NILM Applications: Literature review of learning approaches, recent developments and challenges

Georgios-Fotios Angelis[a,1,*], Christos Timplalexis[a,1], Stelios Krinidis[a,b], Dimosthenis Ioannidis[a], Dimitrios Tzovaras[a]

[a] *Information Technologies Institute / Centre for Research and Technology - Hellas, 6th km Charilaou-Thermi Rd, Thessaloniki, 57001, Thessaloniki, Greece*
[b] *International Hellenic University (IHU), School of Economics and Business Administration, Management Science and Technology Department , Kavala, Greece*

Abstract

This paper presents a critical approach to the non-intrusive load monitoring (NILM) problem, by thoroughly reviewing the experimental framework of both legacy and state-of-the-art studies. Some of the most widely used NILM datasets are presented and their characteristics, such as sampling rate and measurements availability are presented and correlated with the performance of NILM algorithms. Feature engineering approaches are analyzed, comparing the hand-made with the automatic feature extraction process, in terms of complexity and efficiency. The evolution of the learning approaches through time is presented, making an effort to assess the contribution of the latest state-of-the-art deep learning models to the problem. Performance evaluation methods and evaluation metrics are demonstrated and it is attempted to define the necessary requirements for the conduction of fair evaluation across different methods and datasets. NILM limitations are highlighted and future research directions are suggested.

Keywords: NILM, non-intrusive load monitoring, load disaggregation, review, machine learning, deep learning

1. Introduction

The "Net Zero by 2050" report issued by the International Energy Agency (IEA) declares that the energy sector is responsible for three quarters of total greenhouse gas emissions [1]. Moreover, the constant increase in global energy demand compared to power supply, introduces severe challenges for the efficiency and reliability of the traditional power grid systems. On the path towards zero CO_2 emissions by 2050, the transformation of the electrical grids plays a critical role. The increase in computational power combined with novel modeling and simulation capabilities gives the opportunity for a smooth transition from traditional grids to the smart grid era [2]. Rapid advancements in advanced metering equipment, internet of things (IoT) devices and artificial intelligence (AI) algorithms allow the management, monitoring and control of the distribution grid, optimizing energy utilization and thereby saving energy [3]. Smart grid is expected to integrate assets like Distributed Energy Resources (DERs), Electric Vehicles (EVs), Energy Storage Systems (ESSs) and utilize intelligent services aiming to unlock the flexibility potential that will allow the generation, distribution and consumption of energy in a more efficient way. Non-intrusive load monitoring (NILM) is a service that contributes towards this target by estimating the consumption of a building's individual appliances.

NILM, or else load disaggregation, aims to identify the operational state (on/off) and the precise power consumption of individual electrical loads, considering as input only the aggregated consumption of these loads. This concept was firstly introduced by Hart in 1992 [4], but it has been developed extensively over the last decade, due to the progress made in the field of machine learning and deep learning algorithms. Disaggregation algorithms are usually implemented in the residential sector, however there are also studies focused on industrial [5, 6, 7] or shipboard applications [8, 9]. As the term non-intrusive suggests, this method is implemented with minimum interruption of the users' privacy. Measurements are obtained only from a single point (aggregated load), so there is no need for deployment of extra equipment that would increase the complexity and the cost of the installation. The unique energy consumption of each electrical appliance is often called "load signature". Based on their load signatures the appliances can be grouped into the following categories:

- **Type-I:** ON/OFF state. Those appliances are considered to have only two states of operation. Lamps, toaster, boiler and resistive loads in general belong to this category.
- **Type-II:** Finite State Machines (FSM). The appliances belonging to this category have a multiple (finite) number of operating states. Washing machine and stove burner are some examples of type-II appliances. They usually have a repeating pattern of alterations over their operational states, which makes it easier to recognize their load signature.
- **Type-III:** Continuously Variable Devices (CVD). They do not have a fixed number of states, as their consumption

*Corresponding author
Email addresses: angelisg@iti.gr (Georgios-Fotios Angelis), ctimplalexis@iti.gr (Christos Timplalexis), krinidis@iti.gr (Stelios Krinidis), djoannid@iti.gr (Dimosthenis Ioannidis), dimitrios.tzovaras@iti.gr (Dimitrios Tzovaras)
[1] These authors contributed equally to this work.

varies constantly. Dimmer lights are an example of CVD appliances. Due to their way of operation, they are considered hard to disaggregate from the total consumption.

- **Type-IV:** According to [10, 11], a fourth appliance type is identified by the authors, namely permanent consumer devices. As their name indicates, they remain active at all times, consuming energy. TV receivers and smoke detectors are some typical examples of type-IV appliances.

Having access to appliance specific data, rather than whole house measurements, introduces a number of benefits both for consumers and the utility companies. Starting with the consumers, they are able to understand better their energy consumption, as they see which appliances are more energy intensive. Thus, they can make more informed decisions regarding their energy habits. The majority of the consumers are not aware of the energy they consume and do not realize their impact on the environment. Increased awareness can lead to a more rational use of their home appliances. More specifically, consumers may select to use less their top-consuming devices and in some cases they may also want to replace old appliances that are not energy efficient. According to [12, 13], customers consistently achieve energy savings when they receive appliance specific feedback regarding the energy usage of their appliances. Taking it a step further, feedback could be enriched with a set of recommendations that give the opportunity to the customer to achieve specific energy saving goals with measurable rewards. Recommendations could even perform remote diagnostics to a household's appliances, notifying the occupants for unusual usage patterns of the existing appliances. Utility companies could also benefit from NILM, by strengthening their link with customers, providing a better understanding of the usage and consumption of electric power. Discovering certain behaviors, could help utilities effectively implement market segmentation, namely identify clusters of customers having similar needs and demands. This way they can design more efficient marketing strategies, or even diversify their strategies on the basis of personalized services provision. The analysis of historical data, may reveal patterns regarding the time of usage for each appliance, giving the opportunity for personalized recommendations that fit each customer's needs. Understanding how customers are consuming power is essential for the utilities, since more accurate day ahead or short-term load forecasting can be achieved. As a result, the implementation of demand response (DR) strategies is facilitated. Consumers could be given some incentives in order to constrain or time shift the usage of some appliances, giving the opportunity to the grid operators to create a more precise matching between power supply and demand. The energy savings potential deriving from energy disaggregation, is expected to have a positive impact on the environment. Less global energy needs means that our dependence on fossil fuels is reduced, since the energy demand can be covered at a greater percentage from RES.

In the past recent years various publications have reviewed Energy Disaggregation. Each study, demonstrated a different aspect of the literature. For instance, Iqbal et al. [14] presented a critical review of the state-of-the-art residential and commercial datasets. Another study [15] presented an experimental overview of the NILM-API and performed comparisons and evaluation of the state-of-the-art approaches. Finally, the authors of [16] surveyed the application of deep neural networks in low-frequency data. Table 1 presents several impactful review publications for Non-Intrusive Load Monitoring:

Table 1: Energy Dissagregation review publications

Reference	Publication Date	Dataset analysis	Feature Extraction	Learning Approaches	Evaluation Methods
Zeifman and Roth [17]	2011	✗	✓	✓	✓
Klemenjak and Goldsborough [18]	2016	✓	✓	✓	✓
Hosseini et al. [19]	2017	✓	✗	✗	✗
Pereira and Nunes [20]	2018	✓	✗	✗	✓
Bonfigli and Squartini [21]	2020	✓	✓	✓	✓
Donato et al. [22]	2020	✗	✗	✗	✗
Gopinath et al. [3]	2020	✓	✓	✓	✓
Iqbal et al. [14]	2020	✓	✗	✗	✗
Salem et al. [23]	2020	✗	✓	✓	✓
Huber et al. [16]	2021	✗	✗	✓	✓

However, a detailed and up-to-date study that will present a comprehensive and complete overview of the current status and the research gaps is missing. Based on this, the main contribution of this work are:

- An analytical overview of the most widely used datasets. Different characteristics are presented along with appliances availability, sampling rate and measurements duration. In addition, apart from residential or commercial datasets that are usually reviewed in NILM literature we analyze also datasets with measurements collected from the industry domain.

- Comprehensive presentation of all the feature extraction and pre-processing techniques that have been applied in energy disaggregation domain. Approaches that have been presented in deep learning and machine learning studies are analyzed severally.

- Up-to-date detailed overview of the existing NILM approaches. In particular in section 4, all learning approaches are demonstrated, focusing mainly on the latest machine and deep learning methods. Analytical description for each publication and the corresponding methods are depicted in the Tables 3, 4, 5, while current limitations and further research directions are highlighted.

The rest of this paper is organised as follows. Section 2 reviews the characteristics of some of the most widely used NILM datasets and elaborates on data pre-processing techniques. In Section 3, a comparison is made between automated and handcrafted feature extraction methods. Section 4 analyzes the learning approaches for the disaggregation task, mainly focusing on the state-of-the-art deep learning methods. Section 5 investigates the NILM algorithms evaluation process. Current limitations and further work are discussed in section 6 and finally, conclusions are drawn in 7.

2. Data

2.1. NILM Datasets

In the past recent years, the rapid growth of data availability has contributed towards the development of many research areas. NILM was affected by this explosion, with an increasing number of datasets being published lately. Their development and analysis were of paramount importance, to draw meaningful performance comparisons of various NILM algorithms. In the following sections, we present the most widely used publicly available datasets. We analyze and compare the existing datasets from different aspects (sampling rate, measurements capturing period, attributes being registered, number of different houses and devices that each dataset contains, types of buildings that are presented - residential, commercial or industrial).

2.1.1. Residential Datasets

- **REDD:** Reference Energy Disaggregation Data Set (REDD) [24], was published by MIT in 2011, containing both high and low-frequency recordings at 15 kHz and 0.5 Hz, respectively. It includes recordings from six residential buildings in the United States with a total duration of 119 days. Also, ninety-two household appliance measurements were included with a sampling frequency at $1/3$ Hz.

- **UK-DALE:** UK Domestic Appliance-Level Electricity [25] comprises of power consumption data collected from 5 residential buildings in the United Kingdom from 2013 to 2015. More than 10 types of household appliances are included in the dataset. Aggregate consumption frequency ranges depending on the household (low frequency at 1 Hz - high frequency at 16 kHz). All appliances are sub-metered at $1/6$ Hz.

- **AMPds/2:** The Almanac of Minutely Power dataset (Version 2) [26] is an open dataset including 2 years of consumption data for a single household in Canada, sampled at 1 minute. The dataset contains a total of 21 power meters, 2 water meters (with additional appliance usage annotations), and 2 natural gas meters. Power meters describe each appliance by registering 11 electrical parameters (voltage, current, active/ reactive/ apparent power etc.). Billing information for cost analysis is also available.

- **REFIT** [27]: It was published in 2017 and it includes electricity data from 20 households in the United Kingdom. Unlike UK-DALE or REDD, REFIT readings are registered ceaselessly for the full two-year period. REFIT has a low sampling rate with a sampling period of 8 seconds for the mains and individual appliances active power. Also, noteworthy information about this dataset is that in three households (3,11 & 21) solar panel was installed, so most of the research studies on energy disaggregation exclude these houses from the final dataset.

- **Dataport** [28]: It was introduced in 2012, including power consumption recordings from 722 houses and commercial buildings across different cities in the United States. Every home measuring period was different, from 2011 to 2015, and the dataset provides total real and aggregate power consumption and sub-meter appliance level recordings. Dataport dataset is considered also, a low-frequency dataset, with a sampling period of 1 minute for aggregate and appliance signal.

- **ECO** [29]: It is comprised of electricity measurements gathered from 6 residential buildings for over 8 months in Switzerland. Active power, voltage, and current are registered at a low sampling rate of 1 Hz. Each building includes different appliances, data granularity at a different level, and also the dissimilar duration deployment.

- **ENERTALK**

  [30]: It includes data collected from 22 houses in Korea, for a period from 29 to 122 days for each site. Overall, the dataset records readings for 5 home appliances. This dataset also provides active and reactive power measurements for aggregate signal and individual power consumption for each device, with a sampling frequency of 15 Hz. A small amount of readings are not properly registered due to logs and meters issues, so pre-processing techniques for missing values are suggested by the authors.

- **iAWE** The first dataset that included data from households in India published in 2013. The Indian Dataset for Ambient Water and Energy (iAWE) [31] contains ambient, water, and electricity data from a residential building in New Delhi. The dataset's total duration is approximately 73 days and measurements are collected from thirty-three sensors daily. The recordings provide also information about active, reactive, apparent power, voltage, and current with a sampling period from 1 to 6 seconds representing over 63 electrical appliances.

- **BLUED** The Building-level fully labeled dataset for electricity disaggregation (BLUED) [32] is a publicly available dataset utilized for the NILM task. The dataset was released in 2012, and contains one-week electricity data from a domestic site in the United States, with a current and voltage high-sampling frequency of 12 kHz. The aggregated active power measurements are at 60 Hz including more than 50 appliances. Finally, this dataset also captures the state transition of each appliance, labeled with all the necessary time stamps.

- **PLAID** The Plug-Level Appliance Identification Dataset (PLAID) [33] is a load identification dataset that contains current and voltage measurements from 56 domestic sites in the United States. Overall, it contains 1094 observations with a sampling frequency at 30 kHz and 11 power consumption from 11 different household devices. These appliances are: Air Conditioner(AC), Compact Fluorescent, Lamp(CFL), Fridge, Hairdryer, Laptop, Microwave,

Washing Machine, Bulb, Vacuum, Fan, Heater. In addition to PLAID's original version, PLAID II [34] was released to handle the restrictions of the first version. Another 719 observations were added to avoid biasing. Finally, the latest version PLAID III [35] was published, and increased the number of appliances by two. PLAID is the only dataset that excludes active and reactive power recordings.

- **DRED** [36] is a residential dataset that was released in 2015 and contains energy consumption data from a household in Netherlands. The Dutch Residential Energy Dataset includes electricity measurements for the aggregate and submetered signal with a total duration over six months. Regarding the appliance level recordings, twelve devices are measured with a sampling rate of 1 Hz. Moreover household metadata information are provided, number of inhabitants, house layout, mapping between appliance and location.

- **SynD** [37] is a synthetic dataset, constructed for energy disaggregation task, in residential buildings. The creation of real datasets requires long-term capturing periods and a great number of resources spent on recording equipment. Also real datasets, often are affected by equipment malfunctions that produce corrupted measurements. To overcome these boundaries, synthetic datasets were employed. This dataset, simulated electric energy consumption readings for one home for 180 days. The data include aggregate and twenty-one appliance measurements with a sampling frequency at 5 Hz. In this dataset, the measurements campaign was based on the monitoring of twenty-one different home devices from two residential sites in Austria. During the campaign, consumption patterns were observed, and then operation cycles were extracted. After that, the simulation process began where the SynD dataset was generated.

- **Georges Hebrail UCI** [38] is a dataset containing individual household electric power measurements. The sampling rate is at one minute both for the aggregated consumption and the sub-metered appliances, while the whole dataset contains a period of four years. The electrical quantities of active power, reactive power, voltage and current are registered. The house includes three sub-metering points, where each point corresponds to a room and as a result contains more than one sub-metered appliances.

2.1.2. Commercial buildings Datasets

- **COMBED** [39], released in 2014, is the first non-residential dataset in NILM literature. Nowadays, COMBED is one of the few datasets that provides recordings from commercial buildings. Power consumption data were gathered from IIT Delhi, an educational campus, which is constituted of 8 institutional buildings. Smart meters are employed throughout the campus and across the different buildings to collect active power and current measurements. The total recordings acquisition duration is one month and the sampling frequency is 30 seconds.

2.1.3. Industrial Datasets

- **Industrial Machines Dataset for Electrical Load Disaggregation**

  [6] is a public dataset that includes measurements from an industrial site for the energy disaggregation task. It contains power consumption recordings from a poultry feed factory in Brasil for a working period of 111 days. The measurements are collected during the factory's operational hours, Mondays through Fridays, 10:00 PM to 05:00 PM. The aggregate and the appliances power consumption are gathered through eleven meters, with a sampling frequency of 1 Hz. One meter measures the total factory power signal and there are other two collecting power consumptions from the pelletizing and milling processes, correspondingly. The other eight meters record the factory appliances: pelletizer I (PI), pelletizer II (PII), double-pole contactor I (DPCI), double-pole contactor II (DPCII), exhaust fan I (EFI), exhaust fan II (EFII), milling machine I (MI), and milling machine II (MII). Additionally, every meter sends data about RMS voltage, RMS current, active power, reactive power, apparent power, and active energy. All machines are measured through the whole working period except the two milling machines, which were measured only the last 12 days.

- **Aachen Smart Factory** dataset includes electricity recordings from a smart factory and four of its machines at FINESCE trial site Aachen/ Cologne. The data were gathered in the context of EU FINESCE project. It contains active power measurements for a total duration of 68 days.

- **High-resolution Industrial Production Energy (HIPE):**

  Another public dataset from an industrial building is HIPE [40]. Data are collected for a period of three months. Ten smart meters were deployed, measuring ten machines recordings. Also, the total power consumption is derived from the main terminal. The measurements contain values from different electrical quantities e.g. active, reactive, apparent power, voltage, current with a resolution of 5$seconds$. The dataset was released officially in 2018 and it included the following electronic appliances: PickAndPlaceUnit (1P), SolderingOven (3P), WashingMachine (3P), ScreenPrinter (1P), VacuumPump1 (3P), VacuumPump2 (1P), HighTemperatureOven (3P), VacuumOven (3P), ChipSaw (3P), ChipPress (3P). Finally, due to the small number of publications for NILM in an industrial setting, there are not many works that performed disaggregation in the HIPE dataset. However, in comparison with all the related published datasets, it seems more complete and precise.

2.2. Data Harmonization

- **Data Filling:** The problem of missing values in a time-series dataset occurs on several occasions when data are

Table 2: Main characteristics of all the aforementioned datasets in this section

Name	Release Date	Type	Building(s)	Total Appliances	Period	Characteristics	Aggregate Sampling	Device Sampling
REDD	2011	R	6	92	119 days	P,V,I	15 kHz & 0.5 kHz	1/3 Hz
BLUED	2012	R	1	50	1 week	P,Q,V,I	12 kHz	60 Hz
G. Hebrail UCI	2012	R	1	9	4 years	P,Q,V,I	1min	1min
Dataport	2012	R & C	722	8598	4 years	P, S	1 min	1 min
iAWE	2013	R	1	63	73	P,Q,S,V,I	1 – 6 sec	1 – 6 sec
AMPds	2013	R	1	20	360 days	P, Q, S, V, I	1 min	1 min
PLAID	2014	R	65	1876	1-20 sec	V,I,	30 kHz	30 kHz
COMBED	2014	C	8	-	1 month	P,I,	30 sec	30 sec
UK-DALE	2015	R	5	109	2247 days	P,Q,S,V,I	16 kHz & 1/6 kHz	1 Hz
DRED	2015	R	1	12	6 months	P	1 Hz	1 Hz
ASF	2015	I	1	4	68 days	P	1 Hz	1 Hz
ECO	2016	R	6	45	8 months	P, V, I	1 Hz	1 Hz
AMPds2	2016	R	1	20	720 days	P,Q,S,V,I	1 min	1 min
REFIT	2017	R	20	177	2 years	P	8 sec	8 sec
IMD	2018	I	1	8	111 days	P,Q,S,V,I	1 Hz	1 Hz
HIPE	2018	I	1	10	92 days	P,Q,S,V,I	1/5 kHz	1/5 kHz
ENERTALK	2019	R	22	75	1714	P, Q	15 Hz	15 Hz
SynD	2020	R	1	21	180 days	P	5 Hz	5 Hz

gathered from smart meters or sensors. These discontinuities can be a great limitation in handling, analyzing, and extracting important features from the data, which can lead to a biased dataset. To overcome these issues, a variety of methods is used that is aiming at data enhancement. Though the selection of the appropriate data filling technique is an open discussion and it lies heavily on the data characteristics. The method that is preferred with greater frequency is interpolation. Three different interpolation techniques are utilized for missing values calculations, linear, quadratic, and cubic and they can be formulated as first, second, and third-order polynomial equations, respectively [41]. According to, [42] the most appropriate approaches for electricity consumption data are linear and spline interpolation. However, there is not an universal strategy about handing missing values in NILM literature.

