Contribution of Non-Intrusive Load Monitoring to Home Energy Management Systems

Yvon Francou

Doctor of Philosophy

Contribution of Non-Intrusive Load Monitoring to Home Energy Management Systems

Johann FRANCOU

Jury:

Pr. Gilles NOTTON, Université de Corse,
Dr. Manar AMAYRI, Université Concordia
Pr. Stéphane PLOIX, INP Grenoble
Pr. Kaushik DAS, Technical University of Denmark
Pr. Philippe LAURET, Université de La Réunion
Dr. Mathieu DAVID, Université de La Réunion
Dr. Calogine DIDIER, Université de La Réunion
Dr. Oanh CHAU, Université de La Réunion

b

Remerciements

Ma pleine reconnaissance va à ma direction de thèse ; Philippe Lauret et Mathieu David sans qui ces travaux de thèse n'auraient pas eu lieu. Mes pensées vont vers toi Philippe ; tu m'as accordé ta confiance, mais surtout ton temps sur ce projet de thèse et je t'en remercie. Mathieu, je te remercie pour tes précieux conseils, pour le temps passé à travailler ensemble et tes partages de connaissance. Je remercie mon encadrement de thèse, Oanh Chau et Didier Calogine. Didier, merci de ta confiance sans faille et de m'avoir ouvert les portes du monde passionnant de la recherche.

En espérant de fructueuses futures collaborations avec vous.

Je remercie les membres du jury d'avoir accepté d'évaluer cette thèse, Manar Amayri, Gilles Notton, Stéphane Ploix et Kaushik Das.

Sans oublier le comité de thèse, constitué de Manar Amayri, Laurent Bridier, Ghjuvan-Antone Faggianelli et de Harry Boyer. Je vous remercie de votre disponibilité et de vos conseils lors des échanges fructueux de nos précédentes réunions.

Bien que certains d'entre eux n'aient pas contribué directement à ce travail de recherche, je remercie les professeurs de l'Ecole Supérieure d'Ingénieurs Réunion Océan Indien (ESIROI) ; la formation de qualité prodiguée et les valeurs humaines transmises lors de ma période à l'ESIROI ont été précieuses dans l'accomplissement de ces années d'études.

Un grand merci aux collègues et amies/amis de l'Université de La Réunion pour les grandes et petites discussions, les débats, les partages, les encouragements, les collaborations, le foot, les parties d'échecs, tout simplement merci d'avoir été là ; Cédric, Delphine, Tanika, Christian, Karine, Malik, Emeric, Michael, Josselin, Jonathan, Mirhado, Faly, Vanessa, Jérôme, Lionel, Jordy, Toky, Youssouf, Margot, Gabriel, Tristan, Manitra, Arena, Philippe et toute « la bande à Gaël ».

Si j'ai persévéré dans ce long et difficile travail de thèse c'est grâce à mon entourage proche :

Je suis reconnaissant envers mes chers amis, Raoul, Mahafaly, Yann, Noom, portez-vous bien !

Un grand merci aux oncles et tantes, Cédric, Fanny, Gary, Tita & co, Dany & co, Mireille & co, Rondro & co, Naina & co, Erika, Nathalie, Alain & co, Ando & co ; merci de votre partage d'amour qui persiste depuis ma plus tendre enfance.

Je remercie les grands parents, pour leur amour, leurs conseils et leurs prières : Bebe Loty & co, Bebe Line & co, Tonton Roger & co, Dabe Jojo & co, Bebe Lili & co, Dadabe Emile & co, Bebe (Marce)line et Dadabe Robert, Bebe Ramanana. Et une reconnaissance particulière envers ma chère grand-mère Bebe Lala.

À mes cousins, Raphaël et Manon, merci de rendre les dimanches en famille aussi joviales et solaires.

Merci à Tanjona, Morgane et Tatamo pour ces moments de détentes et de rigolades autour des bons jeux de sociétés de Momo, et je l'espère, reprendront de plus belle.

Je remercie mes beaux-parents pour leur soutien infaillible et leur amour.

Avec beaucoup de nostalgie, je remercie la fratrie ; Jonathan, Matthieu, Prisca. Je ne cesse d'espérer ce moment où nous serons tous à nouveau réunis, comme au bon vieux temps. En attendant portez-vous bien.

À mes chers parents, vous avez tout donné pour que nous puissions nous en sortir dans la vie, malgré les difficultés. 1000 mercis ne suffiraient pas pour tous vos sacrifices et pour toutes vos prières. Je vous dois toutes ces années d'études et pas seulement. J'espère qu'à travers ces quelques lignes vous apercevrez une bribe de ma profonde reconnaissance et mon amour.

À mon épouse, Tantely. Sache-le, ton amour et tes encouragements m'ont fait me surpasser pour la réalisation de ce travail. C'est ému que je te le partage ; je suis profondément reconnaissant de t'avoir eu auprès de moi pendant ces 10 dernières années. Je t'aime Tantely.

À mon fils, Timothée ; tu fais le bonheur de tes parents et tu es notre plus grande bénédiction.

Résumé

La crise énergétique actuelle a mis en lumière nos profondes vulnérabilités en matière d'approvisionnement énergétique. Cette crise contribue à renforcer et même à exacerber la dépendance de beaucoup de pays aux énergies fossiles. De nombreux gouvernements ont choisi d'affermir leurs politiques énergétiques, particulièrement sur les progrès que sont la sobriété énergétique et l'usage des énergies renouvelables intermittentes (photovoltaïques et éoliennes). L'énergie photovoltaïque a l'avantage de n'émettre que peu de gaz à effet de serre lors de son fonctionnement mais a le défaut majeur d'être intermittent et variable. Le stockage énergétique peut pallier la variabilité et l'intermittence de l'énergie solaire, cependant les technologies de stockage les plus performantes s'accompagnent de problématiques de coût financier et de recyclage, freinant ainsi leur popularisation.

L'utilisation de Systèmes de Management de l'Énergie pour le Résidentiel (SMER) (ou Home energy management system - HEMS en anglais) vient en complément des énergies solaires en proposant une manière plus intelligente de consommer l'énergie. Développer des SMER est un sujet de recherche très étudié mais leur application dans le quotidien reste très limitée.

De manière pratique, les SMERs s'appuient souvent sur un réseau de capteurs permettant d'obtenir des mesures individuelles d'appareil, afin d'identifier les comportements de consommation ; les durées, les préférences horaires ou aussi l'énergie consommée. Cependant, on peut s'interroger sur la complexité d'installation de réseaux de capteur, la gestion des données, ou aussi sur le coût potentiellement important.

La désagrégation non-intrusive (ou NILM en anglais) est le procédé qui consiste à séparer une mesure générale en signaux d'appareils à l'aide d'algorithmes. Le but premier du NILM est de réduire les coûts associés à l'installation et à la maintenance du réseau de capteurs car le NILM n'a besoin que d'un seul point de mesure placé au niveau du compteur général. A noter que tout au long du manuscrit, seules des simulations de NILM à basse fréquence d'échantillonnage (1 donnée par minute) sont réalisées. Les méthodes utilisant des modèles d'apprentissage profond sont les plus prometteuses, en raison de leurs très bonnes performances. Dans ce travail, nous

expérimentons ce type de modèle et nous constatons l'importance de la diversité des données dans la base d'apprentissage de ces modèles. En même temps, la littérature associée souligne le manque de données publiées et fiables, pour avoir une solution de NILM généralisée, c'est-à-dire efficace pour n'importe quelle maison et pour n'importe quelle saison.

Ce travail propose une nouvelle méthode d'augmentation de données dans le but d'enrichir les bases d'apprentissage, et de se rapprocher de cet objectif de généralisation. La méthode est rigoureusement expérimentée et son apport sur la performance des modèles est démontré. La nouvelle technique d'augmentation de données est ajoutée à l'outil NILMTK, un célèbre outil « open source » dédié à l'évaluation et la comparaison de méthodes de NILM.

Enfin, ce travail de thèse accorde une importance particulière à l'application de ces modèles de NILM en proposant d'évaluer quantitativement l'apport du NILM à un SMER. Le cas d'étude consiste à appliquer une planification de charge à des maisons réelles tirées de données publiques. On se propose d'intégrer un modèle de NILM dans un SMER. Les planifications produites par le SMER avec un NILM intégré, se montrent plus adaptées aux usages habituels de chaque foyer, par conséquent l'ajout de cette méthode permettrait une meilleure acceptabilité des SMER.

En résumé, ce travail apporte deux contributions originales majeures. Premièrement le développement d'une méthode d'augmentation de données pour aider à la généralisation des modèles de NILM. La méthode a été publiée dans le journal SEGAN (Sustainable Energy, Grids and Network). En deuxième lieu, la thèse propose une approche originale pour évaluer quantitativement la contribution du NILM aux SMER.

