0.44	0.45	**0.35**
IMDELD	I	0.79	0.13	0.86	1.14	0.45	0.25	0.22	**0.07**
REDD	II	164.23	0.42	78.30	80.33	0.40	0.31	0.25	**0.19**
					NDE				
HIPE	I	6.36	0.86	6.07	439.50	0.57	**0.49**	0.54	0.50
	II	181.71	0.91	172.87	2631.34	0.72	0.85	0.89	**0.46**
	III	180.34	0.97	84.65	1.71	1.14	0.79	0.73	**0.72**
	I+II+III	142.13	0.93	96.59	975.54	0.87	0.74	0.74	**0.58**
IMDELD	I	0.77	0.54	0.85	15.74	2.71	0.25	0.26	**0.16**
REDD	II	3528.26	0.81	1714.85	5571.96	0.57	0.49	**0.37**	0.39

However, at a more fine-grained level on the HIPE data, the results indicate that S2S is not best for Type I and Type III appliances with SAE. Approaches based on neural networks produce comparable results while outperforming the more "classical" approaches. Mean prediction performs well overall and best in one case.

With many algorithms performing better on Type III than on Type II appliances, we see that the appliance type does not directly influence the quality of load disaggregation. This contradicts opinions from existing studies [5, 22]. In this case, the reason is the poor performance on the industrial washing machine. Looking closer, we assume this is due to its unusual energy demand: After drawing about 10kW initially, the energy usage of the industrial washing machine oscillates heavily before returning to zero.

Appendix A contains the detailed performance per appliance. From Table 2 and 3, we can see that the performance of load disaggregation algorithms mostly depends on the underlying data set and not so much on the individual appliance types.

These results indicate that benchmark data should include appliances of all types: A data set with, say, only Type III appliances, while supposedly difficult to disaggregate, might draw a wrong picture of the performance of an algorithm. Another observation of ours is that, with industrial data sets not only differing from household data but also among each other, diverse benchmarks for load disaggregation algorithms are needed.

6 FUTURE WORK

In the future, we will investigate further the reasons for the large discrepancies between the performance of different algorithms. This might facilitate the development of better load disaggregation algorithms, working across a wider range of appliances and sectors. Next, as extensive domain knowledge of machine behavior is available in industry in many cases, we also aim to incorporate it into the load disaggregation process for further improvement.

ACKNOWLEDGMENTS

This work was supported by the German Research Foundation (DFG) Research Training Group GRK 2153: *Energy Status Data – Informatics Methods for its Collection, Analysis and Exploitation.*

REFERENCES

[1] Pedro Bandeira de Mello Martins, Vagner Barbosa Nascimento, Antônio Renato de Freitas, Pedro Bittencourt e Silva, and Raphael Guimarães Duarte Pinto. 2018. Industrial Machines Dataset for Electrical Load Disagreggation.

[2] Lukas Barth, Veit Hagenmeyer, Nicole Ludwig, and Dorothea Wagner. 2018. How Much Demand Side Flexibility Do We Need?: Analyzing Where to Exploit Flexibility in Industrial Processes. In *E-Energy*.

[3] Nipun Batra, Jack Kelly, Oliver Parson, Haimonti Dutta, William J. Knottenbelt, Alex Rogers, Amarjeet Singh, and Mani B. Srivastava. 2014. NILMTK: An Open Source Toolkit for Non-Intrusive Load Monitoring. In *E-Energy*.

[4] Nipun Batra, Rithwik Kukunuri, Ayush Pandey, Raktim Malakar, Rajat Kumar, Odysseas Krystalakos, Mingjun Zhong, Paulo Meira, and Oliver Parson. 2019. Towards Reproducible State-of-the-Art Energy Disaggregation. In *BuildSys@SenSys*.

[5] Nipun Batra, Oliver Parson, Mario Berges, Amarjeet Singh, and Alex Rogers. 2014. A Comparison of Non-Intrusive Load Monitoring Methods for Commercial and Residential Buildings. *CoRR* abs/1408.6595 (2014).