- **Data Resampling:** Another crucial issue, in data preprocessing is the sampling frequency. In section 2.1 we have presented the original sampling frequencies of all the aforementioned datasets. However in many cases, missing data or noise in the aggregate and the sub-metered signals make data resampling a necessary procedure. According to [43], the frequency resolution is one of the most important aspects in NILM task, because of the non-identical features and characteristics that different ranges of frequency contain. Recent studies [44, 45], performed comparable experiments in several frequency ranges, attempting to define the most suitable sampling frequencies that deliver decent disaggregation accuracy. In [44],the authors executed both classification and regression experiments and concluded that a sampling rate of at least $1Hz$ and $3Hz$ respectively, is required. However, a more detailed analysis presented in [45] suggests that the favourable sampling rate falls within the range of $1Hz$ to $1/30Hz$. Interestingly, in specific datasets there are certain appliances (dishwasher, fridge and washing machine) where the predictive accuracy is improved as the sampling rate is decreased.

2.3. Data Augmentation

A large amount of data is essential for training large and complex neural network architectures. However, in most real-world settings, predictive algorithms address insufficient training sets or datasets with inbalanced classes. In those situations, deep learning models tend to overfit, and consequently, they lack predictive accuracy. As a solution, a method that is called data augmentation is employed that aims at generating synthetic data by applying transformations to the original set [46]. Especially in many fields, such as computer vision (CV), Natural Language Processing (NLP) and speech processing, data augmentation is already an established methodology and has been proved very effective at improving neural nets robustness [47]. Though, data augmentation in time series data has drawn limited attention. Due to time series data complex properties and nature, applying transformations often causes distortions and loss of valuable information. Therefore, classical data augmentation methodologies, that emerged in other domains, such as flipping, zooming, rotation are not well-suited for time series and are resulting in poor synthetic data. Also opposite to computer vision, in time series problems data augmentation strategies are task dependant.

Considering these limitations, a set of different techniques for data augmentation in time series, have been adopted. Most of the existing methodologies manipulate the time series data directly in the time domain. Such transformations are, window-warping [48], cropping [49], noise injection, interpolation, and magnitude-warping. One other approach is data augmentation in the frequency domain. The aim of this method is to map time-series data in frequency and utilize special features. Frequency spectrum is calculated through Discrete Fourier Transform [50] or Discrete Wavelet Transform [51] and perturbations in amplitude and phase spectra are adopted to augment data. Finally, generative models have been used as an efficient data augmentation technique. Several recent works [52, 53, 54], applied Generative Adversarial Networks to create natural-looking synthetic time-series data.

Data collection and annotation is a time-consuming procedure, especially for NILM. Also, often, real-world NILM

datasets contain small amounts of data and suffer from missing measurements. Recently, various studies explore data augmentation in NILM to overcome these boundaries, aiming to improve predictive algorithms generalization ability. The first attempt to generate synthetic data for NILM was made in [55] by extracting appliance activations with random shifts and then arbitrarily assign them in the input vector. The same procedure also adopted in [56, 57, 58], even though this method doesn't imitate precisely the structure and shape of the real aggregate data. Another data augmentation approach is presented in [59, 60], and the intuition behind this method is the summation of random sub-metered data windows to create the synthetic aggregate signals. At last, two recent studies investigated data augmentation techniques for low and high sampling frequency data, respectively. According to [61], the authors proposed four different interpolation techniques to generate augmented high sampling frequency data, from low-frequency datasets, while Delfosse in [62] proposed a data augmentation approach, that performed simulations with appliances signatures to create synthetic data.

3. Feature Engineering

The feature extraction process involves the techniques implemented on the raw input signal and/or other external sources, aiming to construct feature vectors that are able to describe the characteristics of the predicted variable, in our case, the appliances' consumption. Feature extraction was initially conducted manually based on the available expertise on the specific field of study. The development of deep learning has recently unlocked the potential for the automation of the feature extraction process. Automated feature extraction is mainly used with image data due to its effectiveness, however it is common to utilize this technique with other modalities such as text, sound or tabular data.

3.1. Data Preprocessing

As mentioned in Section 2 data collected from smart meters or sensors oftenly need enhancement. Moreover, in order to processed by predictive algorithms another import step is data pre-processing, a process that aims to transform raw data into the appropriate format for training and evaluation. In this subsection core pre-processing steps are going to be analyzed, along with several different strategies that have employed in NILM literature through the years.

3.1.1. Input Configuration

Deep learning became the last few years the prevailing way to address NILM. The neural networks' ability to learn representations and extract complex features from raw data allowed solutions to be designed with the minimum-required handcrafted feature extractions. However, transforming signal data into neural network's input is a crucial matter, that has been paid a lot of attention. The most commonly used technique is the sliding window approach, which splits historical data into overlapping smaller sequences, with a fixed length. It has to be mentioned, that LSTMs have the ability to control the information required by extending or erasing through gating mechanism. This functionality enables the network to retain key information during the training process. However, sequences that contain a large amount of information can increase the computational complexity and also smaller sequences can provide less information to the network. Hence, it necessary the optimal selection of the input lags as it impacts LSTMs [63, 64, 65, 66, 67].

Therefore, one of the most important aspect in time series regression tasks, such as NILM, is the optimal length of the receptive field. It is important to outline that there is not a universal methodology in NILM literature, for the selection of the ideal input size. Most of the studies published have selected experimentally the sliding window's length, considering various characteristics, such as appliance type, individual appliances operational duration, sampling frequency. Moreover, the bigger the receptive field is, the bigger the operational cycles it captures. But in many cases, very large sizes lead to computational complex algorithms and poor disaggregation performance. Of course, another important factor in determining the optimal sequence length is the sampling rate of the input signal. According to [68], the sliding window size can be calculated by a formula that includes, size of the algorithm's input, original data sampling rate, and re-sampling factor. The experimental results showed that their method applied on state-of-the-art models, achieved superior performance, in comparison with the previous work. Finally regarding the input configuration, a different approach was employed with the differential of raw signal as an input. The authors of [69] designed this technique, considering that a neural network tries to disaggregate one appliance per training and sees the other appliances as a noise. Based on this, they proposed to differentiate the raw aggregate signal to make the output signal more distinguishable and improve disaggregation performance.

3.1.2. Feature normalization

A significant procedure that has a great impact on neural networks training routine, is feature scaling or data normalization. Especially in time series, raw data represent a wide range of values that can lead a DNN model in convergence failure. Scaling features, can prevent gradient vanishing or explosion and speed up the learning process [70]. While in time series regression and classification tasks there are several feature scaling techniques employed, in NILM researchers adopted mainly two of them. The first one, is z-score normalization, where mean and standard deviation are calculated at the complete training set both for main and sub-metered signal. Z-score normalization is denoted by the following equation:

$$x_{z-norm} = \frac{X_t - \overline{X}}{\sigma}, \tag{1}$$

where X_t is the aggregate or extracted power consumption at time t, $\overline{X}$ is the mean value of a main or appliance reading and σ is the standard deviation. The second scaling technique that is investigated in NILM literature, is min-max normalization. In this approach, data are scaled to a range of $[0, 1]$ or $[-1, 1]$. Minimum and maximum values of the whole training set are calculated both for mains and per appliance signals and then

the min-max normalization is calculated from the following formula.

$$x_{min-max} = \frac{X - X_{min}}{X_{max} - X_{min}}, \quad (2)$$

where X is the dataset's readings and X_{max} and X_{min}, are the maximum and minimum power consumption values for each signal, correspondingly. Even though, it is a common-used technique in time series data, it is less preferred in the NILM literature because of its sensitivity to outliers. Finally, there is a small amount of studies implementing neural network architectures, applying feature scaling as a pre-processing step. In contrast, normalization layers are adopted in [71, 72, 73] in an effort to prevent convergence issues.

3.1.3. 2D Representations

The recent advances on the Computer Vision field, in combination with the neural networks' ability to extract accurate features and the novel image encoding techniques, enabled researchers to adopt CV high-performance methods in other tasks. In time series data, in order to transform them as images, an encoding technique is required that creates 2-dimensional representations. Several studies [74, 75, 76, 77], explored time series to images encoding techniques and three of them are mainly adopted.

- **Gramian Angular Summation / Difference Fields (GASF / GADF):** In Gramian Angular Summation/Difference Fields time series data are represented in Polar coordinate system, which means that each data point is the the cosine of the summation of angles. In order to apply GAF methods, raw data scaling is needed using min-max normalization. Transformation to polar coordinates is processed with the initial value as the angular cosine, and the timestamp as the radius. Then the Grammar Angular Field is constructed, selecting one of the two main approaches, GASF or GADF where their main difference is that the final angular perspective is extracted by calculating the summation or difference respectively for each timestamp.
- **Markov Transition Fields (MTF):** MTF is a visualization method for time series data, that represents the transition probabilities of the input sequence. The final output is a $Q \times Q$ Markov transition matrix, which is calculated by the weighted assignments of x_i to the corresponding q_j quantile bin. The x_i is an element of the input sequence X and the q_j is an element of the extracted quantile bin Q.
- **Recurrence Plot:** Is the visual that is obtained from a $N \times N$ array of dots when a x_j is sufficiently close to a x_i [78]. The distance between the two data points is calculated through euclidean distance. Finally a threshold is applied and binary representation is constructed.

In NILM literature the first study that tried to encode signal data as an image, was [79]. The authors utilized the current-voltage trajectories to create binary images. These visualizations were used for event detection. Another recent study [80], proposed two novel load signatures to be utilized with the V-I trajectories to create image representations for load classification. In [81] the GADF encoding technique is investigated, in two residential datasets, for state detection. It was the first study, that utilized an encoding technique in active power consumption data. In [82], on the other hand, Recurrence Plot is applied on the aggregate signal and comparisons are performed with previously applied techniques. Finally, the authors of [83] employed signal to image transformations for load disaggregation. They compared all the three techniques mentioned above in three datasets. The GAF technique outperformed the other two approaches.

3.2. Handcrafted feature extraction

According to [84], load signature is the unique pattern of an individual appliance, when it is on an operational state. Each electrical appliance has certain distinct characteristics that determine its electrical behavior and are defined by variables such as voltage, current, active and reactive power. The handcrafted feature extraction process aims at extracting the maximum amount of information from the aggregated metering point (main panel), in a feature space that makes each appliance's signature discrete. Distilling information from handcrafted features requires a level of expertise on the scientific domain in which the problem belongs. More specifically in NILM, the comprehension of the electrical properties of each appliance (resistive, inductive, capacitive loads) is of utmost importance, for the extraction of discriminative characteristics.

3.2.1. Steady-state

Steady state features are extracted when the appliances operate on a steady state. Smart meter data at a low sampling rate usually lead to the exploitation of steady state features. Active power and its temporal variations is the most commonly used electrical measurement in NILM studies. Purely resistive loads, without any capacitive or inductive elements can be disaggregated by using active power as a single feature. However, most of the residential and industrial loads operate with a phase shift between current and voltage waveforms which means that they either generate or consume reactive power. Consequently, the inclusion of reactive power as a feature has been found to increase the models' predictive accuracy [85, 56]. The distribution of residential loads in the P-Q plane creates overlapping among appliances that operate on the same active power levels [86]. Several studies have also included information deriving from the V-I trajectories such as area, slope, curvature, asymmetry, over-shoot and trajectory centroid [87]. The usage of statistical features like mean, min/max, variance, skewness, kurtosis, quantiles and zero-crossings are experimentally found to increase disaggregation accuracy, while optimal results are obtained when combining electrical and statistical features [88]. All of the aforementioned features are extracted on the time domain, but there also studies that utilize the power signal transformation into the frequency domain for steady state features extraction. For example, Fourier analysis for the detection of steady-state current harmonics is used in [84, 89].

3.2.2. Transient

Transient state features are extracted when data are captured with a high sampling rate (kHz range) that allows the detection of the appliances' transient signatures during their operation. Transient features may help distinguish two or more appliances when they operate simultaneously, since the transient state reveals unique information for each appliance and transient signatures are less overlapping between appliances compared to steady state. Capturing data at this rate may introduce some problems since not all hardware equipment is able to process signals at high sampling rates. Moreover, data storage problems and computational issues arise from the processing of such large volume of data for the models' training. Some of the most typical transients features found in most studies are: transient power, transient V-I trajectories, start-up current, transient voltage noise [90]. A common method for analyzing the signal in those high frequencies, is the conversion to the frequency or the wavelet domain. Short-time Fourier Transform (STFT) and fast Fourier Transform (FFT) are widely used for the extraction of current harmonics that can help differentiate transient load signatures [91, 92]. Wavelet Transform (WT) offers a more flexible approach since it enables the representation of the signal both on the frequency and the time domain, allowing access to localized information about the signal [93, 94]. An early study analyzing the STFT and WT approaches, concluded that the formation of feature vectors making use of the Continuous Wavelet Transform (CWT) resulted into optimal results with less computational requirements compared to the STFT method [95].

3.2.3. External

Steady state and transient features usually rely on electrical measurements, however there are also some other external factors that can be modelled and add useful information to the predictive models. The seasonal patterns of usage for each appliance is a common example of those external features. Start time, end time, peak time, time-of-day, day-of-week and day-of-year information is important, since the frequency of usage of certain appliances may vary depending on the profile of a household's residents [96, 97, 98]. For example, home appliances, may operate more often during the evening or during the night, when a household's residents follow a 9-to-5 working schedule during weekdays. An encoding process is required for the appropriate modeling of those temporal features associated to periodic appliance usage. One-hot-encoding is a choice that treats seasonality as a categorical variable and most importantly, it ignores the cyclicality of the temporal features. A more accurate encoding is suggested in [99], where temporal features are considered as cyclically repeated variables being transformed into sines and cosines representations and finally mapped onto a circle to preserve the right relationships between them. Environmental factors may also affect the usage pattern of some loads, and more specifically HVAC loads which tend to operate only when they need to adjust (either hot or cold) outside temperatures [100]. Occupancy information (whether or not the house is occupied by users) has been also used as additional feature that can increase the disaggregation accuracy and moreover reduce the algorithm's computational cost during non-occupancy periods [101, 102].

4. Learning Approaches

NILM studies have tested both supervised and unsupervised approaches, depending on the available information and the implemented predictive algorithm. Supervised methods require a labelled dataset with sub-metered appliances, which may not always be available. Unsupervised methods can be implemented without any prior knowledge of the environment, but the user needs to verify the appliance patterns that are being recognized. HMM-based, optimization and machine learning approaches were predominantly used a decade ago. However, the development of deep learning solutions introduced neural network approaches in the NILM context and it soon became the basis for state-of-the-art implementations, outperforming previous studies. In this section, both classical machine learning and deep learning approaches will be discussed. Emphasis will be given on the latter category, since it is increasingly gaining interest in the field of NILM studies over the recent years.

4.1. Hidden Markov Models

Early NILM studies were mainly based on Hidden Markov Models (HMMs), which are typically used for probabilistic modelling of time series data [103, 104]. A number of states (consumption levels) is typically defined for each appliance, with each state having its own probabilistic distribution. Different variants have been proposed, with the most popular being Factorial HMMs that generalize the HMM state representation by letting the state be represented by a collection of state variables [105]. FHMM implementation for the NILM task are initially presented in [24, 103]. More specifically in [103] the incorporation of additional features to the HMM models, providing information regarding the appliance usage profiles is attempted. The formulated conditional factorial hidden semi-Markov model outperforms typical FHMMs and the authors consider this method a promising fully unsupervised approach for energy disaggregation. Additive Factorial HMMs, which were the basis for AFAMAP algorithm introduce a convex formulation of approximate inference that overcomes the problems of computational efficiency and tackles local optima issues [106]. This algorithm was considered to be a state-of-the-art HMM-based approach, however there are now several studies outperforming it [85, 107]. HMM-based techniques are usually inefficient when the number of disaggregated appliances increases and moreover they suffer from high computational complexity. A low complexity unsupervised solution, inspired by a fuzzy clustering algorithm, called entropy index constraints competitive agglomeration, is presented in [108]. The results indicate that the proposed algorithm enables the generalization of the features and produces a set that can be considered for model learning. A cloud-based solution performing online energy disaggregation using HMMs is presented in [109]. The algorithm consists of an event detection and an appliance modelling part, while it has the advantage that it is a training-less method, based on unsupervised learning.

4.2. Optimization Methods

Optimization methods provide a different approach where the main idea lies in finding the optimal combination of individual appliances that compose the aggregate signal. As stated in [110], the increased complexity from a large number of appliances and the loss of temporal continuity are two major drawbacks in making use of optimization techniques for NILM. Taking a more detailed look at the optimization methods that are being implemented, Genetic algorithms (GAs) seem to have promising performance [111, 112], while a Graphical Signal Processing (GSP) method, proposed in [113], offers a competitive solution with reduced computational complexity. Sparse optimization methods, mainly inspired from image processing, have also gained attention, achieving results that outperform FHMMs [114, 115]. The study presented in [116] treats disaggregation as a single-channel source separation problem. Non-negative matrix factorization is implemented, including additional information from the appliance dependencies. Comparison with other sparse coding approaches found in the literature demonstrated the superiority of the proposed method. The authors of [117] attempted a linear blind source separation strategy, creating clusters of steady-state changes and then employing a matching pursuit algorithm, trying to reconstruct the original power signals using the clusters that were found as the sources. The algorithm had decent performance on energy-intensive appliances, while low-consuming ones were not correctly discriminated due to multiple errors taking place during the clustering process, probably due to poor feature selection. A two stage process is described in [118], where chi-squared GOF and cepstrum smoothing are used for event detection, while the parameters of those two methods are optimized using surrogate-based optimization. Results indicate that both methods clearly outperform standard chi-squared and moreover parameter optimization is executed very fast compared to the brute force approach.

4.3. Shallow learning

The term shallow learning includes all of the traditional classification and regression algorithms (non-deep learning), and implies that a handcrafted feature extraction process was followed, utilizing domain expert knowledge. Those solutions are easier to implement, they have lower computational complexity compared to deep learning approaches and in certain cases they have yielded encouraging results.