Abstract

The current need for a fast decarbonisation of energy production has revealed deep vulnerabilities in energy supplies. It also emphasised the strong fossil fuel dependency of many countries. Some governments decided to tighten their energy policies, particularly on energy efficiency and intermittent renewable energies such as solar and wind. Photovoltaic energy has the advantage of emitting few greenhouse gases in its lifetime operation but has the drawback of being intermittent and variable.
Energy storage can mitigate the variability and intermittency of solar energy. However, energy storage's substantial environmental and financial costs may refrain from investing massively. A second way to face renewable drawbacks involves more intelligent demand management. Developing a Home Energy Management System (HEMS) for residential areas has been a growing research trend but has not been widely applied due to technical, social and financial barriers. It is expected to monitor appliances individually in residential areas to understand the user's behaviours accurately and consequently to determine the preferred hour of use of each appliance, the duration, or the energy demanded. However, Appliance Load Monitoring (ALM) is challenging due to the high complexity and cost of the sensor network involved.
Non-Intrusive Load Monitoring (NILM) is the process of disaggregating the main load into individual appliance loads without additional energy meters. NILM aims to mitigate the cost of sensor installation and maintenance by installing a unique sensor on the main load and retrieving the individual load appliances computationally. Recently, deep learning models have been up-and-coming for NILM tasks. The main meter sampling rate is crucial for the trade-off between disaggregation accuracy and cost efficiency. High-frequency data will carry more detailed signatures but with a costly sensor.
In this work, we assess the generalisation capabilities of low-frequency NILM, which use 1-min time step load monitoring, and identify the settings having the most significant impact on performances. This work demonstrates the importance of diversity in the training dataset to unlock generalisability. At the same time, the literature underlines data scarcity in the domain, mainly due to the complex task of recording reliable supervised data in real conditions. The present work provides a new data

augmentation technique to address this issue by enriching the training set. The technique is thoroughly experimented with to prove its efficiency in improving generalisability. The developed data augmentation add-on module is implemented inside the NILMTK framework, a famous toolkit to evaluate NILM solutions.
This manuscript brought particular care to a practical NILM application. The idea is to assess the contribution of a low-frequency NILM to a HEMS by conducting simulations. The case study applies a load scheduling HEMS to real houses from public databases. The schedules given by the HEMS with NILM are more acceptable for the end-user. The first original contribution of this work relies on developing a new and straightforward data augmentation technique published in the SEGAN (Sustainable Energy, Grids and Network) journal. Secondly, the thesis proposes an original approach to quantitatively evaluating NILM's contribution to HEMS.

Abbreviations

ALM	Appliance Load Monitoring
CNN	Convolutional Neural Network
CO	Combinatorial Optimisation
DAE	Denoising Auto Encoder
DL	Deep-Learning
DLC	Direct Load Control
DR	Demand Response
DSM	Demand Side Management
EMS	Energy Management System
HEMS	Home Energy Management System
Hgrid	A naive HEMS generating random a schedule at off-peak times
Hilm	HEMS with an ILM
HMI	Human Machine Interface
Hnoaug	HEMS with a non-augmented NILM
Hoffsetaug	HEMS with a NILM augmented with the method OFFSETAUG
Hrandom	A naive HEMS generating a random schedule
ILM	Intrusive Load Monitoring
lr	learning rate
ML	Machine-Learning

NILM	Non-Intrusive Load Monitoring
pat	Patience
S2P	Sequence to Point
S2S	Sequence to Sequence

Contents

List of Figures

Introduction

The current need in a fast decarbonisation of the energy production highlights the significant vulnerability of the high fossil-fuel dependency of society [2,3]. Oil is needed for the industry, transportation and electricity production. Without a resilient energy supply chain, the international economy would be severely impacted [4]. Besides, the human use of fossil fuels has been a significant contributor to global warming due to greenhouse gas emissions. Causal relationships between climate and fossil fuels have been demonstrated in recent IPCC reports. [5]. The effects of climate change are disastrous, leading to more frequent heatwaves and extreme weather conditions in some areas. Not only humans but all biodiversity is at risk [6]. Biodiversity is facing a severe threat due to climate change, with potentially devastating consequences. According to the IPCC [5], global temperature rise could increase the extinction risk of up to 30% of Earth's species. To alleviate the impact of climate change, there are several levers to activate, such as instituting profound modifications in the economy in favouring a circular economy, supporting sustainable and local agriculture, developing renewable energy and reducing the consumption of resources per capita. The development of renewable energy sources has been growing remarkably in recent years. This has been driven by the need to alleviate climate change and reduce our reliance on fossil fuels. According to [7], the capacity of renewable energy has expanded rapidly. In 2020, an estimated 260 GW of renewable power capacity were added worldwide. This expansion has been made possible by significant reductions in the cost of renewable technologies, which have made them increasingly competitive. One of the most significant barriers is the variable nature of specific renewables, such as solar and wind power, which can lead to issues for grid integration. The report in [8] suggests significant investments in energy storage technologies (particularly in electrochemical batteries) and the development of smart grids. A smart grid is an advanced electrical system that efficiently manages the generation, distribution, and consumption of electricity using modern information technology [9]. The primary goals of a smart grid are to enhance grid reliability, optimise energy distribution, reduce energy losses, and facilitate the integration of renewable energy sources. Demand Side Management (DSM) is a lever for accommodating renewable energy re-

sources. It consists of strategies that encompass various demand-oriented approaches, such as load shifting or load shedding, and energy efficiency measures, that empower consumers to adapt their electricity consumption patterns with renewable energy generation, reduce their energy bill or more generally optimise the demand profile shape. DSM applies to commercial, industrial and residential [10]. In residential areas, basic DSM begins with a straightforward visualisation of the load given to the dweller. A consumption reduction of 15% is observed in [11], only by providing energy feedbacks. In the literature, developing a more sophisticated DSM tool for residential has been addressed. A Home Energy Management System (HEMS) is an automated system used to control and schedule specific smart appliances [12]. Individual appliance loads are potential inputs for HEMSs because they allow an isolated visualisation of how exactly the device is used, at what time, and at what power amplitude. However, individual demand monitoring is responsible for additional costs from sensor device purchasing, installation, and data management. Monitoring every electrical appliance in a building is challenging, intrusive, and non-robust. In particular, for residential settings, intrusive and invasive monitoring is likely to generate reluctance among residents. Non-Intrusive Load Monitoring (NILM) aims to overcome those barriers by computationally disaggregating the main load measurement into appliance-level loads. NILM has been introduced in works [13, 14] with an algorithm designed to detect on/off events. Those transient sequences are relatively short and require high-frequency data to be discriminated. However, the higher the sampling rate, the higher the monitoring and data processing cost. That is why research on unlocking low-frequency NILM has emerged [15, 16]. Deep Learning (DL) methods have achieved promising results on low-frequency data [17–23]. The first chapter of this work is focused on NILM experimentation to identify the main parameters that impact NILM results. The experiments are conducted on houses from the public dataset REFIT [24]. However, DL needs a high amount of data contrasting with the known lack of supervised data in the realm of NILM [25]. In response to both issues, some researchers in NILM recently proposed Data Augmentation (DA) methods for NILM [21, 25–28]. DA is a technique to enrich the training set in generating a synthetic dataset, habitually based on a transformation of the existing dataset. In this manuscript, I propose a DA technique called OFFSETAUG, which was published in the Sustainable Energy, Grids and Networks (SEGAN) journal [29]. The idea behind this augmentation technique is to generate a synthetic dataset by creating new power profiles with a particular offset process. Compared to the number of papers aiming to enhance NILM performances [13, 17, 18, 21, 22, 30] few authors tackle the application of NILM [31–33]. In Chapter 4, a simulation framework is proposed to assess the contribution of NILM to HEMSs.

This whole work revolves around the following points:

- Sensitivity analysis of NILM to identify the most influential parameters.
- Assessing the performances of NILM for different configurations, particularly when enhanced with the OFFSETAUG technique.
- Evaluating if NILM, and particularly NILM with OFFSETAUG, can improve HEMSs.

CHAPTER 1

Non-Intrusive Load Monitoring (NILM)

In residential areas, consumption behaviour is different from one house to another, depending on the number of people living in the house, the type and number of electrical appliances, and weather conditions. To understand and take action to optimise energy consumption, load monitoring is substantially needed. Load monitoring measures the load in an electrical system [34]. A modern house is an electrical system with electrical wiring to ensure energy distribution, an electrical supply generally from the grid, and outlets and switches for energy consumption. There are two main branches of Appliance Load Monitoring (ALM): Intrusive Load Monitoring (ILM) and Non-Intrusive Load Monitoring (NILM). In this chapter, both approaches are defined and compared qualitatively. A deeper analysis is carried out on NILM to understand the concepts and challenges around this technology. This chapter will be focused on answering the following questions:

- What makes NILM an important research theme?
- What are the main NILM solutions?

Note that this chapter is not wanted to be exhaustive but instead gives the necessary notions to understand future chapters.

1.1 Appliance Load Monitoring

ALM refers to the process of monitoring individual appliances. Following the recent energy crisis, ALM has gained significant interest due to its potential for energy savings. In the residential sector, ALM provides the household with energy feedback, allowing the dwellers to regulate their energy demand and identify causes of

energy wastage. With ALM, inhabitants can decompose their monthly electricity bill to spot the high consumption devices. Malfunctions can occur during appliance operation [35, 36]. For instance, a refrigerator needing to defrost consumes more due to higher compressor work. Appliance operating flaws are identifiable on ALM measurements and can be corrected by fixing or replacing the device. ALM is not only beneficial for the consumer electricity bill, but the process can also help the grid operator for energy management systems (EMS) implementation such as Demand Response (DR) or Direct Load Control (DLC) [31–33]. An accurate and recent estimate of flexible load levels is a significant resource for grid management. The advantages of ALM are not debatable. However, its real-life application still faces barriers such as intrusiveness and cost-effectiveness concerns for ILM approaches while NILM applications encounter accuracy and transferability issues.