[6] BDEW. 2021. Stromverbrauch in Deutschland nach Verbrauchergruppen 2020.

[7] Simon Bischof, Holger Trittenbach, Michael Vollmer, Dominik Werle, Thomas Blank, and Klemens Böhm. 2018. HIPE: An Energy-Status-Data Set from Industrial Production. In *E-Energy*.

[8] K. Carrie Armel, Abhay Gupta, Gireesh Shrimali, and Adrian Albert. 2013. Is Disaggregation the Holy Grail of Energy Efficiency? The Case of Electricity. *Energy Policy* 52 (2013).

[9] Hsueh-Hsien Chang, Hong-Tzer Yang, and Ching-Lung Lin. 2007. Load Identification in Neural Networks for a Non-Intrusive Monitoring of Industrial Electrical Loads. In *CSCWD (Selected Papers)*, Vol. 5236.

[10] Pedro Bandeira de Mello Martins, Raphael Guimaraes Duarte Pinto, and Pedro Bittencourt e Silva. 2018. Load Disaggregation of Industrial Machinery Power Consumption Monitoring Using Factorial Hidden Markov Models. In *4th International NILM Workshop*.

[11] Zoubin Ghahramani and Michael I. Jordan. 1997. Factorial Hidden Markov Models. *Machine Learning* 29, 2-3 (1997).

[12] George W Hart. 1985. *Prototype Nonintrusive Appliance Load Monitor: Progress Report 2*.

[13] G. W. Hart. 1992. Nonintrusive Appliance Load Monitoring. *Proc. IEEE* 80, 12 (1992).

[14] Emil Holmegaard and Mikkel Baun Kjærgaard. 2016. NILM in an Industrial Setting: A Load Characterization and Algorithm Evaluation. In *SMARTCOMP*.

[15] Jordan Holweger, Marina Dorokhova, Lionel Bloch, Christophe Ballif, and Nicolas Wyrsch. 2019. Unsupervised Algorithm for Disaggregating Low-Sampling-Rate Electricity Consumption of Households. *CoRR* abs/1908.10713 (2019).

[16] Jack Kelly and William J. Knottenbelt. 2015. Neural NILM: Deep Neural Networks Applied to Energy Disaggregation. In *BuildSys@SenSys*.

[17] Christoph Klemenjak and Peter Goldsborough. 2016. Non-Intrusive Load Monitoring: A Review and Outlook. *CoRR* abs/1610.01191 (2016).

[18] J Zico Kolter and Matthew J Johnson. 2011. REDD: A Public Data Set for Energy Disaggregation Research. In *Workshop on Data Mining Applications in Sustainability (SIGKDD), San Diego, CA*, Vol. 25.

[19] Odysseas Krystalakos, Christoforos Nalmpantis, and Dimitris Vrakas. 2018. Sliding Window Approach for Online Energy Disaggregation Using Artificial Neural Networks. In *SETN*.

[20] Pedro B. M. Martins, José Gabriel Rodríguez Carneiro Gomes, Vagner B. Nascimento, and Antonio R. de Freitas. 2018. Application of a Deep Learning Generative Model to Load Disaggregation for Industrial Machinery Power Consumption Monitoring. In *SmartGridComm*.

[21] Gopinath Rajendiran, Mukesh Kumar, C. Joshua, and Kota Srinivas. 2020. Energy Management Using Non-Intrusive Load Monitoring Techniques - State-of-the-Art and Future Research Directions. *Sustainable Cities and Society* 62 (2020).

[22] Andreas Reinhardt and Christoph Klemenjak. 2020. How Does Load Disaggregation Performance Depend on Data Characteristics?: Insights from a Benchmarking Study. In *E-Energy*.

[23] Cristina Rottondi, Marco Derboni, Dario Piga, and Andrea Emilio Rizzoli. 2019. An Optimisation-Based Energy Disaggregation Algorithm for Low Frequency Smart Meter Data. *Energy Informatics* 2, 1 (2019).