SVM implementations have attempted to deal with the problem both with linear [119] and non-linear [120] approaches. Naive Bayes classifiers, assuming independence between each of the appliance's state are implemented in [121, 122], obtaining decent accuracy with only a small number of training samples. Variations of the K nearest neighbors (K-nn) algorithm are performed in [123], where the optimal setup is investigated. Tree-based implementations seem to be the best solution in shallow learning approaches, as they are often used until today, challenging the dominance of deep learning methods [124, 125, 126]. In [88] the importance of manually extracted features is initially evaluated. It is concluded that electrical features are more discriminative than the statistical ones. Several regression algorithms are benchmarked, with Random Forests outperforming k-NN, SVM and deep neural networks. The efficiency of Random Forest algorithm in NILM is also realized in [126], where the authors formulate the problem as a multi-label classification task. The suggested approach is more efficient both in terms of accuracy and training time, compared to other multi-label classification NILM approaches found in the literature. Ensemble tree methods have also shown great success, composing a weighted combination of multiple regression trees (weak learners), aiming to form a stronger learner. The work presented in [91] uses a decision tree ensemble based on the bagging technique. A novel set of frequency domain features is utilized and results indicate that appliances with similar consumption are successfully disaggregated. The boosting ensemble technique is implemented in [127], where XGBoost algorithm is used for energy disaggregation. XGBoost outperforms bayesian network, SVM, random forest, neural network and HMM, while it reduces the training overhead of the model.

4.4. Feed-forward Neural Networks

Feed-forward neural networks (FFNN) are the first and the simplest deep learning models. They consist of numerous nodes or units that are arranged in layers. Each layer is connected to the neurons of the next one. All the connections have a different weight, that defines each neuron's importance in the network. The overall structure of a neural network includes the input nodes, that always compose the first layer. There, the raw data are imported to the network, after they are fed in the hidden layers that are responsible for transforming the input data into feature vectors, and then to the output layer which is the last part of an FFNN topology and is responsible to produce the necessary values.

Regarding the adaptation of FFNN in NILM literature, this study [128] proposed a FFNN that aimed at extracting three appliances' power consumption signals. In [129], the authors employed several MLPs with different combinations of hyperparameters for load identification. The final results showed that each device performed better with a combination of different hyperparameters.

4.5. Convolutional Neural Networks

Convolutional Neural Networks are a popular deep learning model that has achieved state-of-the-art performance in many different tasks [130, 131, 132, 133]. A convolutional layer is the major building block of a CNN architecture where the convolution operation is performed. The CNN architectures except for the convolutions, also include several other types of layers, such as pooling layers, normalization layers, activation function layers, Flatten and Fully Connected layers. Also most times, convolutional layers are in combination with other types of DNN layers, such as RNNs, Attention Mechanism.

In terms of the energy disaggregation task, a variety of CNN architectures have been employed both for regression and classification. The study presented in [134], had a great impact on the NILM literature. The authors proposed a new CNN model that was trained to predict the midpoint of the output

signal instead of the whole sequence. They implemented the same architecture for both output strategies and compared the experimental results on two residential datasets. Their new approach performed the best results and shaped a new way to design DNN models for energy disaggregation. The work in [135] is also inspired by the aforementioned sequence to point (s2p) model. The authors applied four pruning algorithms, to shrink the number of weights and reduce the total number of parameters without affecting performance. Moreover, in [7] both sequence to sequence (s2s) and sequence to point (s2p) models are applied in two industrial datasets that have been presented in section 2.1. It was the first study that tried to apply classical NILM approaches from a residential building to an industrial.

As mentioned in section 3.1 the work in [69] designed a CNN architecture that had two inputs, 1D differential input and 1D auxiliary input. The proposed network contained five 1D convolutional layers with 1D max-pooling layers followed, in order to reduce the number of parameters and to prevent overfitting. The differential signal is given as an input and it is then processed from the part of the network described above. The auxiliary input is concatenated with the output of the final max-pooling layer and the resulting representation is fed into the fully connected layers that are responsible to produce the final output sequence.

Besides the works that used multiple inputs for training, these studies [136, 137] adopted two sub-networks that performed regression and classification simultaneously. In the first approach, the authors were inspired by multi-task learning where a neural network learns with multiple tasks, with different loss functions. The parameter sharing is utilized to improve the model's accuracy. The final representation was constructed by multiplying the regression with the classification output. Also, the loss function has been adjusted representing the overall loss from the two sub-networks. In the second study, the authors proposed SCANet, a multi-branch model with two sub-networks and multiple receptive fields that are connected. The model's structure is based on several dilated 1D convolutional layers that are placed in parallel, with different dilation rates. The convolutional branches in each sub-network are connected between them at different scales with a simple gating mechanism. Also, they claimed that performance improvement can be achieved by adding an adversarial loss in the loss function proposed in SGN. The overall results showed that SCANet performed better than SGN overall.

Moreover, two CNN developments that have applied to the NILM literature are dilated convolutions and CNNs with residual blocks. The authors of [138] created a convolutional sequence to sequence network where they used residual blocks to distill the final output. Another study that implemented CNNs with residual connections is presented in [139]. The authors constructed a CNN architecture with several 1D convolutional layers residually connected and with skip connections. Each 1D convolutional layer is followed by Batch Normalization and ReLU as activation function. The final results are obtained in two ways, with classical sequence to sequence and with sequence to short sequence, wherein at the second approach, the output is half of the input signal. Another method proposed a bidirectional dilated ResNet to avoid vanishing gradients [140]. Firstly, the input signal is fed into an 1D convolutional layer to detect low-level features and then to eight residual blocks with dilated convolutional layers. The output of each residual block is connected with the next stack, but it also skips all the subsequent parts and is concatenated before the final linear layer. A sequence to point strategy is adopted, representing the midpoint value of the sub-metered signal. The modification of WaveNet for NILM is presented in [141]. WaveNILM structure contains 1D dilated causal convolutions and their output is fed both into a gating mechanism and a rectified linear activation. The output of the two activations is multiplied and represents the output of the block. WaveNet for NILM has also been implemented for industrial energy disaggregation [6], outperforming FHMM. It was the first study that applied a deep learning method for NILM in an industrial site. Another approach was presented in Kaselimi et al. [142], where a multi-channel hybrid architecture that incorporated RNN properties into a CNN architecture, using a consecutive deep learning model, for enhancing appliance signal estimation.

Furthermore, numerous publications utilized 2D convolutions as a solution on NILM. The procedure of transforming power consumption signals to images is presented briefly in section 3.1.3. This work [143] implemented a 2D CNN architecture for load disaggregation in three residential devices. After that, both studies [81, 144] utilized as a backbone, the VGG-16 architecture. In the first one, the input is fed into the baseline that is used as a standalone feature extractor. Then the output feature vectors have been imported along with appliance labels into a classifier. The machine learning classifier detects on/off events. In the second study, the input is a 320×320 gray-scale image and this method has been designed for event detection. Additionally, in these publications [145, 146, 147, 148] CNN architectures were employed for multi-appliance classification. Especially, in these [146, 147] the authors designed an 1D CNN network, with structure similar to VGG-16, for appliance classification. Moreover, in [146] the authors had come to several conclusions regarding the performance of deep convolutional networks in on/off classification. Specifically, they claimed that power signal contains multiple low level features. Thus, they designed with a considerable depth to capture multiple feature characteristics but also not deep enough in order to add further complexity. Finally, Kong et al. [149] proposed an 1D VGG-16 for load disaggregation. Apart from this, the authors also presented a post processing CNN model, for type II appliances, that classified whether the predicted sequence belonged to the target device or not.

4.6. Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a deep learning architecture that is suitable for sequential data. Their intuition is that information from the previous states is also used as input to the current state. However, RNNs have certain limitations, facing gradient vanishing & explosion problems. To address this, variations of RNNs have been employed, Gated Recurrent Units (GRU) and Long Short Term Memory (LSTM) networks. In NILM literature, various approaches have been

published the last few years using recurrent models. In Kelly and Knottenbelt [55] an RNN architecture containing one convolutional layer and two bidirectional LSTMs after the fully connected output layers is implemented. In [150] a network with two LSTM layers is applied. The studies analyzed in [59, 151] constructed RNN-based networks to perform comparisons. In the first one, a model with one convolutional and two bidirectional GRU layers was created. The second study utilized two parallel 1D convolutional and two bidirectional LSTMs before the two final fully connected layers. The first work that explored state detection with RNN-based architectures was [152]. A GRU-based approach aiming at improving disaggregation accuracy and simultaneously reducing memory usage and computational complexity is presented in [153]. The authors claim that GRU networks are a more appropriate method for NILM compared to LSTMs, due to their computational efficiency. Moreover, these studies [154, 155] designed two LSTM networks for appliance classification. In the first one the model included an 1D convolutional process before the input neurons that fed the input signal into the stacked LSTM layers. The proposed approach was implemented for load classification in ELV dc picogrid. In the second one, a comparison was performed between Markovian models and a simple RNN model. Additionally, another approach for appliance identification was presented in [156]. The authors combined a dAE and a LSTM network, to reconstruct the appliance specific signal and then to identify in which appliance this signal belongs to. Furthermore, several RNN-based approaches have been published [157, 158, 159, 160, 161, 162, 163, 164] for state detection or load disaggregation while in some occasions RNN models also deployed for real-time disaggregation [165, 166] in residential buildings. Though, until [167] was published, there wasn't a study that had employed RNN for industrial energy disaggregation. The authors constructed a DNN for performing classification. Additionally, in Kalinke et al. [7], the authors adopted the models constructed by Kelly and Knottenbelt [55] and the Online-GRU proposed by Krystalakos et al. [153]. The selection of these RNN-models is based on their compatibility with the open framework NILMTK [168]. Finally, it is necessary to mention recent contributions in NILM literature with RNNs. This work [169], proposed two architectures with parallel LSTMs stacks for appliance power consumption estimation. The first network was composed of three parallel stacks, that each stack contained one 1D convolutional and two LSTM layers, that were concatenated before final fully connected layer. The second topology, also comprised three stacks with the first one based on convolutional layers and the other two on LSTM. Both approaches were trained and tested on REDD dataset. Additionally, Kaselimi et al. [170], a LSTM network with a classical structure is proposed, comprising of two biLSTM layers but with a novel hyperparameters optimization method. The study in [171] presents a more composite LSTM network for energy disaggregation. The main innovation proposed is the one-to-many structure that helped towards improving the model's disaggregation accuracy and also repeated training.

4.7. Autoencoders

An autoencoder is a type of neural network architecture that is having three core components: the encoder, the decoder, and the latent-space representation. The encoder compresses the input to a lower latent-space representation and then the decoder reconstructs it. In NILM, the encoder creates from the aggregate signal the latent space representation and then the decoder tries to create the sub-metered signal. In the early years of deep learning applications in energy disaggregation, denoising autoencoders (dAE) have received enough attention. Kelly and Knottenbelt [55] employed first, a dAE model. The main intuition behind the denoising procedure is that the NILM can be defined as a denoising problem. Based on this, the aggregate power consumption is the combination of the individual appliances' consumptions plus the noise. This assumption was employed in several studies [107, 60, 56] where they proposed dAE improvements with convolutional and fully connected layers to improve disaggregation accuracy. Only in [172], they utilized gaussian noise layer before the fully connected encoder. Beyond the dAE architecture, the authors of this work [173] proposed CAEBN-HC, a 1D-convolutional autoencoder with batch normalization and hyperparameter tuning with hill climbing algorithm. Finally, Faustine et al. [174] adopted the traditional autoencoder-like Unet architecture adjusted for one-dimensional data. They also utilized multitask learning in order to predict both appliances state and power consumption and this strategy proved effective, improving models generalization ability.

4.8. Attention Mechanism

Recurrent Neural Nets and their variations have been the prevailing way to address problems with sequential data. Nevertheless, their weakness to perform parallel computations makes them computationally inefficient. Also the fast exponentially gradient growth or decrease that RNNs and their variants suffer from affects the training of the network. To overcome these boundaries the attention mechanism is introduced as a solution [175]. Attention mechanism, contrary to recurrent architectures uses all the prior states of the encoder in sequence-to-sequence architectures, in order to provide a set of features to the decoder to produce the final representation. Despite this, most of the works still utilized attention along with RNN models. In Vaswani et al. [176] the Transformers architecture is presented, which is based completely on self-attention free of recurrent layers. The core of this model is multi-head attention and has achieved great results in a variety of domains, such as computer vision, NLP, time series forecasting. The intuition behind this module is that a sequence representation contains valuable information in different subspaces. Thus, in the Transformer architecture self-attention is calculated simultaneously in multiple heads and then concatenated and projected again, in order to produce the final sequence vector outputs.

In NILM literature, there is an increasing interest in applying Attention and Transformers in energy disaggregation. The authors of [177] presented two different networks, MA-net and MAED-net. The first one is composed of convolutional and

two transformer layers, 6 identical attention blocks, and with fully connected layers. The second one is a multi-head attention encoder-decoder architecture. The MAED-net is an autoregressive model and as we can observe MA-net and MAED-net share the same configurations. MAED outperformed all the other approaches in the REDD dataset. Another work that applied Transformers in NILM is demonstrated in [178]. The authors proposed an architecture that was inspired by BERT and a novel loss function as a solution to energy disaggregation, achieving state-of-the-art results. Furthermore, three studies recently adopted attention and multi-head attention as a part of their architecture. The first study [179] employed the same model with [153], but the authors replaced the one bi-GRU layer with an attention mechanism. They also performed additive and dot attention, to compare the experimental results. The second study was designed one regression and one classification sub-network, for multi-task learning [180]. For the classification, the sub-network that has been adopted is the sequence-to-sequence architecture that was proposed by Zhang et al. [134]. The multiplication of the regression sub-network output with classification sub-network output produces the final result. The third study proposed COLD, which is a model with several residually connected position-wise feed-forward networks and multi-head self-attention applied before the final output layers [181]. COLD was constructed to perform event detection.

In conclusion, the recent trend of applying attention layers and Transformers architecture on a variety of domains influenced researchers to apply them to the NILM task. Despite showing vast improvement contrary to RNNs, in several tasks, both in terms of accuracy and training time, Transformers still remain a hard solution to implement, so the adaptation of their recent advances that enhance their predictive and computational performance [182, 183, 184] in the NILM task is necessary.

4.9. Deep Generative Models

Deep Generative Models (DGM) are a type of a deep neural network that is trained in a large amount of data, that tries to synthesize high-dimensional distributions. The two most commonly implemented approaches are Variational Autoencoders (VAE) and Generative Adversarial Networks (GAN). In Sirojan et al. [185] a Convolutional Variational Autoencoder that extracted appliance-specific signal from the main is proposed. The suggested model is trained in the UK-DALE dataset achieving state-of-the-art results. Variational Autoencoder for NILM is also implemented in [186], where a Variational Recurrent Neural Network (VRNN) generating appliance signals using as an input the total power consumption is analyzed.

This study [57] introduced a model that is trained to synthesize appliance-specific signals from a latent representation Z. The generator G is responsible for producing from the latent representation device sequences for each appliance. Then the produced sequences and the training appliance sequences are observed from the discriminator D, which tries to determine if the generated sequences are real or fake. The current approach is composed of two components: G and the disaggregator Y_α, where the first one is responsible for mapping a latent representation into an appliance sequence and the Y_α at creating the latent representation from the main signal. In [187], the authors proposed a network that was consisted of three components, namely the seeder S that corresponds to the encoder part in a DAE model, G that is responsible to synthesize new signal waveforms from the dAE's latent space, and D that is a binary classifier that aims to clarify which signals are real or fake. Also, an improvement to this architecture was proposed in [188]. The authors added in the D network two Gated Recurrent Units to perform the discrimination. A cGAN for energy disaggregation is implemented in Pan et al. [189]. The network's G was a 1-D UNet that was fed with the main signal as an input and generated an appliance-specific signal. The D was a fully convolutional network, with several convolutional layers. The authors adopted sequence-to-subsequence output, by choosing the optimal output window based on the training process estimated time. Finally, in [190] GAN-NILM is proposed. The network is composed by an autoencoder as G, being responsible to produce the appliance specific signal. The authors concatenated the mains signal with output of G, claiming that this technique ensures training stability. The discriminator aims to identify if the disaggregated signal was real or fake. Also, second-to-last D layer produces the output features that are fed in the next iteration to G. Also using parameter sharing approach they modeled a transferable Generative model and they tested it in three publicly available datasets.

4.10. Transfer Learning

Transfer Learning (TL) is the process of transferring a neural network's knowledge from one domain to another and usually, it utilizes an already pre-trained model or a part of it. TL can be employed in different scenarios, using for instance a large pre-trained model trained in a large dataset as a feature extractor by removing the network's head or fine-tuning the whole network, or using the pre-trained without intervening. Even though transfer learning is an existing methodology in a variety of domains the lack of pre-trained models in a well-established dataset in energy disaggregation, makes it an open affair. Also, bearing in mind the different characteristics of the publicly available datasets, the different appliances, and energy consumption habits between users, the implementation of both appliance and cross-dataset (TL) is necessary. In Murray et al. [191], two multitask network architectures are proposed aiming to accurately solve both NILM tasks and to facilitate successful transfer learning between different datasets. The work in [192] explores Transfer Learning in energy disaggregation from different aspects. Based on the online NILM tool NILMTK the authors trained two models, Seq2Point and Online-GRU in a range of application scenarios. The two networks have been trained and tested on the same dataset, and then across multiple buildings from multiple datasets, combining energy consumption recordings with different appliance characteristics and usage patterns. Another publication that also utilized the Seq2Point model for TL is suggested in [193]. Seq2Point is a lightweight model that can be easily reproduced. The authors, are examining the generalization ability considering two approaches, appliance and cross-domain transfer learning. In

the first case, the model was trained and tested on the same datasets. Seq2Point was trained on REFIT which is the largest one and Transfer Learning was performed on REDD and UK-DALE. For the appliance TL, the models were pre-trained on the washing machine, and for the other devices the convolutional layers were used as feature extractors and the fully connected layers were fine-tuned. With this approach they managed to reduce, even more, the required training time. For the cross-dataset approach, they have performed the base training on the REFIT dataset and utilized UK-DALE and REDD as target datasets. The authors concluded that the adaptation of transfer learning on datasets that are having the same characteristics helps at increasing disaggregation accuracy. In a domain like NILM, where there is a great need for predictive algorithms that are able to generalize well, future progress in TL approaches could be promising.

4.11. Federated Learning

As mentioned in the section 2, NILM is a problem that relies heavily on data availability. However the required equipment power consumption usually differs among regions, users, equipment type. The collection and storage in a remote server of consumer electric recordings data, is often privacy sensitive and risks potential leaks. Along with the progress made over the last few years in machine/deep learning and in cloud infrastructures a new field was introduced to address this issue, Federated Learning (FL). FL is an alternative that keeps stored all the required data locally on devices and trains a shared model, without the need to centrally store it [194]. The main advantage of this method, is that it reduces potential privacy and security risks, by restricting possible attacks only on the devices.

In NILM domain the training data are containing information about the clients (domestic, commercial, industrial) and possible exposition may result, in data manipulation, infer human actions inside a building or data breach. In the last few years several studies have been published that investigated FL, for energy disaggregation. The work presented in Pötter et al. [195] introduces a framework that is able to train and test NILM utilizing FL. The authors implemented a s2p model combining REDD and UK-DALE datasets for three devices. Finally, their proposed DPFL model found out to be more robust in FL than other non-privacy models. Another study exploring NILM-FL is presented in [196], where the authors train a RNN model locally, using metering data and are then sharing the model parameters. Their approach showed that communication costs can be decreased and that FL trained models can approximate the predictive accuracy of regular trained models. In conclusion, FL can overcome the arisen privacy and security issues in NILM, and help towards creating more generalized models.