1.2 Intrusive Load Monitoring

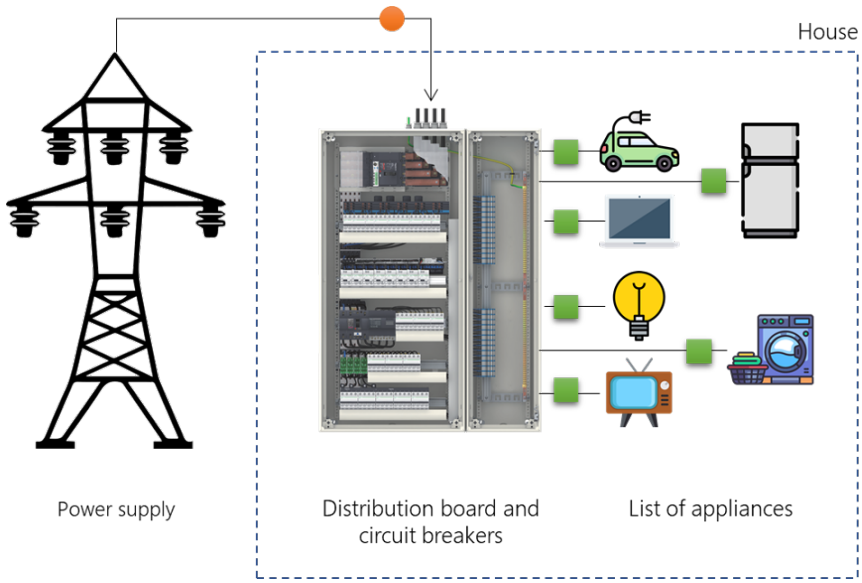


Figure 1.1. Simplified home electrical system diagram, the orange circle represents the metering point for NILM while green squares represent the metering points in a standard ILM-3 approach

.

ILM has been thoroughly reviewed in the works [34, 37, 38].ILM involves the measurement of electricity consumption for one or a few appliances, generally using inexpensive metering devices. The term "intrusive" indicates that the meter is placed within the living space, usually in close proximity to the monitored appliances. The Figure 1.1 gives an overview of the sensor positions in NILM and ILM cases. There are three levels of ILM, described in [37], the ILM-3 level corresponds to the case where the sensor is directly embedded in the appliance itself or installed in the dedicated outlet to individually monitor the appliance. In ILM-2, the sensors are positioned at the plug level, allowing for direct monitoring of the appliances connected to the outlet or multi-outlet, hence each appliance is not necessarily monitored individually. Finally, in ILM-1 each sensor is positioned inside the electrical distribution board to monitor a zone of the house, thus a group of appliance inside the same room. ILM-1 and ILM-2 configurations allow to monitor group of appliances instead of individual appliances, therefore an accuracy drop is expected when it comes to identify the appliance compared to ILM-3 level. Although ILM-3 gives an accurate solution for ALM, the important number of required sensors increases the overall monitoring cost. To achieve cost-effectiveness in ALM applications, a lot of researchers have been focusing on NILM.

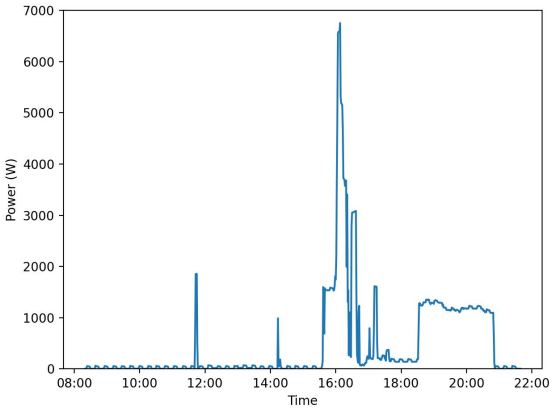
1.3 Non-Intrusive Load Monitoring

NILM is the process of disaggregating the main load into individual appliance loads. The major advantage of NILM is to require a unique measurement spot that can be taken outside the house, as represented in Figure 1.1. This subsection is dedicated to a thorough understanding of the concept of NILM and the challenges involved.

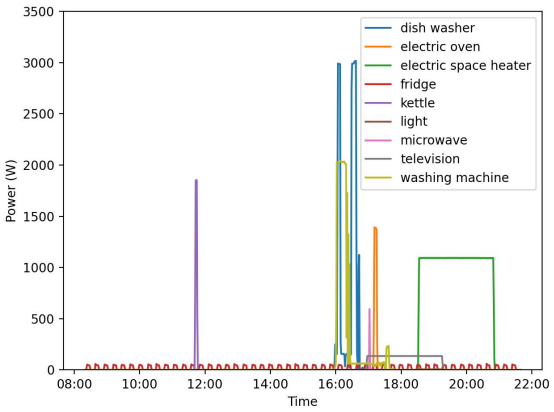
1.3.1 Mathematical formulation

NILM consists in disaggregating a main load measurement, so a time series, let it be $Y_{1:T} = (y_1, y_2, ..., y_T)$ into appliance-level loads such as $X^{(k)}_{1:T} = (x^{(k)}_1, x^{(k)}_2, ..., x^{(k)}_T)$ for the appliance k. Figure 1.2 depicts the disaggregation of the main load. $x^{(k)}_t$ and y_t are power values in Watt. At each timestamp, $t \in [1, T]$, the aggregated signal can be defined by the equation,

$$y_t = \sum_{k=1}^{K} x^{(k)}_t + \epsilon_t. \tag{1.1}$$



(a) Main load Y



(b) Disaggregated load X

Figure 1.2. Observation of a main load and the associated individual loads on the SYND dataset [1]

While (1.2) defines the principle of NILM,

$$\hat{x_t}^{(k)} = f^{(k)}(y_t). \tag{1.2}$$

NILM models, noted $f^{(k)}$, aim at estimating $\hat{x_t}^{(k)}$ values knowing only the main load time series Y. K is the set of all measured appliances and ϵ_t is a noise value. There may be a difference between the main measurement and the sum of individual load appliances, reflected in the noise value. This difference may occur for three potential reasons: not all appliances are necessarily monitored, the submeter and the main sensors may have different sensitivities, or due to minor power fluctuations. Deep

learning models are widespread for NILM due to their promising performances. To train $f^{(k)}$ to identify the load of an appliance k, the supervised training is processed on a dataset where Y and $X^{(k)}$ are fully known. To be consistent in mathematical notation, even if it may be counter-intuitive to predict x with y features, this convention is kept in the whole manuscript.

1.3.2 NILM, a challenging task

The first NILM research began in the 1980s; however, the technology is not popularly used because of the challenges that must be alleviated before a widespread deployment. Main difficulties can be grouped as technical or societal issues.

1.3.2.1 Technically challenging

For disaggregation purposes, a NILM model must have abilities to identify individual appliance activation in a generally complex aggregated load. Several sources of complexity can be pointed out. First, the appliances composing the aggregated load are rarely the same between two houses. Even if certain appliances are typical for residential such as washing machines, fridges, televisions or freezers, less common devices might be retrieved, for instance, aquariums, particular game consoles or DIY tools. Combinations of the fridge and washing machine operations are more ordinary than washing machine and jigsaw. Hence, the consequent diversity of possible appliance combinations adds high complexity to the task of NILM. Secondly, there are uncountable existing appliance brands and versions; finding the same device brand in two distinct houses would be more surprising than having different ones. The appliance's core electrical and mechanical components are relatively common; each standard washing machine has a drum for mixing up the clothes, a rotor to drive the drum mechanically, and a thermal resistance for water heating. However, all the brands do not agree on a defined operating power level or a standard operating duration. Dishwashers, microwaves, or washing machines increase aggregate load diversity as they operate depending on a user-defined program. For instance, dishwashers can run in eco mode or a quick-washing program. For both modes, the warm-up periods and the temperature levels are different; thus, the resulting electric signature is different.

Table 1.1. Types of Appliances

Type	Description	Appliances
I	Only on and off states	Microwaves, Kettles, Toasters
II	Multi-state appliances	Washing machines, Dishwashers
III	Continuously variable devices	Power drill, Dimmer lights
IV	Always-on devices with very low consumption	Appliances on standby mode, Routers

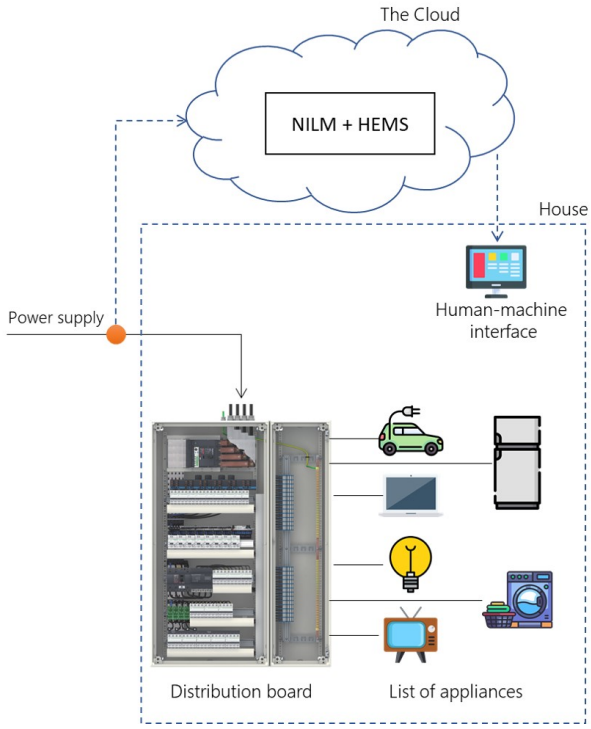
1.3.2.2 Socially challenging

NILM encompasses the social burdens of ALM, be they privacy and acceptance issues. Broken-down monitoring discloses the dweller's activities, which is ethically questionable and requests a trust-based relationship between the monitoring manager and the residents [39, 40]. Observing the profile of an appliance requiring home occupation to function (Television, microwave, oven ..., etc.) indicates when the residents are at home. Surprisingly, according to authors [41], privacy concerns are not the main obstacle to smart-metering roll-out in residential. Instead, installing metering devices and managing data are more challenging due to their high complexity. Ultimately, social burdens meet the technical aspects when considering operational NILM implementations. Figure 1.3 depicts two schemes:

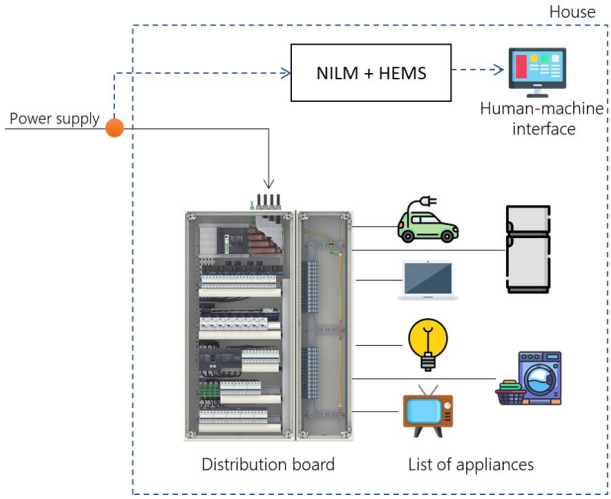
- cloud-based (Figure 1.3(a)), where the aggregate data are sent to the internet for remote processing, be they disaggregation and HEMS (Home Energy Management Sysyem) purposes;
- edge-based (Figure 1.3(b)), in which the data does not physically leave the household. In this configuration, all computing tasks are processed locally.

The cloud-based approach emphasises social vulnerabilities as private data is uploaded online. However, this deployment method allows for the utilisation of higher shared computing resources, thus more complex DL models, compared to edge-based HEMS. The latter approach keeps the data within the household and ensures security at the expense of low computational power available for cost-effectiveness concerns.

Even though some NILM hurdles remain, the technology is a step forward from traditional ILM solutions due to the reduced number of meters installed intrusively.



(a) Cloud-based



(b) Edge-based

Figure 1.3. Main NILM implementations

The use of traditional smart meters has been called socially into question, as noted in a study by [40], due to their low acceptance level. NILM is seen as a potential solution to this ongoing social debate.

1.3.3 Non-ML-based methods

NILM has been a vigorously explored topic. Many authors, [13, 14, 18–21, 30, 42], focused on NILM model development to unlock generalisation capability permitting large-scale deployment. NILM models can be roughly separated into two groups: machine-learning and non-machine-learning-based models. This subsection gives an overview of primary non-ML-based NILM methods to understand the concepts and the pros and cons of the main techniques.

1.3.3.1 Hart method - Edge detection

George W. Hart [13] is the precursor of NILM. To understand the pioneer model, it is worth underlining load profiles that can be characterised in two different states, steady and transient periods. A given appliance can be in several states; for instance, the washing machine can be in washing, water heating or drying phases. Those operating stages induce different patterns on the load profile. Simpler appliances, such as traditional lighting, only have two states, on and off. A state change is a transient phase while non-variant power profiles are associated with steady states. Hart proposes identifying both states on the aggregated load using an edge detection process. An edge refers to a power shift, increasing or decreasing power. Figure 1.4 represents the spotted edges on a subset of the REFIT dataset. Thus, unmarked periods are the steady state moments. The idea of Hart's method is to associate positive and negative edges by comparing the power shift amplitudes. For instance, a 100W positive power shift followed by a 98W negative power shift can be sensibly paired and are potentially caused by the same appliance operation. Figure 1.4 shows that the positive edge noted E9 and negative edge E10 will be associated, similarly for E3 and E4. One problem remains, how to map an appliance operation to the appliance label? Concretely, how can the approach determine if the appliance operation defined by a 100W-positive-shift and a 98W-negative-shift is a fridge or a washing machine operation? Hart separates two setups, the manual and the automatic setups. During manual mode, there is a temporary period where the appliances in question are manually turned on and off to gather data on their power levels and how long they operate for. In automatic setup, typical power level and operation

duration are beforehand determined for each appliance to form a device distinctive map. The aggregate load from Figure 1.4 broadcasts a relatively smooth aggregate load, probably because of the chosen hours, from 00:00 to 03:00, when occupants sleep. However, in most cases, the aggregate load is very noisy. Furthermore, some appliances do not have a clear steady state phase.

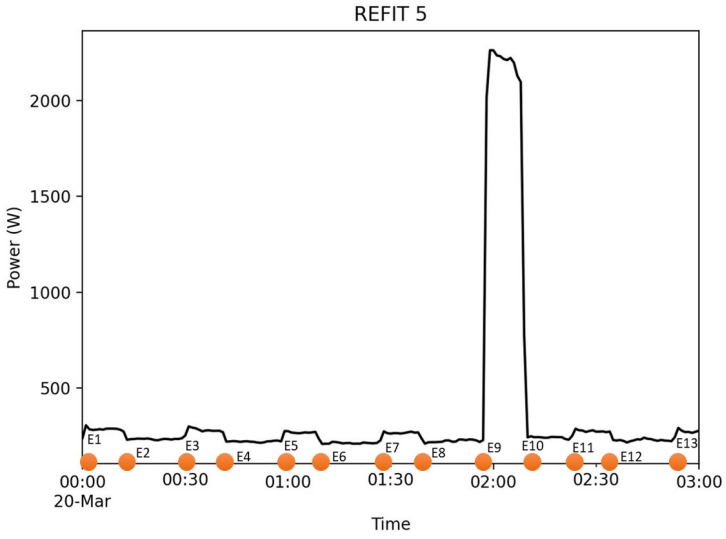


Figure 1.4. Edge detection on house 5 from REFIT dataset, orange circles represent the power shift periods

1.3.3.2 Combinatorial Optimisation (CO)

The CO algorithm, introduced in [14], is an intuitive way to solve NILM problems. Since an aggregate load is a sum of many individual appliance loads, finding the best power-level combination can be a straightforward solution.

$$\hat{z}_t = \underset{\hat{z}_t}{argmin} |y_t - \sum_{k=1}^{K} z_t^{(k)} x_t^{(k)}| \tag{1.3}$$

Equation (1.3) defines the combinatorial approach, with $z^{(k)}$ a binary variable representing the appliance state, if on $z^{(k)} = 1$, otherwise $z^{(k)} = 0$. The combinatorial optimisation approach is multi-target. Consequently, it disaggregates all appliances simultaneously by estimating the state of each appliance, with $\hat{z}_t = \{\hat{z}_t^{(1)}, \hat{z}_t^{(2)}, ..., \hat{z}_t^{(K)}\}$.

This method faces obstacles when the value of K, the number of appliances, is high; the number of possible power combinations expands rapidly, involving massive computational resources. Increasing the number of possible combinations reduces the likelihood of a unique aggregated power level. Both the Hart Section 1.3.3.1 and CO Section 1.3.3.2 methods share a flaw in that they cannot handle negative loads, which has become a problem with the rise of prosumers (consumer and producer at the same time), often involving a two-way connection between the supplier and the prosumer. Prosumers selling their excess photovoltaic energy to the grid can cause negative loads, monitored at the meter level, during high solar irradiation and low energy demand. Those limitations of non-ML-based methods led the researchers to deeply explore the potential of DL-based approaches.

1.3.4 DL-based methods

DL techniques are of great interest in various fields for prediction, estimation or identification tasks. The high complexity of the NILM tasks makes it a platform for machine learning applications, especially for DL, a subset of ML. Unlike ML, DL approaches do not need a crucial feature engineering [43]. Nevertheless, DL-based NILM requires a prior pre-processing of the input when training. Commonly, the sliding window approach is used [44].

1.3.4.1 Sliding-window pre-processing

DL models for NILM are designed to identify patterns on the aggregate load [22]. The whole aggregate sequence is too long and too computationally demanding to input to DL models. Instead, sliced sequences are given to the model as windows. To capture all the information, a window of pre-defined length is slid over the entire aggregate load. In (1.11), the length of the window is W_{in}; thus input length is W_{in}.

1.3.4.2 DL terminology

The present section is intended to be non-exhaustive on DL theory, although major terminologies are defined, and illustrated examples are given. The objective is to provide the necessary understanding to comprehend the NILM models presented in the forthcoming sections Section 1.3.4.4 and Section 1.3.4.5.