[24] Chaoyun Zhang, Mingjun Zhong, Zongzuo Wang, Nigel H. Goddard, and Charles Sutton. 2018. Sequence-to-Point Learning with Neural Networks for Non-Intrusive Load Monitoring. In *AAAI*.

[25] Ahmed Zoha, Alexander Gluhak, Muhammad Ali Imran, and Sutharshan Rajasegarar. 2012. Non-Intrusive Load Monitoring Approaches for Disaggregated Energy Sensing: A Survey. *Sensors* 12, 12 (2012).

A LOAD DISAGGREGATION RESULTS BY MACHINE / APPLIANCE

Table 2 and Table 3 show the means of our results together with the standard deviation of 10 runs each. Deterministic algorithms have results without standard deviation (as it is 0). Lower values are better, the best values are in bold print. The parentheses show the Type.

A.1 Normalized Signal Aggregated Error (SAE)

Table 2: SAE per appliance on the HIPE (industrial), IMDELD (industrial) and REDD (household) data of 10 runs each.

		CO	Mean	FHMM	WGRU	RNN	DAE	S2P	S2S
HIPE	Chip press (II)	39.74	0.69	37.06	0.52 ± 0.18	0.35 ± 0.10	0.32 ± 0.03	0.60 ± 0.12	**0.31 ± 0.11**
	Chip saw (III)	24.78	0.22	21.54	0.96 ± 2.22	1.52 ± 4.50	**0.08 ± 0.02**	0.13 ± 0.06	0.15 ± 0.10
	High temperature oven (III)	19.11	**0.58**	12.55	1.58 ± 1.95	0.70 ± 0.13	0.66 ± 0.01	0.70 ± 0.07	0.72 ± 0.06
	Pick and place unit (II)	5.06	**0.20**	5.09	51.33 ± 114.89	0.35 ± 0.21	0.28 ± 0.00	0.23 ± 0.02	0.28 ± 0.03
	Screen printer (I)	6.90	**0.28**	6.81	8.10 ± 18.82	0.46 ± 0.07	0.29 ± 0.03	0.37 ± 0.07	0.45 ± 0.04
	Soldering oven (III)	51.52	0.10	81.16	2.78 ± 8.31	0.31 ± 0.26	**0.10 ± 0.04**	0.25 ± 0.13	0.14 ± 0.06
	Vacuum oven (III)	442.14	0.73	239.95	2.77 ± 5.61	**0.13 ± 0.00**	0.19 ± 0.08	0.16 ± 0.09	0.26 ± 0.08
	Vacuum pump (I)	5.45	**0.01**	5.21	19.39 ± 59.96	0.26 ± 0.14	0.26 ± 0.00	0.30 ± 0.12	0.26 ± 0.04
	Washing machine (II)	298.43	0.62	298.91	**0.45 ± 0.27**	0.61 ± 0.40	1.82 ± 0.09	1.28 ± 0.46	0.54 ± 0.13