5. Performance evaluation and comparability

5.1. NILM evaluation metrics

Performance evaluation in NILM studies is not standardized, as there is not a commonly accepted way of evaluating the disaggregation models with specific metrics. Studies may utilize metrics that are generally used in machine learning (classification/ regression problems), or apply targeted metrics that have been explicitly proposed for load disaggregation systems. The first category includes common metrics such as True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN), Accuracy, Precision, Recall, f-score, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE). On the contrary, some of the most widely used metrics introduced specifically for NILM are the Total Energy Correctly Assigned (TECA) [24], the Energy Accuracy (EA) [197] and the Energy-based f-score [107].

Another fundamental separation is proposed in [197], where evaluation metrics are categorized as Event Detection (ED) or Energy Estimation (EE). ED metrics evaluate the performance of the algorithm in the identification of an appliance's operational state, treating NILM as a classification problem. EE metrics estimate the precise amount of energy consumed by an appliance at each timestamp, following an approach which is used in regression problems. According to the authors' view, the inclusion of both ED and EE metrics enables a thorough assessment of the performance of a disaggregation model. Thus, it is suggested that EE along with ED metrics should be included in the studies, as state detection and energy estimation are equally important. The selection of the most appropriate evaluation metrics should take into consideration two main factors in order to avoid misleading results: a) appliance's operational power level and b) appliance's frequency of usage. The level of power at which an appliance operates, proportionally affects the value (high or low) of certain EE metrics, such as MAE and RMSE. For example, comparing MAE between two appliances operating at 100W and 4000W respectively is not expected to give any meaningful conclusions, since the appliance that operates on a higher power level is expected to have higher MAE. For this reason, MAE and RMSE are considered inappropriate even if they are commonly used in NILM studies. On the contrary, a great alternative could be TECA which can be applicable to appliances with many states or "smooth" power ons and does not rely on removing edges in the aggregate power signal. However, making use of TECA, high power appliances can be naturally weighted more heavily than low power appliances and this may lead to optimistic results for low power appliances. Alternatively, energy-based precision/recall/f-score can be used as they are able to cover multiple aspects of the performance of a NILM model. Energy-based recall measures the portion of the power consumption that has been accurately identified, whereas energy-based precision denotes the amount of energy that has been assigned to an appliance and actually belongs to it. Energy-based f-score is the geometric mean of precision and recall [107].

The frequency of appliance usage also plays an important role, since most of the appliances may operate for short periods of time. In this case, ED metrics give a complete picture of the active and inactive intervals of each appliance, giving the opportunity to specify type I or type II errors made by the predictive model. This way, explainability of the results is facilitated, which in turn allows taking more targeted actions towards the improvement of the predictive models. In the case

of ED metrics, TPs/ TNs/ FPs and FNs seem to be the most intuitive results, as they are the primary source of information for the extraction of other metrics such as precision, recall and f-score.

The authors in [198] highlight the need to quantify certain properties that are related to the complexity of a dataset and moreover define a common format for reporting testing setup and accuracy results. For this purpose, they suggest two new metrics based on the observation that data properties may vary significantly across different chunks of the same dataset. Test Set Ratio (TSR) is defined as the ratio between the test set duration and the total energy time series duration. It is experimentally found that lower TSR leads to more accurate results, hence it can be concluded that comparing two studies on the same dataset but with different TSR, can be a biased comparison. Similarly, Event Ratio (EVR) gives the ratio between the number of events in the test set and the number of events in the whole dataset.

The increasing number of deep learning implementations in the field of NILM, makes the evaluation of the algorithm's computational cost a necessity. Computationally demanding models may have limitations, since they require huge amount of resources for training, and as a result their deployment on real-time applications is difficult. A study which estimates the processing time and the processing energy of a NILM system is presented in [199]. The authors investigate the potential for large load scale deployment of Cloud-based Online-NILM algorithms by comparing the algorithm's computational performance on a dedicated server and a cloud virtual server. It is concluded that more studies should be encouraged to consider the computation cost perspective and not only the disaggregation accuracy obtained.

The lack of a standardized procedure for the evaluation process, leads to comparability issues among different NILM approaches. According to [198], there are multiple factors that characterize a disaggregation approach, such as dataset selection, performance metrics, disaggregation techniques and the evaluation process. All the previous factors comprise the experimental setup of each method. In order to achieve fair assessment among various methods, the experimental setups need to be totally aligned. An important contribution towards this direction is made by nilmtk framework [200], however the integration of custom NILM algorithms to this framework can be complex.

Summarizing, we list the following suggestions that lead towards more easily comparable NILM implementations:

- A general evaluation framework is important to be defined. Regarding the metrics, taking into consideration relevant studies [198], we propose the wider adoption of Energy F-Score, TECA metrics for load disaggregation and TPs/ TNs/ FPs and FNs for event detection.
- TSR and EVR metrics introduced in [198]should be more widely adopted, so that comparisons conducted among studies utilizing the same dataset are facilitated.
- Concerning datasets, the evaluation procedure through cross validation, especially in datasets that contain several buildings, is necessary since each premise maintains different type of information.

5.2. Tables summary

Regarding shallow learning approaches in Table 3, emphasis is given on the dataset being used (low or high frequency) in combination with the extracted features. Moreover, the predictive algorithms tested in each study are reported, as well as the evaluation metrics which define the approach that has been followed (classification or regression). Results suggest that the majority of the studies utilize data from real-life datasets, while only a few prefer simulated data. Concerning the feature extraction process, a lot of studies use only the active power signal and this probably happens due to data unavailability. The inclusion of additional electrical and statistical features has proven to be beneficial for the model's accuracy in multiple studies. It is observed that NILM is treated as a signal processing problem in some cases, making use of frequency domain transformations of the power signal. Chronologically older studies are largely based on HMM models for the selection of predictive algorithm. HMM methods are sometimes used until today as a benchmark, however their performance is significantly lower compared to machine learning models. Tree-based approaches and some more sophisticated variations of classical optimization methods are the solutions being used more often in recent literature studies, nonetheless deep learning methods are dominant in the field at this point. Analyzing the metrics being used by the examined studies, it becomes apparent that most of them are making use of typical metrics applied to classification or regression tasks (accuracy, precision, recall, f-score, RMSE etc.), even if there are specialized metrics suggested for NILM. Furthermore, the majority of the studies does not use both EE and ED metrics, as suggested in 5.1, and as a consequence the evaluation of the models' performance is not straightforward.

Amongst the characteristics that define deep learning models, the learning framework, the input dimensions and the output strategy are considered. Tables 4, 5 present the main components of each reviewed work according to the above elements. The publication data correspond to the year that each work was published, dataset column presents the dataset type that deep learning models were trained and evaluated on, learning framework includes four attributes, learning method (regression or classification), the main elements of the proposed approach, the number of layers and the training criterion. Finally the last two columns present the input and output shape for each approach, respectively.

One of the core elements in DNN training is the loss function that indicates how well the neural network is trained. In the NILM literature, several different loss functions have been employed. As is depicted in the loss column of the Tables 4, 5 most of the works that are published utilize Mean Squared Error for regression and Binary Cross-Entropy for classification. Also, a smaller percentage of published works used MAE due to its limitations and its sensitivity to outliers. Finally, some studies investigated training DNNs with KL divergence or an MSE and KL divergence combination. Though, in the NILM

domain, there is not a deeper investigation of each loss function's impact on the training procedure.

In Section 3.1 the different pre-processing techniques for the mains signal are presented. Different methodologies and approaches have been proposed, aiming at reducing training time and decreasing error rates. Tables 4, 5 present in the final column the output methodology that was adopted in each publication. S2s is the most widely used output approach and it means that the DNN model receives as an input a fixed-length sequence of the mains signal and outputs a sequence of prediction data. The DNN is trained to remove the power consumption contribution of all appliances except for the device under consideration. For each output sliding window, the mean or median of the predicted values is used as the final result. Instead of training computationally large DNNs, the authors of Zhang et al. [134] proposed s2p, at which the model tries to map an input sequence of data to a single output prediction value. The core idea is that the input is a sequence of the total power consumption and the output value corresponds to the midpoint element of the disaggregated appliance. Additionally, sequence-to-subsequence (seq2subseq) was presented as a compromise between the two approaches, claiming that predicting output signals of a shorter sequence will increase disaggregation performance comparatively to the seq2point and reduce the computation time during the inference against seq2seq [189]. Regarding the evaluation methodology that is utilized in the DNN implementations in the NILM context, most of the studies utilize MAE as their performance metric, while several studies also use Energy F-Score and TECA. However, a systematic comparison of the output approaches for DNNs has not been published yet.

5.3. Quantitative Summary

As mentioned in Subsection 5.1, due to the different metrics that are employed and the variety of used datasets, there is not a standard evaluation methodology in NILM literature. This creates a significant obstacle in determining which method performs the best and also hinders the explainability of the predictive model's performance. Based on this, we try to quantitatively summarize the findings that are presented in Tables 4, 5. The focus is directed towards deep learning methods, since those methods are dominating the literature in recent years. Figure 2 illustrates the number of published studies through the years. We divide deep learning approaches based on the categories that we have previously described in Section 4. AE corresponds to autoencoders where its core concept is described in subsection 4.7, AM represents attention mechanism, that includes attention and transformer based methods that are outlined in 4.8, CNN corresponds to those approaches that utilize convolutional neural networks as their backbone and their analysis is found in Subsection 4.5. DGM refers to Deep Generative Models, specifically Variational Autoencoders and Generative Adversarial Networks that are highlighted in Subsection 4.9, while FFNN indicates Feed-forward neural nets, that are reported in 4.4. Finally RNN stands for Recurrent Neural Networks and their variants (LSTM & GRU), presented in 4.6.

In Figure 1 all of the datasets utilized by each one of the studies shown in Tables 3, 4 and 5 are demonstrated. It can be concluded that residential applications are dominant in NILM literature with REDD, UK-DALE, REFIT and AMPds2 being the most widely used datasets. There is also a number of studies utilizing custom datasets that are not publicly available, while fewer studies use simulated data. Industrial datasets have been released relatively recently, so there is not extensive research interest yet. Many studies use more than one dataset for the evaluation of their predictive model. This is an approach which is encouraged by the authors in order to study how the model performs on datasets and appliances with different properties. Moreover, the generalization capability of the algorithms can be studied which still remains an open research topic in NILM.

Table 3: Shallow Learning Approaches for Energy Disaggregation

Reference	Publication Date	Dataset	Method	Features	Predictive Algorithm	Evaluation Metrics
Kolter et al. [201]	2010	Plugwise (private)	r	P	Total Energy Priors Group Lasso Shift Invariant Sparse Coding Discriminative Disaggregation Sparse Coding	Total-week Accuracy
Kolter and Johnson [24]	2011	REDD	r	P	FHMM	TECA
Parson et al. [104]	2012	REDD	r	P	HMM	Energy norm. error RMSE
Zhong et al. [202]	2014	HES	r	P	AFHMM AFAMAP AFHMM+SAC	NDE SAE
Liao et al. [203]	2014	REDD REFIT	c	P (edges detection)	HMM Decision Trees DTW	TP FP FN Precision Recall F-score
Altrabalsi et al. [204]	2014	REDD	c	P Min/Max value Area Event duration	HMM SVM k-means + SVM	Precision Recall F-score
Nguyen et al. [205]	2015	Simulated data	c	P, Q, S	Decision Trees	Accuracy
Alshareef and Morsi [206]	2015	Simulated data	c	DTW (Daubechies wavelet)	AdaBoost	Accuracy
Gillis et al. [94]	2015	Simulated data	c	Energy of wavelet coefficients	Decision Trees	Accuracy 95% conf. interval
Bonfigli et al. [207]	2016	AMPds	r &c	P, Q	AFAMAP Forward Differential AFAMAP	Precision Recall Energy-based f-score
He et al. [113]	2016	REDD REFIT	r & c	P	Graph method	Precision Recall F-score ANE
Bhotto et al. [208]	2016	REDD AMPds	r	P or S	Aided Linear IP	Accuracy (per appliance) Accuracy (Overall)
Tabatabaei et al. [209]	2016	REDD	r & c	P,S DTW (Haar wavelet)	RAndom k-labELsets (RAkEL) Multi-Label kNN (MLkNN)	F-score Energy error
Batra et al. [210]	2017	Dataport	r	Electrical Household info	FHMM Latent Bayesian Melding Discriminative Disaggregation Sparse Codding Gemello/kNN Matrix Factorization	PEC
Jain et al. [211]	2017	BLUED	c	I	Decision Trees SVM k-NN Random Forest Extra Trees	Accuracy Precision Recall F-score
Batra et al. [212]	2018	Dataport	r	Home Appliance Seasonal	Adagrad (Transfer Learning)	PEC RMS PEC weighted PEC
Machlev et al. [112]	2019	REDD AMPds	r & c	P, Q	HART HART w/MAP AFAMAP(P,Q) Multi-objective evolutionary optimization	Accuracy Precision Recall F-score TPCA
Shi et al. [213]	2019	REDD Dataport REFIT	r	P DFT (freq. domain)	Similar Time Window (STW) FHMM Powerlet-based Energy Disaggregation Multilabel Classification Sparse Coding (SC) Discriminative SC Greedy Deep SC Extract Deep SC	Accuracy
Yuan et al. [214]	2019	AMPds	r & c	P	Optimized SVR (OSVR) SVR FHMM DTW	F-score PCE RMSE
Kang et al. [91]	2020	PLAID Private dataset	c	FFT-based (Magnitude and phase of 3rd, 5th and 7th current harmonics)	Bagging Decision Tree	Accuracy Precision Recall F-score
Puente et al. [215]	2020	Georges Hebrail UCI UK-DALE	c	P	Fuzzy Clustering	TP, TN, FP, FN Precision Recall F-score

Table 4: Part I. Deep Learning Approaches for Energy Disaggregation

Reference	Publication Date	Dataset	Learning Framework				Input	Output
			Method	Main Components	Layers	Loss		
Kelly and Knottenbelt [55]	2015	UK-DALE	r & c	RNN dAE CNN	6 6 8	MSE	1D	s2s
Mauch and Yang [150]	2015	REDD	r	LSTM	3	Least Squares	1D	s2p
do Nascimento [59]	2016	REDD	r & c	CNN RCNN LSTM GRU ResNet	17 8 4 4 8	CCE	1D	s2c
Mauch and Yang [216]	2016	REDD	r	FF-HMM	3	Negative log-likelihood	1D	s2s
Kim et al. [152]	2016	UK-DALE	c	GRU RNN	2 2	BCE	1D	on/off
Mottahedi and Asadi [143]	2016	Dataport	c	2D CNN	12	BCE	2D	on/off
He and Chai [151]	2016	UK-DALE	r	dAE CNN-LSTM	2 2	BCE	1D	on/off
Garcia et al. [172]	2017	Custom	r	sdAE	5	Quadratic	1D	s2s
Kim et al. [157]	2017	UK-DALE REDD Custom	r	LSTM	2	MSE	1D	on/off
de Paiva Penha and Castro [145]	2017	REDD	r	CNN	4	BCE	2D	on/off
Zhang et al. [134]	2018	UK-DALE REDD	r	CNN	8	MSE	1D	s2s & s2p
Bonfigli et al. [107]	2018	UK-DALE REDD AMPDs2	r	dAE	14	MSE	1D	s2s
Tsai et al. [60]	2018	Custom commercial	r	dAE	7	MSE	1D	s2s on/off
Krystalakos et al. [153]	2018	UK-DALE	r	CNN-LSTM CNN-GRU	8	MSE	1D	s2p
Chen et al. [138]	2018	REDD	r	CNN-GLU-ResNet	13	MAE	1D	s2s
Valenti et al. [56]	2018	UK-DALE AMPDs2	r & c	dAE	8	MSE	1D	s2s on/off
Martins et al. [6]	2018	IMD	r & c	WaveNet	9 blocks	MAE	1D	s2s on/off
Rafiq et al. [166]	2018	UK-DALE	r & c	CNN-LSTM CNN-GRU	5	MSE	1D	s2p on/off
Sirojan et al. [185]	2018	UK-DALE	r	CVAE	7	MSE w/ KL divergence	1D	s2s
Chang et al. [217]	2018	Custom	r	dAE	5	MSE	1D	s2s
Cho et al. [158]	2018	Dataport	r & c	FF LSTM	2 2	MSE BCE	1D	s2p on/off
Harell et al. [165]	2018	Custom	r & c	LSTM	4	MSE	1D	s2p on/off
Bao et al. [57]	2018	UK-DALE	r	GAN	6 & 6	MSE & BCE	1D	s2s
Bejarano et al. [186]	2019	DataPort REDD	r	VRNN	14	KL divergence	1D	s2s
Harell et al. [141]	2019	AMPds2	r	WaveNet	9 blocks	MSE	1D	s2s
Kaselimi et al. [142]	2019	AMPds	r	CNN	5	MSE	1D	s2s
Xia et al. [218]	2019	UK-DALE WikiEnergy	r	ResNet	6 units	MSE	1D	s2s
Xia et al. [219]	2019	UK-DALE WikiEnergy	r	ResNet w/ Attention	6 units	MSE	1D	s2s
Liang et al. [139]	2019	UK-DALE	r	CNN	24	MSE	1D	s2s s2ss
Buchhop and Ranganathan [129]	2019	Custom	c	FF	2	BCE	1D	on/off
Kyrkou et al. [81]	2019	UK-DALE	c	VGG-16	21	BCE	2D	on/off
Shin et al. [136]	2019	UK-DALE REDD	r & c	CNN w/ Multitask L	8 & 8	MSE & BCE	1D	s2s w/ on/off
Chen et al. [137]	2019	UK-DALE REDD	r & c	CNN w/ Multitask L	8	MSE & BCE & adversarial loss	1D	s2s w/ on/off
Jasiński [128]	2019	ECO	r	FF	3	MSE	1D	s2p
Zhang et al. [69]	2019	REDD	r	CNN	8	MSE	1D	s2p