- Neurons and dense layers

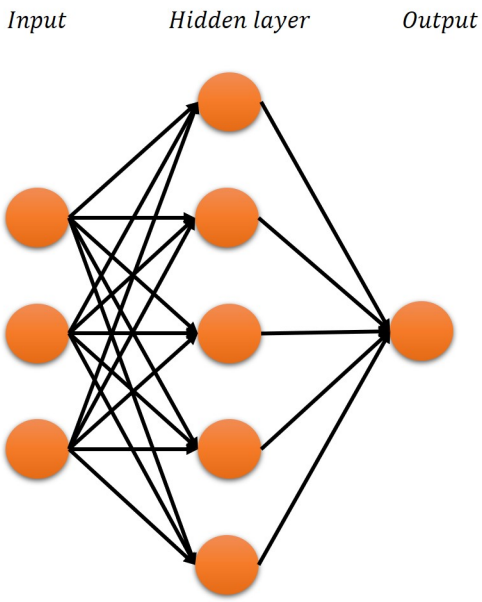


Figure 1.5. Example of a neural network

A neuron, also called a node, is a fundamental unit in DL models. A DL neuron simulates the activity of a biological neuron seen in the human brain. In the context of DL, the building blocks of neural networks are neurons, which are essential to process and transmit information. Figure 1.6 shows a single neuron representation. A neuron receives input data (y_i), and each input is associated with a weight (w_i) which quantifies the input importance. In the training phase, these weights are constantly updated to optimise the network performance. Inside each neuron, a straightforward operation is processed. Each input is multiplied by its associated weights, and the results are summed up; the obtained weighted sum is defined in equation (1.4). However, this operation remains linear and does not grant the ability to adapt to complex non-linear functions. For this reason, the weighted sum noted S, the summation is passed through an activation function. The activation function brings non-linearity to the neuron's output, which is indispensable for the network to grasp and estimate complex relationships within the data. Different activation functions can be used, such as $ReLu, Sigmoid$ or $tanh$; see equations (1.5), (1.6) and (1.7). The result of the activation operation is the output of the neuron, noted x. Moreover, a bias term b, updated during the training phase, gives an additional degree of freedom to the activation function; for instance,

bias grants to always-positive functions, like $ReLu$, the ability to activate the neuron even if the output is negative.

$$W = \sum_{i=0} w_i.y_i \tag{1.4}$$

$$g_{relu}(W) = max(0, W) \tag{1.5}$$

$$g_{sigmoid}(W) = \frac{1}{1 + e^{-W}} \tag{1.6}$$

$$g_{tanh}(W) = \frac{e^{W} - e^{-W}}{e^{W} + e^{-W}} \tag{1.7}$$

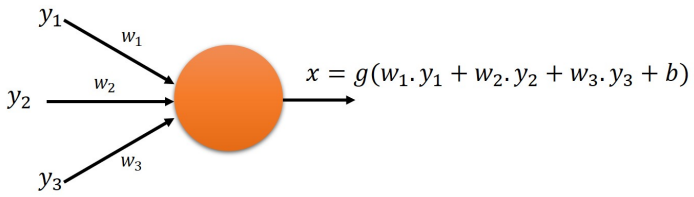


Figure 1.6. Single neuron representation, with y_i, g, w_i, b and x respectively the i^{th} input, the activation function, the i^{th} weight, the bias value and the neuron output.

A dense layer or a fully connected layer is composed of neurons. It is one of the most prevalent kinds of layers utilized in a variety of artificial intelligence activities such as time-series analysis or image recognition. The dense layer is useful for linking neurons between neighbouring layers of a neural network. In Figure 1.5, the hidden layer is dense, as this layer is connected to each of the previous and the posterior nodes.

- Convolutional Neural Network

  A Convolutional Neural Network (CNN) is a deep learning model that processes grid-like data, including images, audio, and time series data. CNNs have revolutionised computer vision tasks, achieving state-of-the-art results in tasks such as image classification and speech recognition [43, 45]. A CNN applies convolutional layers to learn spatial hierarchies of features from input data through automatic and adaptive learning. Convolutional layers are composed

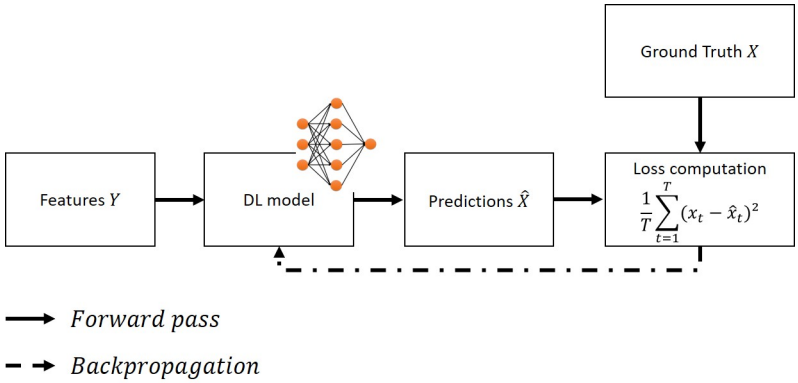


Figure 1.7. One training iteration representation; forward pass and backpropagation alternation

of adaptable filters called kernels that convolve with the input data, screening for patterns and features at the local level. The Figure 1.8 illustrates the convolution operation for an image Figure 1.8-(a) or a 1D-sequence Figure 1.8-(b). The idea of CNN training is to capture patterns by updating the kernel elements, the j_i values.

- Backpropagation

  How do the DL, and generally the ML models, learn information? DL models can learn intricate patterns and representations from enormous amounts of data thanks to the underlying process known as backpropagation [46]. Backpropagating errors automatically modify weights and biases to minimise prediction errors. The input data is fed into the neural network during the forward pass. As the information moves through the layers, it undergoes several mathematical operations, including linear transformations and activation functions (Figure 1.6), to obtain the neural network output. A loss function calculates the difference between the network's output $\hat{x}_t$ and the target power value x_t. In NILM applications, the ground truth is the individual appliance load measurement. There are two significant types of loss functions: regression and classification. For regression cases, one prefers the Mean Squared Error (equation (1.9)) and the Mean Absolute Error (equation (1.8)). In contrast, the binary-cross entropy and the categorical-cross entropy losses are used for classification.

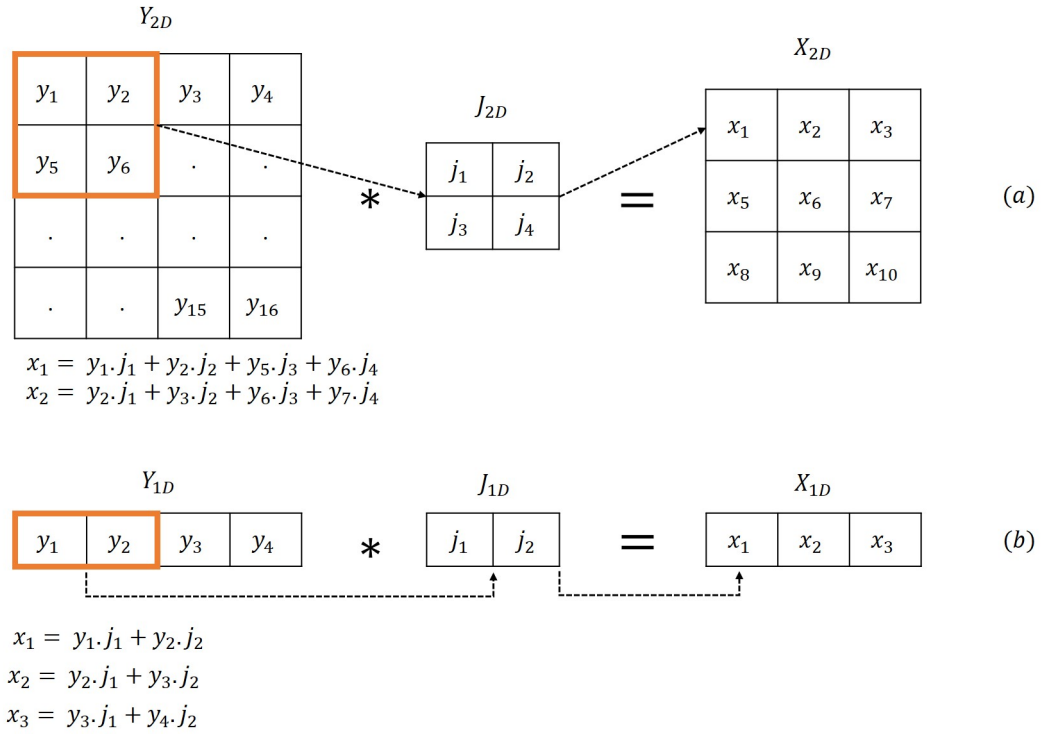


Figure 1.8. Convolutional operation (a) on a 2D element and (b) a 1D element

$$MAE_{loss} = \frac{1}{T}\sum_{t=1}^{T}|x_t - \hat{x}_t| \tag{1.8}$$

$$MSE_{loss} = \frac{1}{T}\sum_{t=1}^{T}(x_t - \hat{x}_t)^2 \tag{1.9}$$

During the backward pass, the algorithm computes the gradient of the loss function concerning each weight value by working backwards through the network. The gradient shows how responsive each weight's change is to the loss function. To minimise the loss, the neural network's weights are updated using the calculated gradients in the opposite direction of the gradient. The learning rate hyperparameter controls the range of these weight updates. Finally, the network performs the forward pass, loss computation, backward pass, and weight updates iteratively over the training data until the performance converges to a satisfactory level.

1.3.4.3 Hyperparameters for DL models

In DL domain, hyperparameters are user-defined settings for model training configuration. The hyperparameters are manually adjustable and highly suggested to be fine-tuned; the hyperparameter values need a prior adjustment to optimise the model learning. Following is a non-exhaustive list of parameters needing a prior adjustment:

- Patience for early stopping

  When should the model training be stopped? When its performances are satisfying regarding the ability to predict the output accurately. Unless in a biased experiment, the testing set cannot be seen during the training phase, so performances must be evaluated on a section of the training set, called the validation set. The learning ends in diverse possible conditions; it can be when it reaches a fixed-epoch value or if the model does not demonstrate further improvement during a defined number of epochs (patience value). The patience hyperparameter permits an earlier learning stop at the expense of higher risks of under-fitting when the patience value is set too low. Fine-tuning the patience parameter is essential to ward off non-converging models and time-consuming training.