IMDELD	Double-pole Contactor I (I)	1.40	0.25	1.47	0.24 ± 0.39	0.28 ± 0.04	0.49 ± 0.03	0.39 ± 0.31	**0.13 ± 0.05**
	Double-pole Contactor II (I)	1.39	0.25	1.46	2.68 ± 7.34	0.43 ± 0.35	0.43 ± 0.05	0.27 ± 0.15	**0.10 ± 0.05**
	Exhaust Fan I (I)	0.87	0.17	1.01	0.34 ± 0.44	0.34 ± 0.15	0.35 ± 0.02	0.33 ± 0.05	**0.09 ± 0.03**
	Exhaust Fan II (I)	0.97	0.15	0.90	2.92 ± 8.54	0.29 ± 0.06	0.33 ± 0.03	0.27 ± 0.09	**0.05 ± 0.02**
	Milling Machine I (I)	0.65	0.09	0.86	0.52 ± 0.02	**0.03 ± 0.02**	0.05 ± 0.04	0.33 ± 0.02	0.04 ± 0.01
	Milling Machine II (I)	0.75	**0.01**	0.85	0.69 ± 0.21	0.06 ± 0.04	0.05 ± 0.03	0.06 ± 0.02	0.02 ± 0.01
	Pelletizer I (I)	**0.05**	0.05	0.07	0.51 ± 0.55	1.76 ± 5.28	0.10 ± 0.02	0.06 ± 0.04	0.08 ± 0.02
	Pelletizer II (I)	0.25	0.10	0.22	0.44 ± 0.36	0.23 ± 0.08	0.23 ± 0.02	**0.04 ± 0.04**	0.05 ± 0.03
REDD	Dish washer (II)	4.81	0.73	2.82	0.78 ± 2.00	0.29 ± 0.42	0.03 ± 0.04	0.03 ± 0.02	**0.03 ± 0.02**
	Electric oven (II)	1.81	0.95	2.06	0.61 ± 0.19	**0.06 ± 0.03**	0.81 ± 0.02	0.76 ± 0.04	0.85 ± 0.02
	Electric space heater (II)	1237.17	0.33	639.63	14.76 ± 45.16	1.32 ± 0.00	0.51 ± 0.09	0.26 ± 0.17	**0.08 ± 0.03**
	Electric stove (II)	210.55	0.32	42.77	661.60 ± 893.36	1.24 ± 0.00	0.64 ± 0.04	0.39 ± 0.25	**0.03 ± 0.03**
	Fridge (II)	3.62	0.29	2.26	25.42 ± 57.86	0.09 ± 0.08	**0.07 ± 0.08**	0.11 ± 0.09	0.13 ± 0.06
	Light (II)	4.37	**0.04**	4.29	0.10 ± 0.04	0.04 ± 0.02	0.09 ± 0.04	0.09 ± 0.04	0.12 ± 0.04
	Microwave (II)	10.73	0.39	8.10	19.27 ± 56.14	0.27 ± 0.27	0.09 ± 0.03	**0.09 ± 0.07**	0.15 ± 0.09
	Sockets (II)	0.14	0.02	0.14	0.11 ± 0.28	0.00 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	**0.00 ± 0.00**
	Washer dryer (II)	4.89	0.67	2.62	0.33 ± 0.40	**0.29 ± 0.34**	0.51 ± 0.02	0.50 ± 0.04	0.36 ± 0.04