Table 5: Part II. Deep Learning Approaches for Energy Disaggregation

Reference	Publication Date	Dataset	Learning Framework				Input	Output
			Method	Main Components	Layers	Loss		
Davies et al. [147]	2019	PLAID	c	FF & CNN	2,3,4,5 5,7,9,11	BCE	1D	on/off
Linh and Arboleya [159]	2019	REDD	r & c	RNN	10	Levenberg-Marquardt	1D	s2p on/off
Gopu et al. [160]	2019	AMPds	r & c	RNN	5	nD	1D	s2s
Kaselimi et al. [162, 170]	2019, 2020	AMPds2, AMPds2 REFIT	r	Bi-LSTM w/ Bayesian Opt	4	KL divergence	1D	s2s
Kaselimi et al. [187]	2020	AMPds2, REFIT	r	GAN	6 & 4	MSE	1D	s2s
Kaselimi et al. [188]	2020	AMPds2, REFIT	r	GAN w/ CNN-GRU	6 & 4	MSE	1D	s2s
Yue et al. [178]	2020	REDD UK-DALE	r & c	Transformer	8	MSE + KL divergence + soft-margin	1D	s2s
Faustine et al. [174]	2020	UK-DALE	r & c	UNet	8 blocks	CE + QuantileLoss	1D	s2s on/off
Ahmed et al. [190]	2020	REFIT REDD UK-DALE	r & c	GAN	7 & 5	MSE+adversarial + BCE	1D	s2s
Ayub and M. [220]	2020	ENERTALK	r & c	CNN	4 & 5	MSE	1D	multi-s2p
Lin et al. [177]	2020	REDD	r & c	Multi-Head Attention	8 & 15	MSE	1D	s2s
Bousbiat et al. [83]	2020	SynD UK-DALE REFIT	r	2D CNN	8	MSE	2D	s2s
Barber et al. [135]	2020	REDD UK-DALE REFIT	r	Pruned CNN	8	MSE	1D	s2p
García-Pérez et al. [221]	2020	Custom Commercial	r	dAE	13	MSE	1D	s2s
Yang et al. [144]	2020	REDD UK-DALE	c	VGG-16	21	CCE	2D	on/off
Gkalinikis et al. [179]	2020	REDD UK-DALE	c	CNN-GRU w/ Attention	6	CCE	1D	s2p on/off
Ciancetta et al. [222]	2020	BLUED	c	CNN	8	CCE	1D	on/off
Yadav et al. [167]	2020	Custom Industrial	c	CNN	8	BCE	1D	on/off
Rafiq et al. [163]	2020	UK-DALE, ECO	r & c	bi-LSTM	4	MSE	1D	s2p
Massidda et al. [223]	2020	UK-DALE	c	AE	18	BCE	1D	on/off
Pan et al. [189]	2020	UK-DALE REFIT	c	GAN	8 blocks 4	L_1+BC CCE	1D	s2subseq
Zhang et al. [224]	2020	REDD	c	CNN & clustering	6	CCE	1D	on/off
Kukunuri et al. [225]	2020	UK-DALE REFIT Dataport DRED	r & c	CNN	7	MSE	1D	s2p on/off
Reinhardt and Bouchur [68]	2020	UK-DALE REDD	r	CNN	8	MSE	1D	s2s & s2p
Zhou et al. [226]	2020	UK-DALE	c	ResNet	14 blocks	CCE	1D	on/off
Jiang et al. [227]	2021	REFIT	c	CNN	8	CCE	1D	on/off
Jia et al. [140]	2021	REDD UK-DALE	r	ResNet	8 blocks	MSE	1D	s2p
Çimen et al. [228]	2021	REFIT	r & c	Multitask CNN-GRU	7	MSE & BCE	1D	s2p on/off
Song et al. [164]	2021	REDD UK-DALE REFIT	r & c	LSTM	4	MSE	1D	s2p
Piccialli and Sudoso [180]	2021	REDD UK-DALE	r & c	Attention	10 & 10	MSE BCE	1D	s2s on/off
Moradzadeh et al. [148]	2021	REDD	c	CNN	8	CCE	1D	on/off
Huang et al. [229]	2021	REDD	c	LSTM-BP	4	CCE	1D	on/off
Athanasiadis et al. [230]	2021	REDD	c	CNN	11	BCE	1D	on/off
Kalinke et al. [7]	2021	IMD HIPE	r	CNN DAE RNN GRU	8 6 6 8	MSE	1D	s2s s2p
Bucci et al. [231]	2021	Custom BLUED	c	CNN	8	CCE	1D	on/off
de Diego-Otón et al. [232]	2021	BLUED	c	LSTM & FF	4 & 3	CCE	1D	on/off
Jia et al. [233]	2021	PLAID	c	CNN	4	CCE	2D	on/off

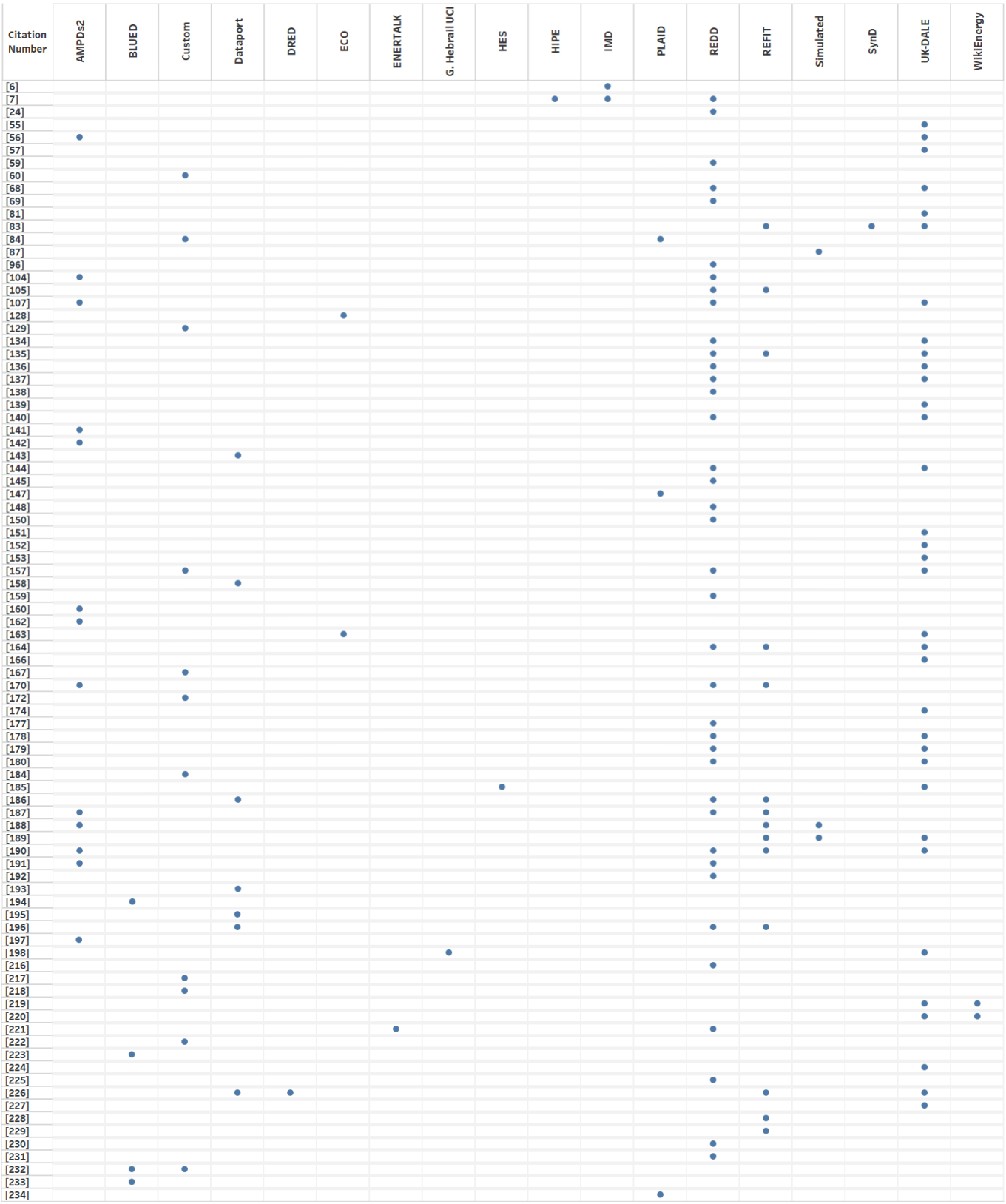

Figure 1: Distribution of NILM datasets across studies

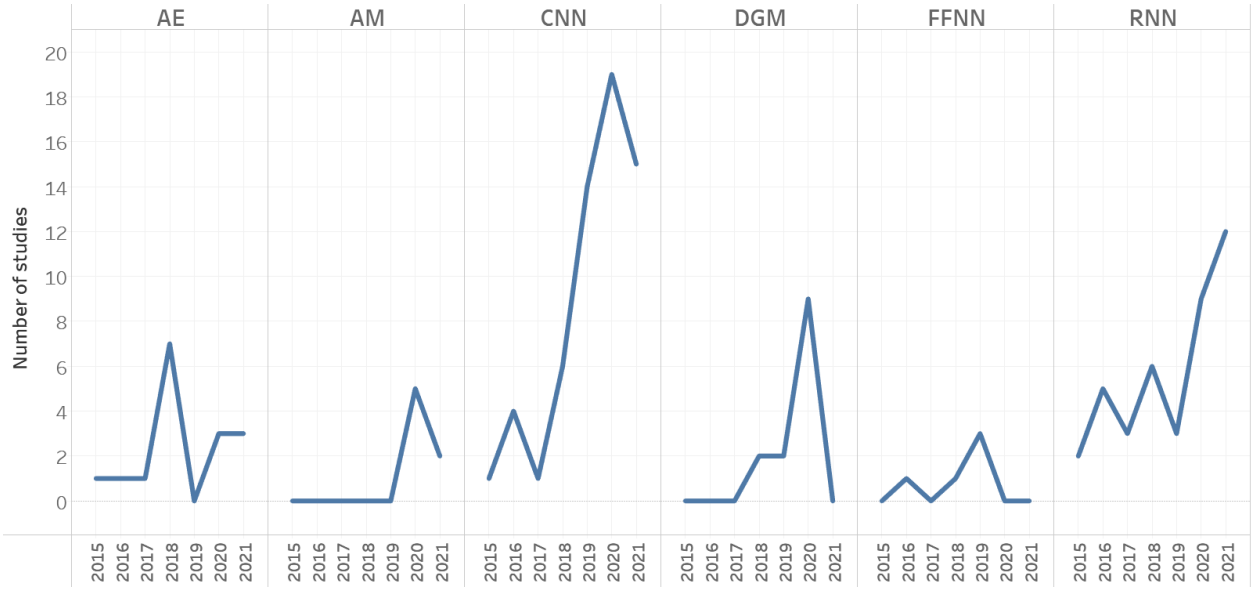

Figure 2: Breaking down DL approaches across years

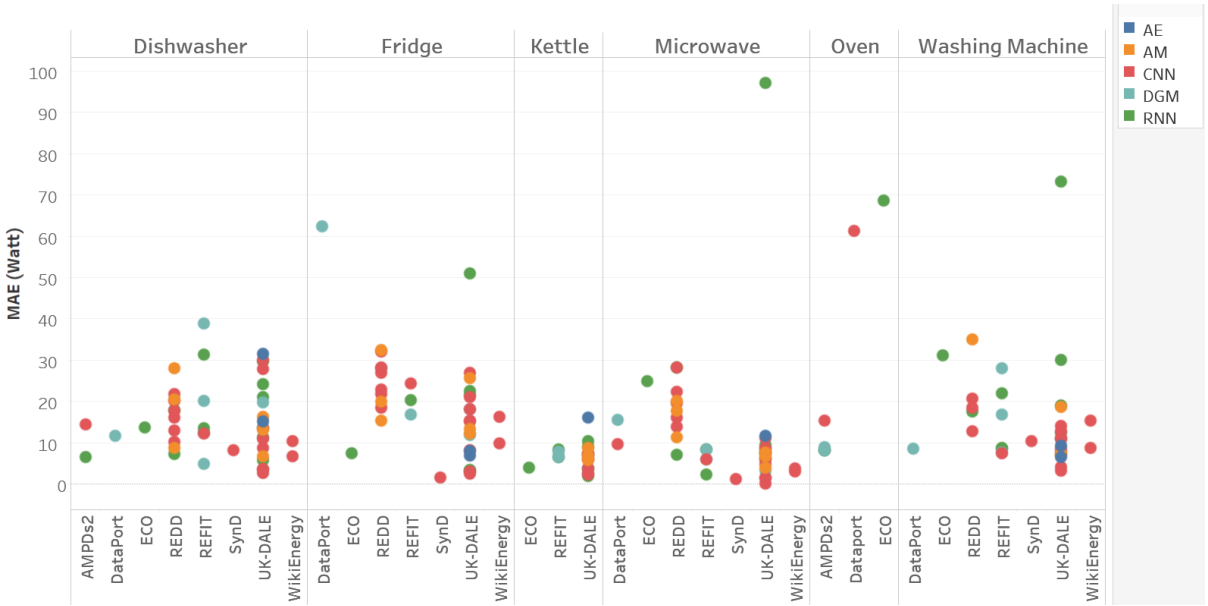

Figure 3: MAE metric per appliance and dataset across various DL approaches

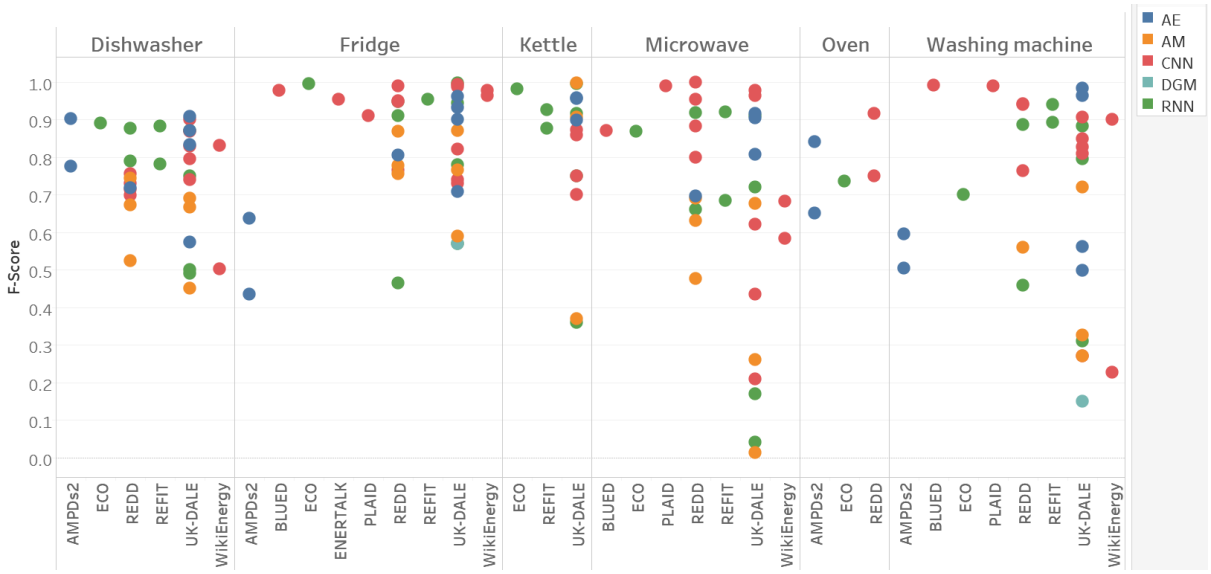

Figure 4: F-score metric per appliance and dataset across various DL approaches

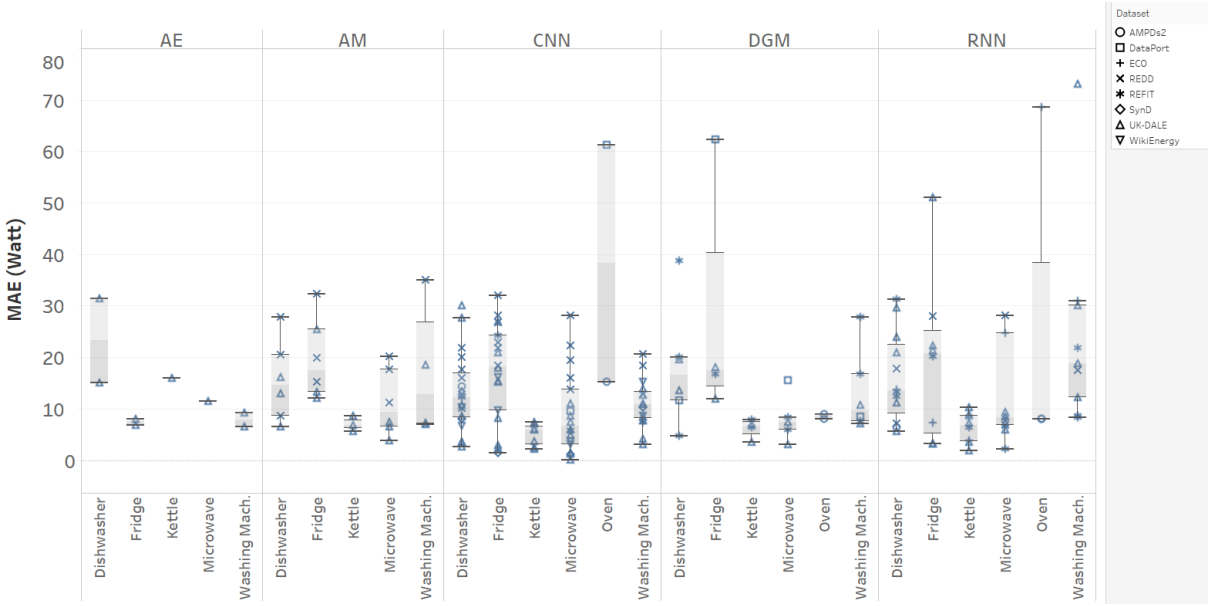

Figure 5: MAE boxplot indicating the performance range of DL approaches per appliance across datasets

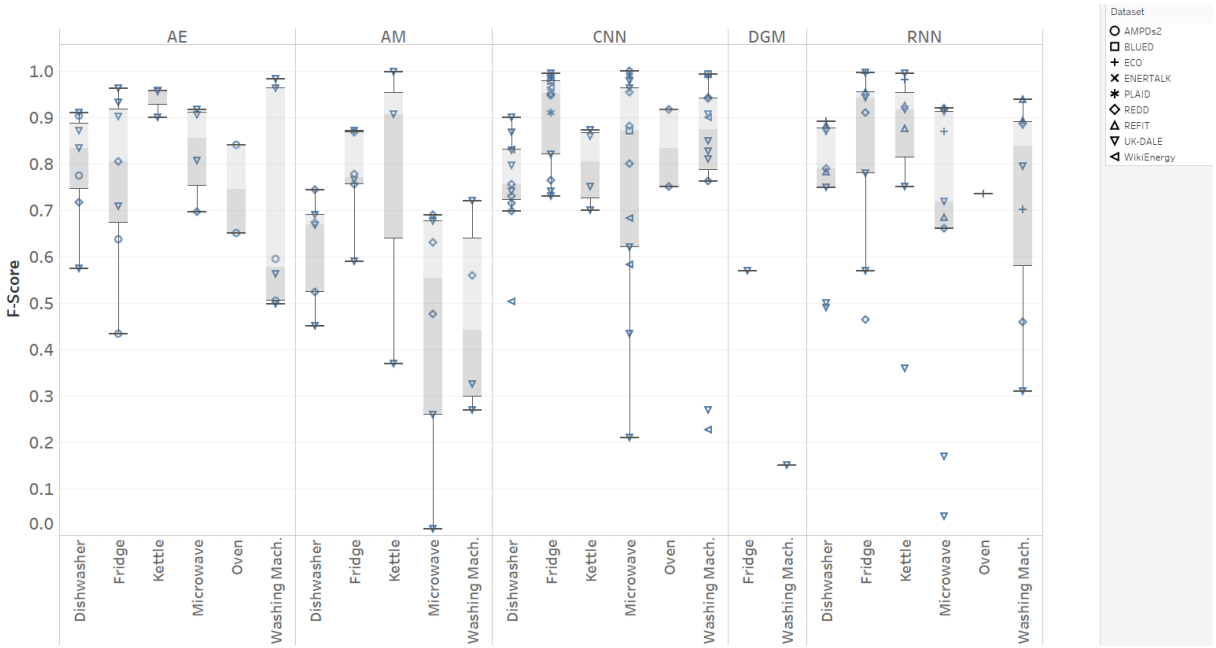

Figure 6: F-score boxplot indicating the performance range of DL approaches per appliance across datasets

Observing Figure 2 we can conclude that Deep Neural Networks are continuously gaining more and more attention. Especially over the last three years a significant number of studies have been published in NILM, adopting also recent deep learning developments such as, GANs and Attention-based networks. Moreover, we observe that CNN architectures are by far the most widely used deep learning approach in NILM literature, representing almost 50% of the implementations that are illustrated in Tables 4 and 5. Additionally, we can notice that there is a trend regarding the application of CNNs in NILM domain, thanks to their multiple variants (i.e s2p, s2s, dilated convolutions, residual nets). Furthermore, as it is shown in Figure 2 RNNs are the second most widely used DL approach. Their proven ability in handling sequential data [234] made them and their variants (LSTM, GRU) an appropriate candidate for the NILM problem. Also, due to their simplicity, Feed-forward networks haven't received much attention, while autoencoder architectures either on their denoised or classical format were proposed as a solution in several studies through the years. Finally, as already mentioned, DGMs, attention mechanism and Transformers have been recently adopted in NILM, since their success in other tasks makes them a promising solution in the energy disaggregation domain.