- Learning rate

  During training, the optimisation algorithm adjusts the model weights (Figure 1.7), with the learning rate controlling the step size. A higher learning rate can result in faster convergence but may also cause overshooting of the optimal weights. On the other hand, a lower learning rate can prevent overshooting but may lead to slower convergence.

- Model architecture parameters

  The number of layers and neurons in a DL model is tightly related to the performance and the training time needed. The more parameters, the more computational resource is required, but a high parameter number only sometimes warrants good accuracy. Heavy iterative computations might come up with the perfect architecture, although in NILM scholarship, many architectures have proven effectiveness with reasonable computational resources [17, 18]. As explained in Section 1.3.4.1, classical-DL-based-NILM models rely on sliding-window preprocessing. Hence, the model does not see the whole training sequence but sees a series of subsets through a sliding window. The importance of the sliding window length has been pointed out by authors [47, 48]. A

small window captures non-discriminant patterns, while the model is bound to manage more noise for large windows. The window length is addressed in a section dedicated to model architecture because the window-length hyperparameter is correlated to the number of trainable parameters; see equations (1.10) and (1.11), the longer the sliding window, the higher the trainable parameters. Sliding-window-length adjustment is central in DL NILM to adapt to appliance pattern duration and to ensure a reasonable training time.

- Dropouts

  In deep learning, dropout is a technique to avoid overfitting in neural networks. Overfitting happens when a neural network becomes highly skilled at tasks on the training data but struggles to perform effectively on new, unseen data. Dropout helps prevent this problem by introducing randomness during training. This reduces the network's dependence on specific neurons. The technique was first introduced in work [49]. In every training iteration, dropout randomly disables a certain proportion of neurons, meaning that they are set to zero on the forward pass (Figure 1.7) and their values are frozen on the backward pass, making them insensitive to the gradient. This is done to improve generalisation abilities on unseen data.

1.3.4.4 Sequence-to-sequence (S2S)

Kelly and Knottenbelt [21] identified the potential of CNN-S2S models. Those DL models showed significant improvement compared to classical methods such as the Combinatorial Optimisation approach or the Factorial Hidden Markov Model [50]. S2S for the NILM task disaggregates a W_{in}-length-main-load sequence to a W_{out}-length output sequence in the equation (1.10). Usually, CNN-S2S approaches respect $W_{in} = W_{out}$, although some researchers [30] explored sequence-to-subsequence architectures where $W_{in} > W_{out}$.

$$f^{(k)}_{s2s}(Y_{t:t+W_{in}-1}, \theta) = \hat{X}^{(k)}_{t:t+W_{out}-1}, \tag{1.10}$$

with $Y^{(k)}_{t:t+W_{in}-1}$ is a main load chunk and $\hat{X}^{(k)}_{t:t+W_{in}-1}$ being the load of the appliance k. θ is the hyperparameter set. Denoising Autoencoder (DAE) architecture for NILM issues, illustrated in Figure 1.10(a), has been introduced in [21]. The general idea in DAE is to add noise artificially in the input sequence (Autoencoder) to train the model to retrieve the clean signal (Denoising). In NILM problems, the noise is already

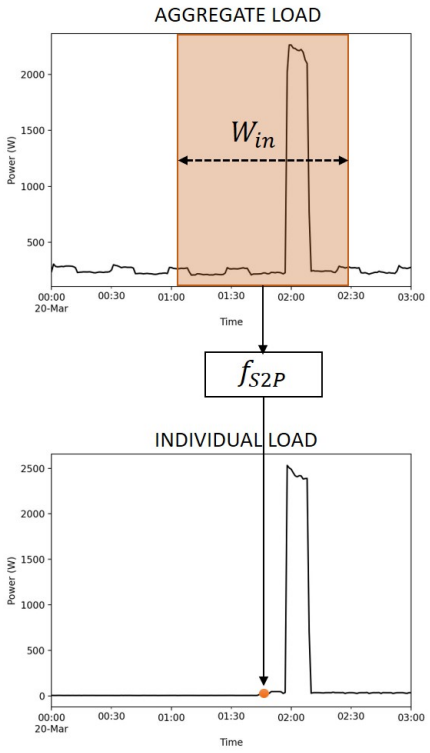


Figure 1.9. Principle of the CNN-S2P approach and representation of a sliding window of width W_{in}.

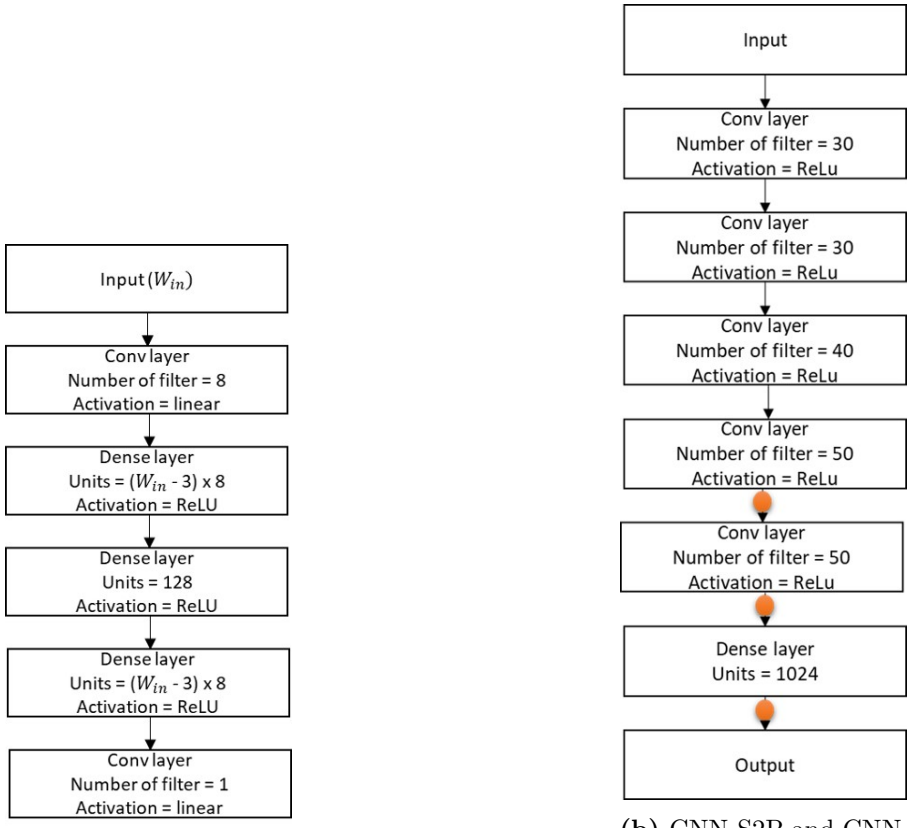
generated by the operations of other devices besides the one(s) to identify; thus, the architecture is designed only to denoise the aggregate load. Another S2S method is the CNN-S2S [21], which utilises the same architecture depicted in Figure 1.10(b).

1.3.4.5 Sequence-to-point (S2P)

In work [18], the S2P approach overtakes S2S methods and is among the most studied in literature [17, 22, 48]. The classical S2P architecture is represented in Figure 1.10(b). The first layers are the CNN needed for feature capturing.

$$f_{s2p}^{(k)}(Y_{t:t+W_{in}-1}, \theta) = \hat{X}_{\tau}^{(k)}, \tag{1.11}$$

The concept behind the S2P is that the centre point of the window can be understood through the values preceding and succeeding it, as represented in Figure 1.9. In



(a) DAE architecture

(b) CNN-S2P and CNN-S2S approaches

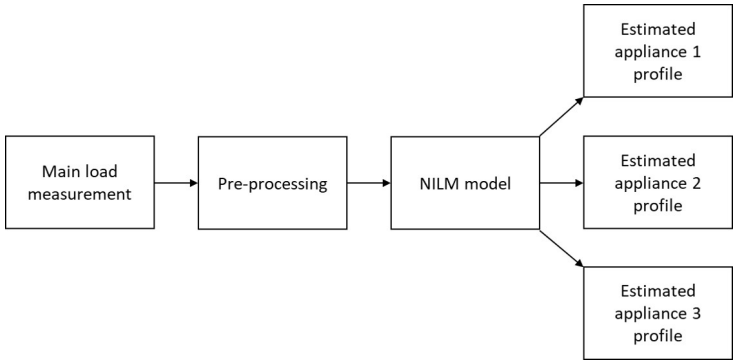
Figure 1.10. Two model architectures, orange circles represent the dropouts

[17,18], the window length is set as an odd number to have a midpoint. Throughout this work, W_{in} is taken odd, so let be $\tau = \frac{W_{in}+1}{2}$.

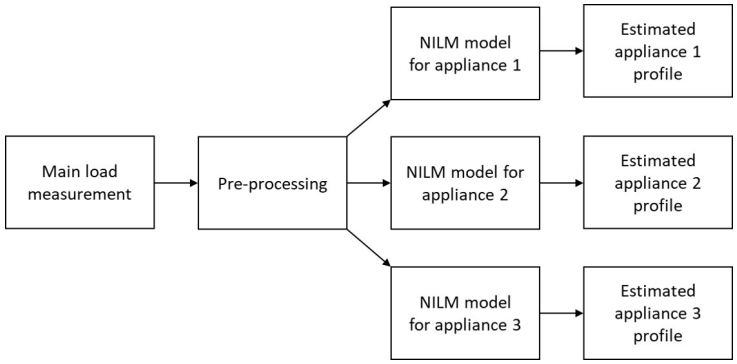
1.3.5 Single-target and multi-target models

The disaggregation task is intuitively associated with separating an aggregated load into many individual loads. Multi-target approaches, illustrated in Figure 1.11(a), predict the individual appliance profiles using a unique model. Hence, the idea is to build a model capable of identifying a finite number of appliance types. A unique model is implemented in a multi-target approach, making it a more computationally effective method. However, the model cannot be tuned differently for each

appliance type. To contrast, the significant advantage of the single-state approach Figure 1.11(b) is that a particular tuning is possible according to the appliance to detect. For instance, a shorter sliding window (see Section 1.3.4.1) is preferable for a microwave estimation to a washing machine detection, as the microwave generally runs for a short time.