A.2 Normalized Disaggregation Error (NDE)

Table 3: NDE per appliance on the HIPE (industrial), IMDELD (industrial) and REDD (household) data of 10 runs each.

		CO	Mean	FHMM	WGRU	RNN	DAE	S2P	S2S
HIPE	Chip press (II)	62.61	0.99	39.75	0.50 ± 0.19	0.46 ± 0.02	0.55 ± 0.01	0.82 ± 0.14	**0.32 ± 0.07**
	Chip saw (III)	30.25	0.95	23.09	1.22 ± 1.06	2.02 ± 4.28	**0.64 ± 0.00**	0.69 ± 0.01	0.71 ± 0.13
	High temperature oven (III)	23.61	0.96	10.72	1.50 ± 1.65	0.86 ± 0.07	**0.79 ± 0.00**	0.84 ± 0.04	0.83 ± 0.03
	Pick and place unit (II)	7.26	0.76	7.33	7892.89 ± 24 220.49	0.70 ± 0.14	0.62 ± 0.00	**0.62 ± 0.01**	0.63 ± 0.01
	Screen printer (I)	7.08	0.88	6.92	53.53 ± 158.88	0.66 ± 0.13	**0.57 ± 0.00**	0.60 ± 0.02	0.60 ± 0.02
	Soldering oven (III)	49.81	0.98	122.16	1.96 ± 4.83	0.66 ± 0.12	0.69 ± 0.02	0.41 ± 0.02	**0.32 ± 0.01**
	Vacuum oven (III)	617.68	**1.00**	182.62	2.17 ± 3.43	1.01 ± 0.00	1.03 ± 0.01	1.01 ± 0.00	1.00 ± 0.00
	Vacuum pump (I)	5.63	0.84	5.22	825.46 ± 2608.66	0.48 ± 0.19	0.42 ± 0.00	0.48 ± 0.20	**0.40 ± 0.01**
	Washing machine (II)	475.27	1.00	471.53	0.64 ± 0.18	1.02 ± 0.02	1.37 ± 0.06	1.24 ± 0.39	**0.43 ± 0.07**
IMDELD	Double-pole Contactor I (I)	1.48	0.53	1.58	0.49 ± 0.27	0.22 ± 0.03	0.41 ± 0.03	0.32 ± 0.17	**0.20 ± 0.02**
	Double-pole Contactor II (I)	1.43	0.56	1.53	40.68 ± 120.82	0.37 ± 0.29	0.39 ± 0.06	0.27 ± 0.05	**0.23 ± 0.04**
	Exhaust Fan I (I)	0.87	0.48	1.01	0.40 ± 0.38	0.30 ± 0.27	0.25 ± 0.02	0.31 ± 0.04	**0.13 ± 0.01**
	Exhaust Fan II (I)	0.95	0.47	0.89	69.46 ± 218.31	0.22 ± 0.02	0.24 ± 0.03	0.26 ± 0.05	**0.13 ± 0.01**
	Milling Machine I (I)	0.54	0.64	0.76	0.33 ± 0.02	0.15 ± 0.01	**0.14 ± 0.00**	0.46 ± 0.00	0.17 ± 0.01
	Milling Machine II (I)	0.58	0.63	0.73	0.61 ± 0.26	0.13 ± 0.01	0.12 ± 0.00	**0.07 ± 0.00**	0.12 ± 0.00
	Pelletizer I (I)	**0.12**	0.49	0.13	0.64 ± 0.82	17.96 ± 56.22	0.18 ± 0.01	0.17 ± 0.05	0.13 ± 0.01
	Pelletizer II (I)	0.19	0.52	0.18	0.46 ± 0.30	0.29 ± 0.12	0.25 ± 0.01	**0.17 ± 0.02**	0.18 ± 0.02
REDD	Dish washer (II)	2.95	0.97	2.53	0.98 ± 2.44	0.44 ± 0.44	0.15 ± 0.04	**0.06 ± 0.01**	0.07 ± 0.02
	Electric oven (II)	1.77	1.00	1.85	0.58 ± 0.24	**0.08 ± 0.03**	0.81 ± 0.01	0.70 ± 0.03	0.78 ± 0.03
	Electric space heater (II)	30 747.76	0.98	15 367.16	42.19 ± 130.43	1.27 ± 0.00	0.57 ± 0.03	**0.39 ± 0.06**	0.45 ± 0.01
	Electric stove (II)	978.37	0.98	41.31	46 303.92 ± 95 099.02	1.23 ± 0.00	0.59 ± 0.01	**0.38 ± 0.06**	0.44 ± 0.01
	Fridge (II)	5.41	0.69	2.31	3323.79 ± 7723.61	0.51 ± 0.01	0.50 ± 0.01	0.52 ± 0.04	**0.48 ± 0.02**
	Light (II)	7.39	0.70	7.10	**0.67 ± 0.01**	0.73 ± 0.01	0.70 ± 0.02	0.71 ± 0.01	0.69 ± 0.01
	Microwave (II)	8.60	0.98	9.31	475.16 ± 1423.43	0.51 ± 0.18	0.69 ± 0.01	**0.31 ± 0.02**	0.43 ± 0.01
	Sockets (II)	0.03	0.01	0.03	0.09 ± 0.26	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	**0.01 ± 0.00**
	Washer dryer (II)	2.05	0.98	2.09	0.26 ± 0.39	0.32 ± 0.47	0.36 ± 0.02	0.27 ± 0.02	**0.19 ± 0.01**