In Figures 3,4 visualization for MAE and F-score metrics respectively, per appliance and dataset across deep learning methods are illustrated. These two metrics are selected since they are the most commonly used in the publications that are presented in Tables 4 and 5, as each of them is appeared in 45 % of the studies. They are responsible for measuring performance for energy estimation and event detection respectively. For each study, the optimal model was selected, ignoring any other benchmarking models that could possibly be included in the study. Moreover, we obtain results for nine most frequently utilized datasets (Fig. 1) and six appliances. It has to be highlighted that Figures 3, 4 don't present a straightforward comparison across each approach, since the experimental setup and evaluation settings may significantly differentiate across studies. For instance, as it was also mentioned in Section 5.1 regarding MAE metric, individual dataset characteristics such as appliance properties, sampling rate, data balancing, amount of test data, directly affect the evaluation result. The kettle and microwave appliance seem to have the lowest MAE. This is partly explained by the fact that they operate for very short periods of time (usually less than five minutes), hence they remain inactive for the largest part of the day. Focusing on the microwave appliance, we observe that even though MAE is low, f-score is also low. This is justified by the fact that, the great majority of events in microwave appliances are classified as TNs (low MAE), however the existence of FP and FN events affects negatively the f-score metric (low f-score). Examining the fridge appliance, it is obvious that even if it operates on a low power level (100 W) it presents a similar MAE with other energy-intensive appliances. This can be interpreted by the fact that the fridge operates in cycles, so it has many active intervals during the day. Another interesting finding here is that fridge performance can vary significantly across different datasets, which highlights once again the need for conducting comparisons utilizing the same datasets and testing intervals. The dishwasher and washing machine have a similar operation, since their load signature comprises of an energy-intensive part (water pumping and heating) and a low consuming part (spinning cycle). It is quite common for disaggregation models to detect accurately the most energy-intensive part of the operation and fail to identify correctly the low consuming part (FNs). This large number of FN events is depicted in low f-score performance for those two appliances in several studies. Finally, oven is included in relatively few studies for comparison, yet a generally high MAE and f-score are observed since it is an appliance that is easily distinguishable but operates on a high power level. Considerably lower MAE is observed in AMPds dataset, due to the fact that oven appliance operates rarely in this dataset.

For the evaluation of the predictive models, we supplementarily examine the results depicted in MAE and f-score boxplots, depicted in Figures 5 and 6 respectively. It is obvious that concerning MAE, the top-performing studies for each appliance are mostly CNN implementations, while the performance remains high when it comes to state detection (f-score). AM approaches are able to deliver high accuracy results, especially in energy estimation task, even if it is a relatively new methodology followed by a small number of studies up to this point. On the contrary, RNNs perform remarkably better in state detection as it can be seen from the f-score metric. DGMs demonstrate an average performance compared to the rest methods in MAE metric, while only 2 studies explored state detection with DGMs resulting in poor performance. Lastly, AEs implementations also showcase moderate performance both for energy estimation and state detection tasks. It is further observed that appliances that can be approximated as On/Off type, such as kettle and microwave have a fairly consistent range of performance across different studies, while this range becomes much wider when multiple state, or continuous varying consumption appliances are examined. Analyzing the boxplots in more detail, it is confirmed that the same type of appliance may be disaggregated with varying accuracy across different datasets. In many cases, the studies that are found in the whiskers (lower and upper quartiles) of the boxplots belong on the same dataset, which means that this specific appliance is more hard/easy to disaggregate, compared to other appliances of the same type found in other datasets. This observation emphasizes the need of using complexity metrics for the datasets used in NILM research.

6. Further work and limitations

Observing the the current status in NILM literature, it can be identified that deep learning approaches are gaining interest. Most of the presented publications consider NILM as a single task problem, trying to detect either the device status or the precise power consumption of individual electrical loads. Three of the demonstrated methods are trained to perform simultaneously both regression and classification [136, 137, 228]. Their structure shares the same methodology, with a shared model as a feature extractor and two output vectors that are responsible to produce the final representation for each learning approach. The output results are obtained via the multiplication of these

two vectors. This functionality of the corresponding architectures, according to the authors, tends to yield more successful results and reduces noisy estimations. Another noteworthy aspect is that most of the approaches are implemented to disaggregate each device separately. This produces a great limitation since the training of multiple models is a time-consuming procedure. One publication [220] has investigated the estimation of multiple devices at the same time, training a CNN in single and multiple regression setting. They compared both approaches in ENERTALK and REDD dataset. They concluded that in overall, the single point model achieves better performance due to the unique features that are needed for each appliance to perform disaggregation accurately.

However, the multi-target approach is more computationally efficient. The further investigation of multi-target and lightweight deep learning models, in parallel with the increase of computational power, can overcome one of the biggest limitations in NILM domain, the time-consuming training routines. However, all deep learning studies approach NILM as a supervised learning problem. To the best of our knowledge, there is not an existing publication that utilized deep neural networks for unsupervised or non-supervised NILM. Only Hsu et al. [235] investigated self-supervised learning, to perform event detection based on the location of the resident in the home over time. Even though this publication had a great impact on the NILM literature it causes privacy issues.

Finally, the generalization ability of the proposed approaches is an open concern. As demonstrated, in the previous sections of this chapter several deep learning methods deliver decent disaggregation accuracy. But the different evaluation frameworks, the lack of a universal dataset, and the absence of publicly released pre-trained models create benchmarking issues. The last few years several comparison studies have been published in a residential [192, 168, 236, 237, 15] and in an industrial [7] setting. However, most of these publications performed evaluations through the standard literature baselines. A more complete and comprehensive comparison is needed, that will include the recent developments and state-of-the-art models which have been presented in the literature. Summing up, the confrontation of the aforementioned limitations, should be a priority since it will lead to more robust models and a more accurate real-time NILM.

7. Conclusions

The emerging concept of NILM seems to have a dominant role as a service of future smart energy grids, enabling customers to gain control upon their energy usage through increased awareness. Breakdown of energy usage at the appliance level could also help identify anomalies of malfunctioning appliances. The current survey attempts to assess the progress made and the limitations encountered on multiple aspects of NILM studies. The most widely used datasets in the literature are listed, analyzing their characteristics. The advancement of the learning algorithms from HMMs, to shallow learning and ultimately deep learning techniques is presented. Strengths and weaknesses of the commonly used evaluation methods are demonstrated, and comparability across different approaches is discussed. Future research studies are expected to face a number of challenges in trying to suggest approaches that could succeed as real-life applications. Data transmission and storage issues dictate the utilization of relatively low sampling rate on the smart meter data, while at the same time the computational cost of the models' training should not prevent the algorithm's scaling capability. The main focus of future research directions could address one of the biggest weaknesses of NILM algorithms, which is the generalization capability across different datasets/buildings. Finally, current disaggregation solutions are mostly tested on a residential level, while commercial buildings and industrial facilities could have a much larger savings potential.

Acknowledgement

This work is partially funded by the European Union's Horizon 2020 Innovation Action Programme through the PRECEPT H2020 project under Grant Agreement No. 958284.

References

[1] IEA, Net zero by 2050, `https://www.iea.org/reports/net-zero-by-2050`, 2021.

[2] K. Aurangzeb, S. Aslam, S. M. Mohsin, M. Alhussein, A fair pricing mechanism in smart grids for low energy consumption users, IEEE Access 9 (2021) 22035–22044.

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

[4] G. W. Hart, Nonintrusive appliance load monitoring, Proceedings of the IEEE 80 (1992) 1870–1891.

[5] E. Holmegaard, M. B. Kjaergaard, Nilm in an industrial setting: A load characterization and algorithm evaluation, in: 2016 IEEE International Conference on Smart Computing (SMARTCOMP), IEEE, 2016, pp. 1–8.

[6] P. B. Martins, J. G. Gomes, V. B. Nascimento, A. R. de Freitas, Application of a deep learning generative model to load disaggregation for industrial machinery power consumption monitoring, in: 2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), IEEE, 2018, pp. 1–6.

[7] F. Kalinke, P. Bielski, S. Singh, E. Fouché, K. Böhm, An evaluation of nilm approaches on industrial energy-consumption data, in: Proceedings of the Twelfth ACM International Conference on Future Energy Systems, 2021, pp. 239–243.

[8] R. W. Cox, P. L. Bennett, T. D. McKay, J. Paris, S. B. Leeb, Using the non-intrusive load monitor for shipboard supervisory control, in: 2007 IEEE Electric Ship Technologies Symposium, IEEE, 2007, pp. 523–530.

[9] J. Paris, Z. Remscrim, K. P. Douglas, S. B. Leeb, R. W. Cox, S. T. Galvin, S. G. Coe, J. R. Haag, J. A. Goshorn, Scalability of non-intrusive load monitoring for shipboard applications (2009).

[10] M. Zeifman, K. Roth, Nonintrusive appliance load monitoring: Review and outlook, IEEE transactions on Consumer Electronics 57 (2011) 76–84.

[11] M. Baranski, J. Voss, Nonintrusive appliance load monitoring based on an optical sensor, in: 2003 IEEE Bologna Power Tech Conference Proceedings,, volume 4, IEEE, 2003, pp. 8–pp.

[12] B. Neenan, J. Robinson, R. Boisvert, Residential electricity use feedback: A research synthesis and economic framework, Electric Power Research Institute 3 (2009).

[13] K. Ehrhardt-Martinez, K. A. Donnelly, S. Laitner, et al., Advanced metering initiatives and residential feedback programs: a meta-review for household electricity-saving opportunities, American Council for an Energy-Efficient Economy Washington, DC, 2010.

[14] H. K. Iqbal, F. H. Malik, A. Muhammad, M. A. Qureshi, M. N. Abbasi, A. R. Chishti, A critical review of state-of-the-art non-intrusive load monitoring datasets, Electric Power Systems Research (2020) 106921.

[15] A. Verma, A. Anwar, M. Mahmud, M. Ahmed, A. Kouzani, A comprehensive review on the nilm algorithms for energy disaggregation, arXiv preprint arXiv:2102.12578 (2021).

[16] P. Huber, A. Calatroni, A. Rumsch, A. Paice, Review on deep neural networks applied to low-frequency nilm, Energies 14 (2021) 2390.

[17] M. Zeifman, K. Roth, Nonintrusive appliance load monitoring: Review and outlook, IEEE transactions on Consumer Electronics 57 (2011) 76–84.

[18] C. Klemenjak, P. Goldsborough, Non-intrusive load monitoring: A review and outlook, arXiv preprint arXiv:1610.01191 (2016).

[19] S. S. Hosseini, K. Agbossou, S. Kelouwani, A. Cardenas, Non-intrusive load monitoring through home energy management systems: A comprehensive review, Renewable and Sustainable Energy Reviews 79 (2017) 1266–1274.

[20] L. Pereira, N. Nunes, Performance evaluation in non-intrusive load monitoring: Datasets, metrics, and tools—a review, Wiley Interdisciplinary Reviews: data mining and knowledge discovery 8 (2018) e1265.

[21] R. Bonfigli, S. Squartini, Machine learning approaches to non-intrusive load monitoring, Springer, 2020.

[22] P. G. Donato, Á. Hernández, M. A. Funes, I. Carugati, R. Nieto, J. Ureña, Review of nilm applications in smart grids: power quality assessment and assisted independent living, in: 2020 Argentine Conference on Automatic Control (AADECA), IEEE, 2020, pp. 1–6.

[23] H. Salem, M. Sayed-Mouchaweh, M. Tagina, A review on non-intrusive load monitoring approaches based on machine learning, in: Artificial Intelligence Techniques for a Scalable Energy Transition, Springer, 2020, pp. 109–131.

[24] J. Z. Kolter, M. J. Johnson, Redd: A public data set for energy disaggregation research, in: Workshop on data mining applications in sustainability (SIGKDD), San Diego, CA, volume 25, 2011, pp. 59–62.

[25] J. Kelly, W. Knottenbelt, The uk-dale dataset, domestic appliance-level electricity demand and whole-house demand from five uk homes, Scientific data 2 (2015) 1–14.

[26] S. Makonin, F. Popowich, L. Bartram, B. Gill, I. V. Bajić, Ampds: A public dataset for load disaggregation and eco-feedback research, in: 2013 IEEE electrical power & energy conference, IEEE, 2013, pp. 1–6.

[27] D. Murray, L. Stankovic, V. Stankovic, An electrical load measurements dataset of united kingdom households from a two-year longitudinal study, Scientific data 4 (2017) 1–12.

[28] O. Parson, G. Fisher, A. Hersey, N. Batra, J. Kelly, A. Singh, W. Knottenbelt, A. Rogers, Dataport and nilmtk: A building data set designed for non-intrusive load monitoring, in: 2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2015, pp. 210–214. doi:10.1109/GlobalSIP.2015.7418187.

[29] C. Beckel, W. Kleiminger, R. Cicchetti, T. Staake, S. Santini, The eco data set and the performance of non-intrusive load monitoring algorithms, in: Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings, BuildSys '14, Association for Computing Machinery, New York, NY, USA, 2014, p. 80–89. URL: https://doi.org/10.1145/2674061.2674064. doi:10.1145/2674061.2674064.

[30] C. Shin, E. Lee, J. Han, J. Yim, W. Rhee, H. Lee, The enertalk dataset, 15 hz electricity consumption data from 22 houses in korea, Scientific data 6 (2019) 1–13.

[31] N. Batra, M. Gulati, A. Singh, M. B. Srivastava, It's different: Insights into home energy consumption in india, in: Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings, 2013, pp. 1–8.

[32] K. Anderson, A. Ocneanu, D. Benitez, D. Carlson, A. Rowe, M. Berges, Blued: A fully labeled public dataset for event-based non-intrusive load monitoring research, Proceedings of the 2nd KDD Workshop on Data Mining Applications in Sustainability (SustKDD) (2012) 1–5.

[33] J. Gao, S. Giri, E. C. Kara, M. Bergés, Plaid: A public dataset of high-resoultion electrical appliance measurements for load identification research: Demo abstract, in: Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings, BuildSys '14, Association for Computing Machinery, New York, NY, USA, 2014, p. 198–199. URL: https://doi.org/10.1145/2674061.2675032. doi:10.1145/2674061.2675032.

[34] L. D. Baets, C. Develder, T. Dhaene, D. Deschrijver, J. Gao, M. Bergés, Handling imbalance in an extended plaid, 2017 Sustainable Internet and ICT for Sustainability (SustainIT) (2017) 1–5.

[35] R. Medico, L. De Baets, J. Gao, S. Giri, E. Kara, T. Dhaene, C. Develder, M. Bergés, D. Deschrijver, A voltage and current measurement dataset for plug load appliance identification in households, Scientific data 7 (2020) 1–10.

[36] A. S. Uttama Nambi, A. Reyes Lua, V. R. Prasad, Loced: Location-aware energy disaggregation framework, in: Proceedings of the 2nd acm international conference on embedded systems for energy-efficient built environments, 2015, pp. 45–54.

[37] C. Klemenjak, C. Kovatsch, M. Herold, W. Elmenreich, A synthetic energy dataset for non-intrusive load monitoring in households, Scientific data 7 (2020) 1–17.

[38] G. Hebrail, Uci machine learning repository: Individual household electric power consumption data set, https://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption, 2012.

[39] N. Batra, O. Parson, M. Berges, A. Singh, A. Rogers, A comparison of non-intrusive load monitoring methods for commercial and residential buildings, arXiv preprint arXiv:1408.6595 (2014).

[40] S. Bischof, H. Trittenbach, M. Vollmer, D. Werle, T. Blank, K. Böhm, Hipe – an energy-status-data set from industrial production, in: Proceedings of ACM e-Energy (e-Energy 2018), ACM, New York, NY, USA, 2018, pp. 599–603.

[41] M. Noor, A. Yahaya, N. A. Ramli, A. M. M. Al Bakri, Filling missing data using interpolation methods: Study on the effect of fitting distribution, volume 594, Trans Tech Publ, 2014.

[42] A. Allik, A. Annuk, Interpolation of intra-hourly electricity consumption and production data, in: 2017 IEEE 6th International Conference on Renewable Energy Research and Applications (ICRERA), 2017, pp. 131–136. doi:10.1109/ICRERA.2017.8191254.

[43] K. C. Armel, A. Gupta, G. Shrimali, A. Albert, Is disaggregation the holy grail of energy efficiency? the case of electricity, Energy Policy 52 (2013) 213–234.

[44] C. Shin, S. Rho, H. Lee, W. Rhee, Data requirements for applying machine learning to energy disaggregation, Energies 12 (2019) 1696.

[45] J. Huchtkoetter, A. Reinhardt, On the impact of temporal data resolution on the accuracy of non-intrusive load monitoring, in: Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, 2020, pp. 270–273.

[46] A. Antoniou, A. Storkey, H. Edwards, Data augmentation generative adversarial networks, arXiv preprint arXiv:1711.04340 (2017).

[47] J. Wang, L. Perez, et al., The effectiveness of data augmentation in image classification using deep learning, Convolutional Neural Networks Vis. Recognit 11 (2017).

[48] A. Le Guennec, S. Malinowski, R. Tavenard, Data augmentation for time series classification using convolutional neural networks, in: ECML/PKDD workshop on advanced analytics and learning on temporal data, 2016.

[49] Z. Cui, W. Chen, Y. Chen, Multi-scale convolutional neural networks for time series classification, arXiv preprint arXiv:1603.06995 (2016).

[50] J. Gao, X. Song, Q. Wen, P. Wang, L. Sun, H. Xu, Robusttad: Robust time series anomaly detection via decomposition and convolutional neural networks, arXiv preprint arXiv:2002.09545 (2020).

[51] X. Teng, T. Wang, X. Zhang, L. Lan, Z. Luo, Enhancing stock price trend prediction via a time-sensitive data augmentation method, Complex. 2020 (2020) 6737951:1–6737951:8.

[52] C. Esteban, S. L. Hyland, G. Rätsch, Real-valued (medical) time series generation with recurrent conditional gans, arXiv preprint arXiv:1706.02633 (2017).

[53] S. K. Lim, Y. Loo, N.-T. Tran, N.-M. Cheung, G. Roig, Y. Elovici, Doping: Generative data augmentation for unsupervised anomaly detection with gan, in: 2018 IEEE International Conference on Data Mining (ICDM), IEEE, 2018, pp. 1122–1127.

[54] J. Yoon, D. Jarrett, M. van der Schaar, Time-series generative adversarial networks, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/

file/c9efe5f26cd17ba6216bbe2a7d26d490-Paper.pdf.

[55] J. Kelly, W. Knottenbelt, Neural nilm: Deep neural networks applied to energy disaggregation, in: Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments, BuildSys '15, Association for Computing Machinery, New York, NY, USA, 2015, p. 55–64. URL: https://doi.org/10.1145/2821650.2821672. doi:10.1145/2821650.2821672.