(a) Multi-target



(b) Single-target

Figure 1.11. Two NILM solution structures

1.4 Conclusion

This Chapter 1 comprises a thorough definition of NILM, to understand what makes NILM an important research theme. Standard ILM solutions are not cost-efficient and are less adapted for large-scale deployment. NILM can overcome this barrier in the conditions that current NILM solutions can give accurate disaggregation despite the well-known NILM technical complexity. In the literature, the authors separate DL-based and non-DL-based approaches, as DL has gained much attention in the NILM domain. In the next chapter, a series of experiments is conducted to understand how NILM models perform and what are the significant settings impacting those models.

CHAPTER 2

NILM experimentation

Developing and proposing new models for NILM has recently become a prominent trend in academic research. Researchers widely agree that DL approaches offer the most effective methodology for NILM tasks. However, the results obtained from these models tend to be unstable due to variations in the experimental setup, selection of houses, and appliances. A NILM model must demonstrate high generalisation capability or adaptability to diverse households for successful implementation. The operational efficiency of NILM at a low implementation cost is a significant prerequisite for practical deployment. The efficient functioning of the model at a high sampling period should reduce the need for expensive meters. This study experiments with a 1-minute sampled NILM approach under various experimental configurations. The initial configuration comprises a "seen" setting, while the latter is centred around an "unseen" scenario. A single unique house dataset is used for training and evaluation in the seen configuration. In contrast, in the unseen configuration, the NILM models are evaluated based on the data from a house that has never been used for training. For this investigation, the experimental platform used is the NILM toolkit (NILMTK), a widely used toolkit in the field. Effects of the sample period, time of the day, weekdays and weekends, and length of training data on NILM performances are also evaluated with short NILM experiments. This series of experiments gives an overview of significant factors that determine NILM performances and points out levers for improvement. The research questions tackled by the Chapter 2 are:

- What are the major parameters involved in NILM performance?
- To what extent can a NILM solution replace a standard ILM?

2.1 NILMTK

NILMTK is a significant contribution in the realm of NILM [17, 50, 51] to facilitate the evaluation of NILM algorithm performances. This open-source toolkit serves as a valuable resource for researchers, enabling them to compare different energy disaggregation algorithms in a reproducible manner. A substantial contribution of NILMTK is its effort to compare multiple disaggregation approaches across various publicly available datasets, thus promoting greater transparency and collaboration within the research community. NILMTK encompasses parsers for various existing datasets. NILMTK data parsing consists of converting the raw public datasets to the NILMTK-DF (NILMTK Data Format), a Hierarchical Data Format (HDF) file [52]. Additionally, it includes a collection of preprocessing algorithms to handle missing data and data standardisation, benchmark disaggregation algorithms from the scholarship, and a standardised set of evaluation metrics. To ease user handling of NILMTK, an API (Application Programming Interface) has been developed to allow the user to set up the experiment quickly and easily without the need to handle complex coding. By providing these essential components, NILMTK facilitates rigorous evaluations and benchmarking of NILM methodologies.

2.2 NILM datasets

DL models are known to need a considerable amount of data. For DL-based NILM models, the data needed are composed of an aggregate load and the associated ground truths, i.e. the individual appliance loads. One of the most used datasets is REDD (Reference Energy Disaggregation Data Set) [53] composed of 6 houses from the USA. In European regions, UKDALE appears as the notorious dataset [54], with 5 houses recorded from the United Kingdom from 2013 to 2015. Nevertheless, only house 1 has a high data length, contrary to the rest. The limited data regarding house variety and temporal length makes REDD and UKDALE less appropriate for thorough evaluations. Indeed, in the training phase, having a variety of houses enhances the model's ability to generalise due to the diversity of seen data. Testing on several houses is required to establish strong empirical evidence of generalisation capabilities. A sufficient data length for the NILM experiment ensures to make the model less seasonal-dependent. For these reasons, some researchers have sought to contribute to recording more diverse and long data. In 2015, DATAPORT [55] was claimed to be the largest source of disaggregated energy data with 722 houses.

The REFIT dataset [24] is a compelling collection of data from 20 houses, each with almost a year and a half of monitored activity. REFIT and DATAPORT were monitored in actual conditions where data from real homes with real families were collected. Referring to ILM properties (Section 1.2), this data collection method has its drawbacks, specifically on possible poor data quality (missing data and outliers) and inaccurate data labelling. To address these issues, researchers have investigated synthetical datasets. SynD [56] is a synthetic energy dataset that combines traces of actual household appliances. The SynD system was created with around 20 different types of appliances, enabling it to simulate various appliance combinations.

2.3 Metrics

As per [57–59], the evaluation of load disaggregation can be broken down into two tasks: event detection and energy estimation. For event detection, the primary metric considered in this paper is the f1-score defined in the equation (2.1) calculated with the True Positive Rate (TPR), see equation (2.3) and the False Positive Rate (FPR), see equation (2.4). As NILM is originally a regression problem, a user-defined threshold separates on-events from off-events. This work selects three thresholds to evaluate the F1Score metric at 25W, 100W and 500W. The thresholds compute the confusion matrix variables in figure Figure 2.1. The Mean Absolute Error (MAE) assesses an average power deviation between ground truth and prediction. However, MAE metric comparison can be biased by low-amplitude noises when the device is not operating and by the operating power level. Then, the Normalized Disaggregation Error (NDE_k), see equation (2.6), is also applied in this work. As MAE and NDE are error metrics, a 0 value is sought for ideal performances.

- F1score

$$\text{F1Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{2.1}$$

- Precision

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.2}$$

- Recall or sensitivity or True Positive Rate (TPR)

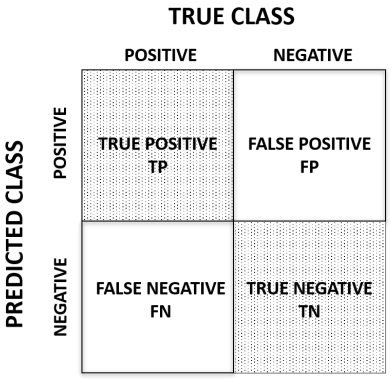


Figure 2.1. Confusion matrix

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.3}$$

- False Positive Rate (FPR) or fall-out

$$\text{FPR} = \frac{FP}{FP + TN} \tag{2.4}$$

- Mean Absolute Error

$$\text{MAE}^{(k)} = \frac{1}{T} \sum_{t=1}^{T} |\hat{x}_t^{(k)} - x_t^{(k)}| \tag{2.5}$$

- Normalized Disaggregation Error

$$\text{NDE}^{(k)} = \frac{\sum_{t=1}^{T} (x_t^{(k)} - \hat{x}_t^{(k)})^2}{\sum_{t=1}^{T} (x_t^{(k)})^2}, \tag{2.6}$$

2.4 Model fine-tuning

DL model performances rely on user-defined hyperparameters requiring a fine-tuning process. The aim is to determine the optimal configuration parameters regarding the model's accuracy. Generally, a grid search approach is applied to deal with the many

hyperparameters involved. The model performances are assessed for each hyperparameter combination. This grid-searching process requires heavy computational resources. A randomised grid search can address this issue by reducing the number of assessed combinations by random selection. However, this type of selection faces the risk of missing the optimum. Building a less computational-demanding grid-searching space is possible by selecting only the most meaningful hyperparameters and defining reasonable ranges for each value. For NILM applications, the sliding window length for feature capturing is an additional hyperparameter to be tuned. The present subsection aims to accurately determine the hyperparameter values and observe the model sensitivity to different adjustable settings. The fine-tuning study focuses on three critical variables: the learning rate, the patience for early stopping and the window length (Section 1.3.4.3).

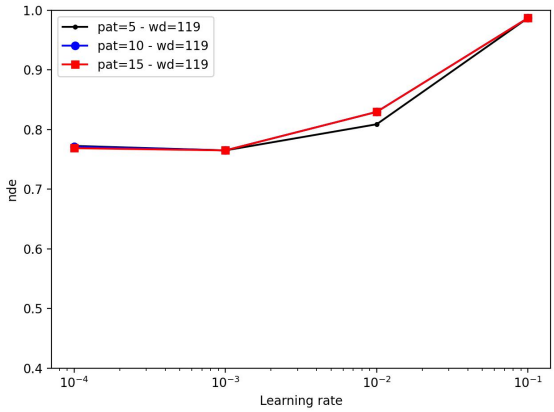
Table 2.1. Fixed parameters for fine-tuning experimentation. Dates are formatted as YYYY-MM-DD.