[56] M. Valenti, R. Bonfigli, E. Principi, S. Squartini, Exploiting the reactive power in deep neural models for non-intrusive load monitoring, in: 2018 International Joint Conference on Neural Networks (IJCNN), IEEE, 2018, pp. 1–8.

[57] K. Bao, K. Ibrahimov, M. Wagner, H. Schmeck, Enhancing neural non-intrusive load monitoring with generative adversarial networks, Energy Informatics 1 (2018) 18.

[58] F.-Y. Chang, W.-J. Ho, An analysis of semi-supervised learning approaches in low-rate energy disaggregation, in: 2019 3rd International Conference on Smart Grid and Smart Cities (ICSGSC), IEEE, 2019, pp. 145–150.

[59] P. P. M. do Nascimento, Applications of deep learning techniques on nilm, Diss. Universidade Federal do Rio de Janeiro (2016).

[60] C.-W. Tsai, C.-W. Yang, W.-J. Ho, Z.-X. Yin, K.-C. Chiang, Using autoencoder network to implement non-intrusive load monitoring of small and medium business customer, in: 2018 IEEE International Conference on Applied System Invention (ICASI), IEEE, 2018, pp. 433–436.

[61] T. L. Quy, S. Zerr, E. Ntoutsi, W. Nejdl, Data augmentation for dealing with low sampling rates in nilm, arXiv preprint arXiv:2104.02055 (2021).

[62] A. Delfosse, G. Hebrail, A. Zerroug, Deep learning applied to nilm: is data augmentation worth for energy disaggregation?, in: ECAI 2020, IOS Press, 2020, pp. 2972–2977.

[63] G. H. Ribeiro, P. S. d. M. Neto, G. D. Cavalcanti, R. Tsang, Lag selection for time series forecasting using particle swarm optimization, in: The 2011 International Joint Conference on Neural Networks, IEEE, 2011, pp. 2437–2444.

[64] S. Bouktif, A. Fiaz, A. Ouni, M. A. Serhani, Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches, Energies 11 (2018) 1636.

[65] L. Munkhdalai, T. Munkhdalai, K. H. Park, T. Amarbayasgalan, E. Batbaatar, H. W. Park, K. H. Ryu, An end-to-end adaptive input selection with dynamic weights for forecasting multivariate time series, IEEE Access 7 (2019) 99099–99114.

[66] H. Abbasimehr, M. Shabani, M. Yousefi, An optimized model using lstm network for demand forecasting, Computers & industrial engineering 143 (2020) 106435.

[67] K. A. Koparanov, K. K. Georgiev, V. A. Shterev, Lookback period, epochs and hidden states effect on time series prediction using a lstm based neural network, in: 2020 28th National Conference with International Participation (TELECOM), IEEE, 2020, pp. 61–64.

[68] A. Reinhardt, M. Bouchur, 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, NILM'20, Association for Computing Machinery, New York, NY, USA, 2020, p. 75–78. URL: https://doi.org/10.1145/3427771.3427857. doi:10.1145/3427771.3427857.

[69] Y. Zhang, G. Yang, S. Ma, Non-intrusive load monitoring based on convolutional neural network with differential input, Procedia CIRP 83 (2019) 670–674.

[70] S. Bhanja, A. Das, Impact of data normalization on deep neural network for time series forecasting, arXiv preprint arXiv:1812.05519 (2018).

[71] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International conference on machine learning, PMLR, 2015, pp. 448–456.

[72] D. Ulyanov, A. Vedaldi, V. Lempitsky, Instance normalization: The missing ingredient for fast stylization, arXiv preprint arXiv:1607.08022 (2016).

[73] L. J. Ba, J. R. Kiros, G. E. Hinton, Layer normalization, CoRR abs/1607.06450 (2016). URL: http://arxiv.org/abs/1607.06450.

[74] Z. Wang, T. Oates, Imaging time-series to improve classification and imputation, in: Twenty-Fourth International Joint Conference on Artificial Intelligence, 2015.

[75] O. B. Sezer, A. M. Ozbayoglu, Algorithmic financial trading with deep convolutional neural networks: Time series to image conversion approach, Applied Soft Computing 70 (2018) 525–538.

[76] J.-H. Chen, Y.-C. Tsai, Encoding candlesticks as images for pattern classification using convolutional neural networks, Financial Innovation 6 (2020) 1–19.

[77] S. Barra, S. M. Carta, A. Corriga, A. S. Podda, D. R. Recupero, Deep learning and time series-to-image encoding for financial forecasting, IEEE/CAA Journal of Automatica Sinica 7 (2020) 683–692. doi:10.1109/JAS.2020.1003132.

[78] J.-P. Eckmann, S. O. Kamphorst, D. Ruelle, et al., Recurrence plots of dynamical systems, World Scientific Series on Nonlinear Science Series A 16 (1995) 441–446.

[79] L. De Baets, C. Develder, T. Dhaene, D. Deschrijver, Automated classification of appliances using elliptical fourier descriptors, in: 2017 IEEE International Conference on Smart Grid Communications (SmartGridComm), IEEE, 2017, pp. 153–158.

[80] R. de Paula Rodrigues, P. M. da Silveira, Curvature scale space-based signatures for electrical load classification in nilm, Electrical Engineering 103 (2021) 1239–1252.

[81] L. Kyrkou, C. Nalmpantis, D. Vrakas, Imaging time-series for nilm, in: International Conference on Engineering Applications of Neural Networks, Springer, 2019, pp. 188–196.

[82] D. L. Cavalca, R. A. S. Fernandes, Recurrence plots and convolutional neural networks applied to nonintrusive load monitoring, in: 2020 IEEE Power Energy Society General Meeting (PESGM), 2020, pp. 1–5. doi:10.1109/PESGM41954.2020.9281660.

[83] H. Bousbiat, C. Klemenjak, W. Elmenreich, Exploring time series imaging for load disaggregation, in: Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, 2020, pp. 254–257.

[84] J. Liang, S. K. Ng, G. Kendall, J. W. Cheng, Load signature study—part i: Basic concept, structure, and methodology, IEEE transactions on power Delivery 25 (2009) 551–560.

[85] R. Bonfigli, E. Principi, M. Fagiani, M. Severini, S. Squartini, F. Piazza, Non-intrusive load monitoring by using active and reactive power in additive factorial hidden markov models, Applied Energy 208 (2017) 1590–1607.

[86] A. Zoha, A. Gluhak, M. A. Imran, S. Rajasegarar, Non-intrusive load monitoring approaches for disaggregated energy sensing: A survey, Sensors 12 (2012) 16838–16866.

[87] B. M. Mulinari, D. P. de Campos, C. H. da Costa, H. C. Ancelmo, A. E. Lazzaretti, E. Oroski, C. R. Lima, D. P. Renaux, F. Pottker, R. R. Linhares, A new set of steady-state and transient features for power signature analysis based on vi trajectory, in: 2019 IEEE PES Innovative Smart Grid Technologies Conference-Latin America (ISGT Latin America), IEEE, 2019, pp. 1–6.

[88] P. A. Schirmer, I. Mporas, Statistical and electrical features evaluation for electrical appliances energy disaggregation, Sustainability 11 (2019) 3222.

[89] A. Cole, A. Albicki, Nonintrusive identification of electrical loads in a three-phase environment based on harmonic content, in: Proceedings of the 17th IEEE Instrumentation and Measurement Technology Conference [Cat. No. 00CH37066], volume 1, IEEE, 2000, pp. 24–29.

[90] N. Sadeghianpourhamami, J. Ruyssinck, D. Deschrijver, T. Dhaene, C. Develder, Comprehensive feature selection for appliance classification in nilm, Energy and Buildings 151 (2017) 98–106.

[91] H. Kang, H. Kim, et al., Household appliance classification using lower odd-numbered harmonics and the bagging decision tree, IEEE Access 8 (2020) 55937–55952.

[92] A. S. Bouhouras, P. A. Gkaidatzis, E. Panagiotou, N. Poulakis, G. C. Christoforidis, A nilm algorithm with enhanced disaggregation scheme under harmonic current vectors, Energy and Buildings 183 (2019) 392–407.

[93] L. Guo, S. Wang, H. Chen, Q. Shi, A load identification method based on active deep learning and discrete wavelet transform, IEEE Access 8 (2020) 113932–113942.

[94] J. M. Gillis, S. M. Alshareef, W. G. Morsi, Nonintrusive load monitoring using wavelet design and machine learning, IEEE Transactions on Smart Grid 7 (2015) 320–328.

[95] C. Duarte, P. Delmar, K. W. Goossen, K. Barner, E. Gomez-Luna, Non-

intrusive load monitoring based on switching voltage transients and wavelet transforms, in: 2012 Future of Instrumentation International Workshop (FIIW) Proceedings, IEEE, 2012, pp. 1–4.

[96] H. Kim, S. Lim, Temporal patternization of power signatures for appliance classification in nilm, Energies 14 (2021) 2931.

[97] C. Nalmpantis, D. Vrakas, Machine learning approaches for non-intrusive load monitoring: from qualitative to quantitative comparation, Artificial Intelligence Review 52 (2019) 217–243.

[98] M. Zeifman, Disaggregation of home energy display data using probabilistic approach, IEEE Transactions on Consumer Electronics 58 (2012) 23–31.

[99] G. Petneházi, Recurrent neural networks for time series forecasting, arXiv preprint arXiv:1901.00069 (2019).

[100] Z. Wang, G. Zheng, Residential appliances identification and monitoring by a nonintrusive method, IEEE transactions on Smart Grid 3 (2011) 80–92.

[101] R. Brown, N. Ghavami, M. Adjrad, M. Ghavami, S. Dudley, et al., Occupancy based household energy disaggregation using ultra wideband radar and electrical signature profiles, Energy and Buildings 141 (2017) 134–141.

[102] G. Tang, Z. Ling, F. Li, D. Tang, J. Tang, Occupancy-aided energy disaggregation, Computer Networks 117 (2017) 42–51.

[103] H. Kim, M. Marwah, M. Arlitt, G. Lyon, J. Han, Unsupervised disaggregation of low frequency power measurements, in: Proceedings of the 2011 SIAM international conference on data mining, SIAM, 2011, pp. 747–758.

[104] O. Parson, S. Ghosh, M. Weal, A. Rogers, Non-intrusive load monitoring using prior models of general appliance types, in: Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012.

[105] Z. Ghahramani, M. I. Jordan, Factorial hidden markov models, Machine learning 29 (1997) 245–273.

[106] J. Z. Kolter, T. Jaakkola, Approximate inference in additive factorial hmms with application to energy disaggregation, in: Artificial intelligence and statistics, PMLR, 2012, pp. 1472–1482.

[107] R. Bonfigli, A. Felicetti, E. Principi, M. Fagiani, S. Squartini, F. Piazza, Denoising autoencoders for non-intrusive load monitoring: improvements and comparative evaluation, Energy and Buildings 158 (2018) 1461–1474.

[108] Q. Liu, K. M. Kamoto, X. Liu, M. Sun, N. Linge, Low-complexity non-intrusive load monitoring using unsupervised learning and generalized appliance models, IEEE Transactions on Consumer Electronics 65 (2019) 28–37.

[109] M. A. Mengistu, A. A. Girmay, C. Camarda, A. Acquaviva, E. Patti, A cloud-based on-line disaggregation algorithm for home appliance loads, IEEE Transactions on Smart Grid 10 (2018) 3430–3439.

[110] A. Cominola, M. Giuliani, D. Piga, A. Castelletti, A. E. Rizzoli, A hybrid signature-based iterative disaggregation algorithm for non-intrusive load monitoring, Applied energy 185 (2017) 331–344.

[111] Y.-H. Lin, M.-S. Tsai, C.-S. Chen, Applications of fuzzy classification with fuzzy c-means clustering and optimization strategies for load identification in nilm systems, in: 2011 IEEE international conference on fuzzy systems (FUZZ-IEEE 2011), IEEE, 2011, pp. 859–866.

[112] R. Machlev, J. Belikov, Y. Beck, Y. Levron, Mo-nilm: A multi-objective evolutionary algorithm for nilm classification, Energy and Buildings 199 (2019) 134–144.

[113] K. He, L. Stankovic, J. Liao, V. Stankovic, Non-intrusive load disaggregation using graph signal processing, IEEE Transactions on Smart Grid 9 (2016) 1739–1747.

[114] E. Elhamifar, S. Sastry, Energy disaggregation via learning powerlets and sparse coding, in: Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

[115] D. Piga, A. Cominola, M. Giuliani, A. Castelletti, A. E. Rizzoli, Sparse optimization for automated energy end use disaggregation, IEEE Transactions on Control Systems Technology 24 (2015) 1044–1051.

[116] M. Figueiredo, B. Ribeiro, A. de Almeida, Electrical signal source separation via nonnegative tensor factorization using on site measurements in a smart home, IEEE Transactions on Instrumentation and Measurement 63 (2013) 364–373.

[117] H. Gonçalves, A. Ocneanu, M. Bergés, R. Fan, Unsupervised disaggregation of appliances using aggregated consumption data, in: The 1st KDD Workshop on Data Mining Applications in Sustainability (SustKDD), 2011.

[118] L. De Baets, J. Ruyssinck, C. Develder, T. Dhaene, D. Deschrijver, On the bayesian optimization and robustness of event detection methods in nilm, Energy and Buildings 145 (2017) 57–66.

[119] M. Singh, S. Kumar, S. Semwal, R. Prasad, Residential load signature analysis for their segregation using wavelet—svm, in: Power Electronics and Renewable Energy Systems, Springer, 2015, pp. 863–871.

[120] F. Gong, N. Han, Y. Zhou, S. Chen, D. Li, S. Tian, A svm optimized by particle swarm optimization approach to load disaggregation in non-intrusive load monitoring in smart homes, in: 2019 IEEE 3rd Conference on Energy Internet and Energy System Integration (EI2), IEEE, 2019, pp. 1793–1797.

[121] C. C. Yang, C. S. Soh, V. V. Yap, A non-intrusive appliance load monitoring for efficient energy consumption based on naive bayes classifier, Sustainable Computing: Informatics and Systems 14 (2017) 34–42.

[122] P. Meehan, C. McArdle, S. Daniels, An efficient, scalable time-frequency method for tracking energy usage of domestic appliances using a two-step classification algorithm, Energies 7 (2014) 7041–7066.

[123] F. Hidiyanto, A. Halim, Knn methods with varied k, distance and training data to disaggregate nilm with similar load characteristic, in: Proceedings of the 3rd Asia Pacific Conference on Research in Industrial and Systems Engineering 2020, 2020, pp. 93–99.

[124] D. Chowdhury, M. M. Hasan, Non-intrusive load monitoring using ensemble empirical mode decomposition and random forest classifier, in: Proceedings of the International Conference on Digital Image and Signal Processing (DISP), Oxford, UK, 2019, pp. 29–30.

[125] Z. Xiao, W. Gang, J. Yuan, Y. Zhang, C. Fan, Cooling load disaggregation using a nilm method based on random forest for smart buildings, Sustainable Cities and Society 74 (2021) 103202.

[126] X. Wu, Y. Gao, D. Jiao, Multi-label classification based on random forest algorithm for non-intrusive load monitoring system, Processes 7 (2019) 337.

[127] Z. Chen, J. Chen, X. Xu, S. Peng, J. Xiao, H. Qiao, Non-intrusive load monitoring based on feature extraction of change-point and xgboost classifier, in: 2020 IEEE 4th Conference on Energy Internet and Energy System Integration (EI2), 2020, pp. 2652–2656. doi:10.1109/EI250167.2020.9347014.

[128] T. Jasiński, Modelling the disaggregated demand for electricity at the level of residential buildings with the use of artificial neural networks (deep learning approach), in: MATEC Web of Conferences, volume 282, EDP Sciences, 2019, p. 02077.

[129] S. J. Buchhop, P. Ranganathan, Residential load identification based on load profile using artificial neural network (ann), in: 2019 North American Power Symposium (NAPS), IEEE, 2019, pp. 1–6.

[130] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[131] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[132] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, Wavenet: A generative model for raw audio, arXiv preprint arXiv:1609.03499 (2016).

[133] X. Bai, B. Shi, C. Zhang, X. Cai, L. Qi, Text/non-text image classification in the wild with convolutional neural networks, Pattern Recognition 66 (2017) 437–446.

[134] C. Zhang, M. Zhong, Z. Wang, N. Goddard, C. Sutton, Sequence-to-point learning with neural networks for nonintrusive load monitoring, in: Thirty-Second AAAI Conference on Artificial Intelligence, volume 32, AIII Press, 2018, p. 2604.

[135] J. Barber, H. Cuayáhuitl, M. Zhong, W. Luan, Lightweight non-intrusive load monitoring employing pruned sequence-to-point learning, in: Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring, 2020, pp. 11–15.

[136] C. Shin, S. Joo, J. Yim, H. Lee, T. Moon, W. Rhee, Subtask gated networks for non-intrusive load monitoring, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 1150–1157.

[137] K. Chen, Y. Zhang, Q. Wang, J. Hu, H. Fan, J. He, Scale-and context-aware convolutional non-intrusive load monitoring, IEEE Transactions on Power Systems 35 (2019) 2362–2373.

[138] K. Chen, Q. Wang, Z. He, K. Chen, J. Hu, J. He, Convolutional sequence

to sequence non-intrusive load monitoring, the Journal of Engineering 2018 (2018) 1860–1864.

[139] J. Liang, Z. Ren, L. Wang, B. Tang, J. Liu, Y. Liu, Deep neural network in sequence to short sequence form for non-intrusive load monitoring, in: 2019 IEEE 3rd Conference on Energy Internet and Energy System Integration (EI2), IEEE, 2019, pp. 565–570.

[140] Z. Jia, L. Yang, Z. Zhang, H. Liu, F. Kong, Sequence to point learning based on bidirectional dilated residual network for non-intrusive load monitoring, International Journal of Electrical Power & Energy Systems 129 (2021) 106837.

[141] A. Harell, S. Makonin, I. V. Bajić, Wavenilm: A causal neural network for power disaggregation from the complex power signal, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 8335–8339.

[142] M. Kaselimi, E. Protopapadakis, A. Voulodimos, N. Doulamis, A. Doulamis, Multi-channel recurrent convolutional neural networks for energy disaggregation, IEEE Access 7 (2019) 81047–81056.

[143] M. Mottahedi, S. Asadi, Non-intrusive load monitoring using imaging time series and convolutional neural networks, in: 16th International Conference on computing in civil and building engineering, 2016, pp. 705–710.

[144] D. Yang, X. Gao, L. Kong, Y. Pang, B. Zhou, An event-driven convolutional neural architecture for non-intrusive load monitoring of residential appliance, IEEE Transactions on Consumer Electronics 66 (2020) 173–182.

[145] D. de Paiva Penha, A. R. G. Castro, Convolutional neural network applied to the identification of residential equipment in non-intrusive load monitoring systems, in: 3rd International Conference on Artificial Intelligence and Applications, 2017, pp. 11–21.

[146] P. Dash, K. Naik, A very deep one dimensional convolutional neural network (vdocnn) for appliance power signature classification, in: 2018 IEEE Electrical Power and Energy Conference (EPEC), IEEE, 2018, pp. 1–6.

[147] P. Davies, J. Dennis, J. Hansom, W. Martin, A. Stankevicius, L. Ward, Deep neural networks for appliance transient classification, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 8320–8324.

[148] A. Moradzadeh, B. Mohammadi-Ivatloo, M. Abapour, A. Anvari-Moghaddam, S. G. Farkoush, S.-B. Rhee, A practical solution based on convolutional neural network for non-intrusive load monitoring, Journal of Ambient Intelligence and Humanized Computing (2021) 1–15.

[149] W. Kong, Z. Y. Dong, B. Wang, J. Zhao, J. Huang, A practical solution for non-intrusive type ii load monitoring based on deep learning and post-processing, IEEE Transactions on Smart Grid 11 (2019) 148–160.

[150] L. Mauch, B. Yang, A new approach for supervised power disaggregation by using a deep recurrent lstm network, in: 2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP), IEEE, 2015, pp. 63–67.

[151] W. He, Y. Chai, An empirical study on energy disaggregation via deep learning, Advances in Intelligent Systems Research 133 (2016) 338–342.

[152] J. Kim, H. Kim, et al., Classification performance using gated recurrent unit recurrent neural network on energy disaggregation, in: 2016 international conference on machine learning and cybernetics (ICMLC), volume 1, IEEE, 2016, pp. 105–110.

[153] O. Krystalakos, C. Nalmpantis, 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.

[154] Y. T. Quek, W. L. Woo, T. Logenthiran, Load disaggregation using one-directional convolutional stacked long short-term memory recurrent neural network, IEEE Systems Journal 14 (2019) 1395–1404.

[155] S. Hosseini, N. Henao, S. Kelouwani, K. Agbossou, A. Cardenas, A study on markovian and deep learning based architectures for household appliance-level load modeling and recognition, in: 2019 IEEE 28th International Symposium on Industrial Electronics (ISIE), IEEE, 2019, pp. 35–40.

[156] T. Wang, T. Ji, M. Li, A new approach for supervised power disaggregation by using a denoising autoencoder and recurrent lstm network, in: 2019 IEEE 12th International Symposium on Diagnostics for Electrical Machines, Power Electronics and Drives (SDEMPED), IEEE, 2019, pp. 507–512.

[157] J. Kim, T.-T.-H. Le, H. Kim, Nonintrusive load monitoring based on advanced deep learning and novel signature, Computational intelligence and neuroscience 2017 (2017).

[158] J. Cho, Z. Hu, M. Sartipi, Non-intrusive a/c load disaggregation using deep learning, in: 2018 IEEE/PES Transmission and Distribution Conference and Exposition (T&D), IEEE, 2018, pp. 1–5.

[159] N. V. Linh, P. Arboleya, Deep learning application to non-intrusive load monitoring, in: 2019 IEEE Milan PowerTech, IEEE, 2019, pp. 1–5.

[160] R. Gopu, A. Gudimallam, N. Thokala, M. G. Chandra, On electrical load disaggregation using recurrent neural networks, in: Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, 2019, pp. 364–365.

[161] J. Wang, S. El Kababji, C. Graham, P. Srikantha, Ensemble-based deep learning model for non-intrusive load monitoring, in: 2019 IEEE Electrical Power and Energy Conference (EPEC), IEEE, 2019, pp. 1–6.

[162] M. Kaselimi, N. Doulamis, A. Doulamis, A. Voulodimos, E. Protopapadakis, Bayesian-optimized bidirectional lstm regression model for non-intrusive load monitoring, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 2747–2751.

[163] H. Rafiq, X. Shi, H. Zhang, H. Li, M. K. Ochani, A deep recurrent neural network for non-intrusive load monitoring based on multi-feature input space and post-processing, Energies 13 (2020) 2195.

[164] J. Song, H. Wang, M. Du, L. Peng, S. Zhang, G. Xu, Non-intrusive load identification method based on improved long short term memory network, Energies 14 (2021) 684.

[165] A. Harell, S. Makonin, I. V. Bajic, A recurrent neural network for multisensory non-intrusive load monitoring on a raspberry pi, in: IEEE MMSP, volume 18, 2018.

[166] H. Rafiq, H. Zhang, H. Li, M. K. Ochani, Regularized lstm based deep learning model: first step towards real-time non-intrusive load monitoring, in: 2018 IEEE International Conference on Smart Energy Grid Engineering (SEGE), IEEE, 2018, pp. 234–239.

[167] A. Yadav, A. Sinha, A. Saidi, C. Trinkl, W. Zörner, Nilm based energy disaggregation algorithm for dairy farms, in: Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring, 2020, pp. 16–19.

[168] R. Kukunuri, N. Batra, A. Pandey, R. Malakar, R. Kumar, O. Krystalakos, M. Zhong, P. Meira, O. Parson, Nilmtk-contrib: Towards reproducible state-of-the-art energy disaggregation, in: Proc. AI Social Good Workshop, 2020, pp. 1–5.

[169] M. Mobasher-Kashani, N. Noman, S. Chalup, Parallel lstm architectures for non-intrusive load monitoring in smart homes, in: 2020 IEEE Symposium Series on Computational Intelligence (SSCI), 2020, pp. 1272–1279. doi:10.1109/SSCI47803.2020.9308592.

[170] M. Kaselimi, N. Doulamis, A. Voulodimos, E. Protopapadakis, A. Doulamis, Context aware energy disaggregation using adaptive bidirectional lstm models, IEEE Transactions on Smart Grid 11 (2020) 3054–3067.

[171] M. Xia, K. Wang, W. Song, C. Chen, Y. Li, et al., Non-intrusive load disaggregation based on composite deep long short-term memory network, Expert Systems with Applications 160 (2020) 113669.

[172] F. C. C. Garcia, C. M. C. Creayla, E. Q. B. Macabebe, Development of an intelligent system for smart home energy disaggregation using stacked denoising autoencoders, Procedia Computer Science 105 (2017) 248–255.

[173] H. Chen, Y.-H. Wang, C.-H. Fan, A convolutional autoencoder-based approach with batch normalization for energy disaggregation, The Journal of Supercomputing 77 (2021) 2961–2978.

[174] A. Faustine, L. Pereira, H. Bousbiat, S. Kulkarni, Unet-nilm: A deep neural network for multi-tasks appliances state detection and power estimation in nilm, in: Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring, 2020, pp. 84–88.

[175] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL: http://arxiv.org/abs/1409.0473.

[176] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in

neural information processing systems, 2017, pp. 5998–6008.
[177] N. Lin, B. Zhou, G. Yang, S. Ma, Multi-head attention networks for non-intrusive load monitoring, in: 2020 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), IEEE, 2020, pp. 1–5.
[178] Z. Yue, C. R. Witzig, D. Jorde, H.-A. Jacobsen, Bert4nilm: A bidirectional transformer model for non-intrusive load monitoring, in: Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring, 2020, pp. 89–93.
[179] N. V. Gkalinikis, C. Nalmpantis, D. Vrakas, Attention in recurrent neural networks for energy disaggregation, in: International Conference on Discovery Science, Springer, 2020, pp. 551–565.
[180] V. Piccialli, A. M. Sudoso, Improving non-intrusive load disaggregation through an attention-based deep neural network, Energies 14 (2021) 847.
[181] I. Kamyshev, D. Kriukov, E. Gryazina, Cold: Concurrent loads disaggregator for non-intrusive load monitoring, arXiv preprint arXiv:2106.02352 (2021).
[182] N. Kitaev, Ł. Kaiser, A. Levskaya, Reformer: The efficient transformer, arXiv preprint arXiv:2001.04451 (2020).
[183] A. Katharopoulos, A. Vyas, N. Pappas, F. Fleuret, Transformers are rnns: Fast autoregressive transformers with linear attention, in: International Conference on Machine Learning, PMLR, 2020, pp. 5156–5165.
[184] Z.-H. Jiang, W. Yu, D. Zhou, Y. Chen, J. Feng, S. Yan, Convbert: Improving bert with span-based dynamic convolution, Advances in Neural Information Processing Systems 33 (2020).
[185] T. Sirojan, B. T. Phung, E. Ambikairajah, Deep neural network based energy disaggregation, in: 2018 IEEE International Conference on Smart Energy Grid Engineering (SEGE), IEEE, 2018, pp. 73–77.
[186] G. Bejarano, D. DeFazio, A. Ramesh, Deep latent generative models for energy disaggregation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 850–857.
[187] M. Kaselimi, A. Voulodimos, E. Protopapadakis, N. Doulamis, A. Doulamis, Energan: A generative adversarial network for energy disaggregation, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 1578–1582.
[188] M. Kaselimi, N. Doulamis, A. Voulodimos, A. Doulamis, E. Protopapadakis, Energan++: A generative adversarial gated recurrent network for robust energy disaggregation, IEEE Open Journal of Signal Processing 2 (2020) 1–16.
[189] Y. Pan, K. Liu, Z. Shen, X. Cai, Z. Jia, Sequence-to-subsequence learning with conditional gan for power disaggregation, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 3202–3206.
[190] A. M. Ahmed, Y. Zhang, F. Eliassen, Generative adversarial networks and transfer learning for non-intrusive load monitoring in smart grids, in: 2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), IEEE, 2020, pp. 1–7.
[191] D. Murray, L. Stankovic, V. Stankovic, S. Lulic, S. Sladojevic, Transferability of neural network approaches for low-rate energy disaggregation, in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 8330–8334. doi:10.1109/ICASSP.2019.8682486.
[192] N. Batra, R. Kukunuri, A. Pandey, R. Malakar, R. Kumar, O. Krystalakos, M. Zhong, P. Meira, 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.
[193] M. D'Incecco, S. Squartini, M. Zhong, Transfer learning for non-intrusive load monitoring, IEEE Transactions on Smart Grid 11 (2020) 1419–1429. doi:10.1109/TSG.2019.2938068.
[194] B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Artificial intelligence and statistics, PMLR, 2017, pp. 1273–1282.
[195] H. Pötter, S. Lee, D. Mossé, Towards privacy-preserving framework for non-intrusive load monitoring, in: Proceedings of the Twelfth ACM International Conference on Future Energy Systems, 2021, pp. 259–263.
[196] N. Hudson, M. J. Hossain, M. Hosseinzadeh, H. Khamfroush, M. Rahnamay-Naeini, N. Ghani, A framework for edge intelligent smart distribution grids via federated learning, in: 2021 International Conference on Computer Communications and Networks (ICCCN), 2021, pp. 1–9. doi:10.1109/ICCCN52240.2021.9522360.
[197] E. T. Mayhorn, G. P. Sullivan, J. M. Petersen, R. S. Butner, E. M. Johnson, Load disaggregation technologies: real world and laboratory performance, Pacific Northwest National Laboratory (PNNL), Richland, WA (US), Tech. Rep. PNNL-SA-116560 (2016).
[198] C. Klemenjak, S. Makonin, 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.
[199] M. W. Asres, L. Ardito, E. Patti, Computational cost analysis and data-driven predictive modeling of cloud-based online nilm algorithm, IEEE Transactions on Cloud Computing (2021).
[200] N. Batra, J. Kelly, O. Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, M. Srivastava, Nilmtk: An open source toolkit for non-intrusive load monitoring, in: Proceedings of the 5th international conference on Future energy systems, 2014, pp. 265–276.
[201] J. Kolter, S. Batra, A. Ng, Energy disaggregation via discriminative sparse coding, Advances in neural information processing systems 23 (2010) 1153–1161.
[202] M. Zhong, N. Goddard, C. Sutton, Signal aggregate constraints in additive factorial hmms, with application to energy disaggregation, Advances in Neural Information Processing Systems 27 (2014) 3590–3598.
[203] J. Liao, G. Elafoudi, L. Stankovic, V. Stankovic, Non-intrusive appliance load monitoring using low-resolution smart meter data, in: 2014 IEEE International Conference on Smart Grid Communications (SmartGridComm), IEEE, 2014, pp. 535–540.
[204] H. Altrabalsi, J. Liao, L. Stankovic, V. Stankovic, A low-complexity energy disaggregation method: Performance and robustness, in: 2014 IEEE Symposium on Computational Intelligence Applications in Smart Grid (CIASG), IEEE, 2014, pp. 1–8.
[205] M. Nguyen, S. Alshareef, A. Gilani, W. G. Morsi, A novel feature extraction and classification algorithm based on power components using single-point monitoring for nilm, in: 2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE), IEEE, 2015, pp. 37–40.
[206] S. Alshareef, W. G. Morsi, Application of wavelet-based ensemble tree classifier for non-intrusive load monitoring, in: 2015 IEEE Electrical Power and Energy Conference (EPEC), IEEE, 2015, pp. 397–401.
[207] R. Bonfigli, M. Severini, S. Squartini, M. Fagiani, F. Piazza, Improving the performance of the afamap algorithm for non-intrusive load monitoring, in: 2016 IEEE Congress on Evolutionary Computation (CEC), IEEE, 2016, pp. 303–310.
[208] M. Z. A. Bhotto, S. Makonin, I. V. Bajić, Load disaggregation based on aided linear integer programming, IEEE Transactions on Circuits and Systems II: Express Briefs 64 (2016) 792–796.
[209] S. M. Tabatabaei, S. Dick, W. Xu, Toward non-intrusive load monitoring via multi-label classification, IEEE Transactions on Smart Grid 8 (2016) 26–40.
[210] N. Batra, H. Wang, A. Singh, K. Whitehouse, Matrix factorisation for scalable energy breakdown, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 31, 2017.
[211] A. K. Jain, S. S. Ahmed, P. Sundaramoorthy, R. Thiruvengadam, V. Vijayaraghavan, Current peak based device classification in nilm on a low-cost embedded platform using extra-trees, in: 2017 IEEE MIT Undergraduate Research Technology Conference (URTC), IEEE, 2017, pp. 1–4.
[212] N. Batra, Y. Jia, H. Wang, K. Whitehouse, Transferring decomposed tensors for scalable energy breakdown across regions, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[213] X. Shi, H. Ming, S. Shakkottai, L. Xie, J. Yao, Nonintrusive load monitoring in residential households with low-resolution data, Applied Energy 252 (2019) 113283.
[214] Q. Yuan, H. Wang, B. Wu, Y. Song, H. Wang, A fusion load disaggregation method based on clustering algorithm and support vector regression optimization for low sampling data, Future Internet 11 (2019) 51.
[215] C. Puente, R. Palacios, Y. González-Arechavala, E. F. Sánchez-Úbeda, Non-intrusive load monitoring (nilm) for energy disaggregation using soft computing techniques, Energies 13 (2020) 3117.

[216] L. Mauch, B. Yang, A novel dnn-hmm-based approach for extracting single loads from aggregate power signals, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, pp. 2384–2388.

[217] F.-Y. Chang, C. Chen, S.-D. Lin, An empirical study of ladder network and multitask learning on energy disaggregation in taiwan, in: 2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI), IEEE, 2018, pp. 86–89.

[218] M. Xia, K. Wang, X. Zhang, Y. Xu, et al., Non-intrusive load disaggregation based on deep dilated residual network, Electric Power Systems Research 170 (2019) 277–285.

[219] M. Xia, Y. Xu, K. Wang, X. Zhang, et al., Dilated residual attention network for load disaggregation, Neural Computing and Applications 31 (2019) 8931–8953.

[220] M. Ayub, E.-S. M., Multi-Target Energy Disaggregation using Convolutional Neural Networks, International Journal of Advanced Computer Science and Applications 11 (2020). URL: http://thesai.org/Publications/ViewPaper?Volume=11&Issue=10&Code=IJACSA&SerialNo=85. doi:10.14569/IJACSA.2020.0111085.

[221] D. García-Pérez, D. Pérez-López, I. Díaz-Blanco, A. González-Muñiz, M. Domínguez-González, A. A. C. Vega, Fully-convolutional denoising auto-encoders for nilm in large non-residential buildings, IEEE Transactions on Smart Grid 12 (2020) 2722–2731.

[222] F. Ciancetta, G. Bucci, E. Fiorucci, S. Mari, A. Fioravanti, A new convolutional neural network-based system for nilm applications, IEEE Transactions on Instrumentation and Measurement 70 (2020) 1–12.

[223] L. Massidda, M. Marrocu, S. Manca, Non-intrusive load disaggregation via a fully convolutional neural network: improving the accuracy on unseen household, in: 2020 2nd IEEE International Conference on Industrial Electronics for Sustainable Energy Systems (IESES), volume 1, IEEE, 2020, pp. 317–322.

[224] Y. Zhang, B. Yin, Y. Cong, Z. Du, Multi-state household appliance identification based on convolutional neural networks and clustering, Energies 13 (2020) 792.

[225] R. Kukunuri, A. Aglawe, J. Chauhan, K. Bhagtani, R. Patil, S. Walia, N. Batra, Edgenilm: towards nilm on edge devices, in: Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, 2020, pp. 90–99.

[226] G. Zhou, Z. Li, M. Fu, Y. Feng, X. Wang, C. Huang, Sequence-to-sequence load disaggregation using multiscale residual neural network, IEEE Transactions on Instrumentation and Measurement 70 (2020) 1–10.

[227] J. Jiang, Q. Kong, M. D. Plumbley, N. Gilbert, M. Hoogendoorn, D. M. Roijers, Deep learning-based energy disaggregation and on/off detection of household appliances, ACM Transactions on Knowledge Discovery from Data (TKDD) 15 (2021) 1–21.

[228] H. Çimen, N. Çetinkaya, J. C. Vasquez, J. M. Guerrero, A microgrid energy management system based on non-intrusive load monitoring via multitask learning, IEEE Transactions on Smart Grid 12 (2020) 977–987.

[229] L. Huang, S. Chen, Z. Ling, Y. Cui, Q. Wang, Non-invasive load identification based on lstm-bp neural network, Energy Reports 7 (2021) 485–492.

[230] C. Athanasiadis, D. Doukas, T. Papadopoulos, A. Chrysopoulos, A scalable real-time non-intrusive load monitoring system for the estimation of household appliance power consumption, Energies 14 (2021) 767.

[231] G. Bucci, F. Ciancetta, E. Fiorucci, S. Mari, A. Fioravanti, Multi-state appliances identification through a nilm system based on convolutional neural network, in: 2021 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), IEEE, 2021, pp. 1–6.

[232] L. de Diego-Otón, D. Fuentes-Jimenez, Á. Hernández, R. Nieto, Recurrent lstm architecture for appliance identification in non-intrusive load monitoring, in: 2021 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), IEEE, 2021, pp. 1–6.

[233] D. Jia, Y. Li, Z. Du, J. Xu, B. Yin, Non-intrusive load identification using reconstructed voltage–current images, IEEE Access 9 (2021) 77349–77358.

[234] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to sequence learning with neural networks, in: Advances in neural information processing systems, 2014, pp. 3104–3112.

[235] C.-Y. Hsu, A. Zeitoun, G.-H. Lee, D. Katabi, T. Jaakkola, Self-supervised learning of appliance usage, in: International Conference on Learning Representations, 2019.

[236] G. Herath, T. Thilakanayake, M. Liyanage, C. Angammana, Comprehensive analysis of convolutional neural network models for non-instructive load monitoring, in: 2020 International Conference and Utility Exhibition on Energy, Environment and Climate Change (ICUE), IEEE, 2020, pp. 1–11.

[237] H. Ren, F. M. Bianchi, J. Li, R. L. Olsen, R. Jenssen, S. N. Anfinsen, Towards applicability: A comparative study on non-intrusive load monitoring algorithms, in: 2021 IEEE International Conference on Consumer Electronics (ICCE), IEEE, 2021, pp. 1–5.