Training house	refit2
Validation house	refit2
Testing house	refit2
Dates train	2014-07-01 to 2014-07-20
Dates test	2014-09-01 to 2014-09-30
Sampling period (s)	60
Validation split	15%
Batch size	1000
Maximum epoch	100

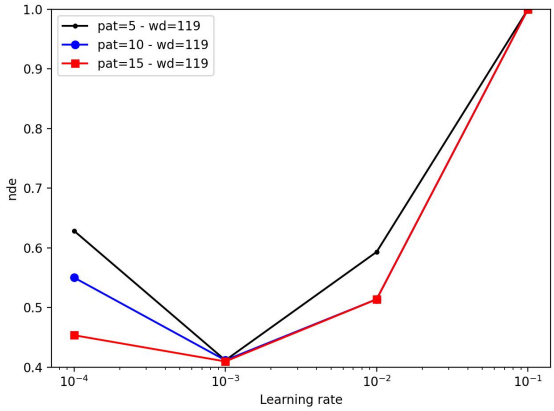
Adjustable variables are not necessarily fully independent when performing fine-tuning tasks, i.e., learning rate variations might impact the best patience value. There is no standard hyperparameter value combination for NILM experimentation; for instance, some authors [17] do not use patience for early stopping while [22] do. Diversity in the choices of datasets, houses, training periods, or NILM models makes it ambitious to emerge "the best" hyperparameter combination. However, there are unfavourable model settings that lead to skewed NILM results. The present investigation is conducted on two common appliances, the kettle and the dishwasher.

- Learning rate lr

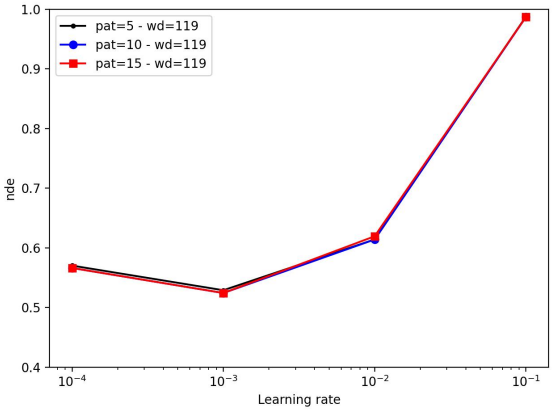
  Figure 2.2 and Figure 2.3 give an overview of the effect of learning rate variations on NDE-error (2.6). Notably, the errors increase when a high learning rate is set. The dishwasher and kettle disaggregations perform better when



(a) DAE

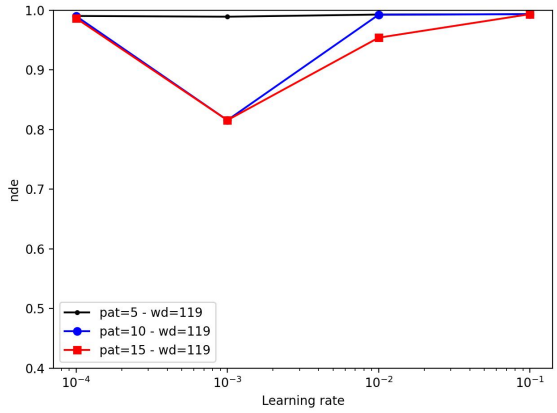


(b) CNN-S2P

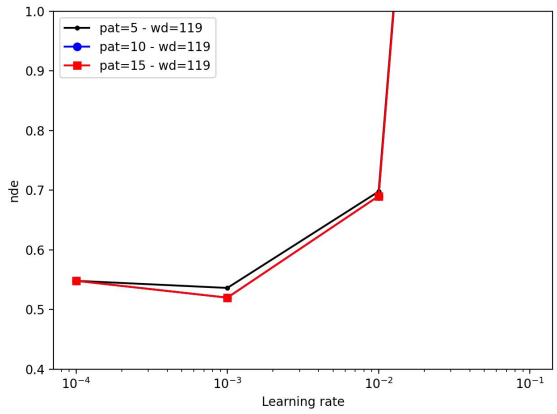


(c) CNN-S2S

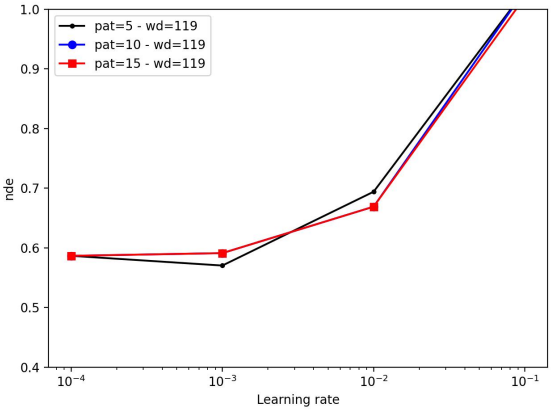
Figure 2.2. Model fine-tuning patience (pat), learning rate (lr) for a fixed window length (wd) on the dishwasher.



(a) DAE

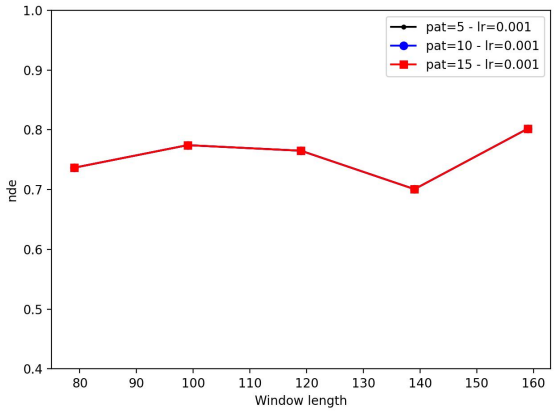


(b) CNN-S2P

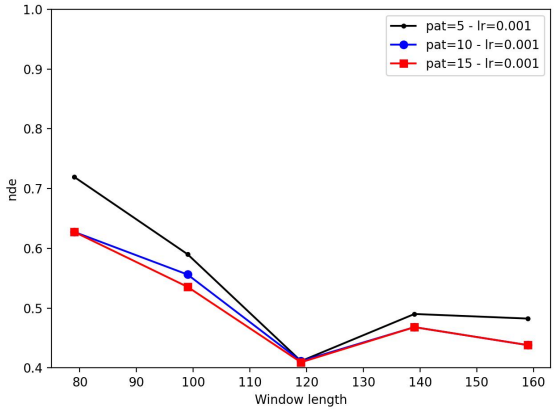


(c) CNN-S2S

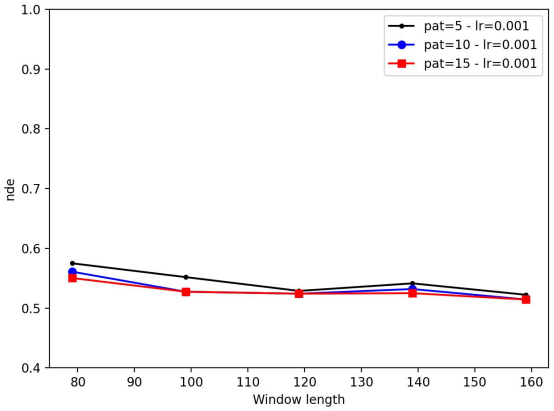
Figure 2.3. Model fine-tuning patience (pat), learning rate (lr) for a fixed window length (wd) on the kettle.



(a) DAE

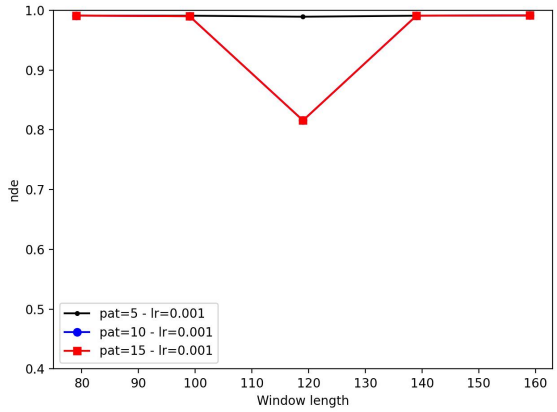


(b) CNN-S2P

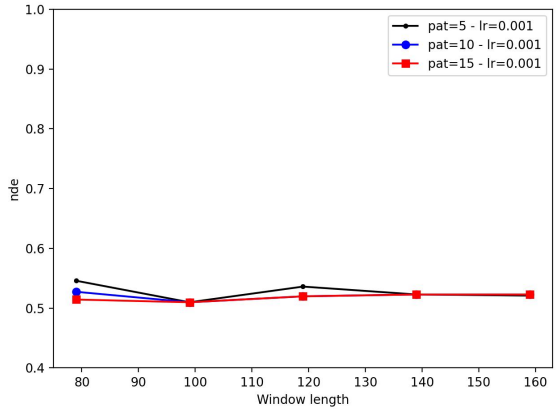


(c) CNN-S2S

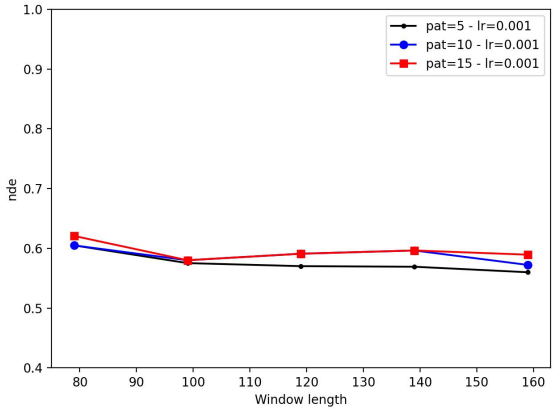
Figure 2.4. Model fine-tuning patience (pat), window (wd) for a fixed learning rate (lr) for dishwasher detection.



(a) DAE



(b) CNN-S2P



(c) CNN-S2S

Figure 2.5. Model fine-tuning patience (pat), window (wd) for a fixed learning rate (lr) for kettle detection.

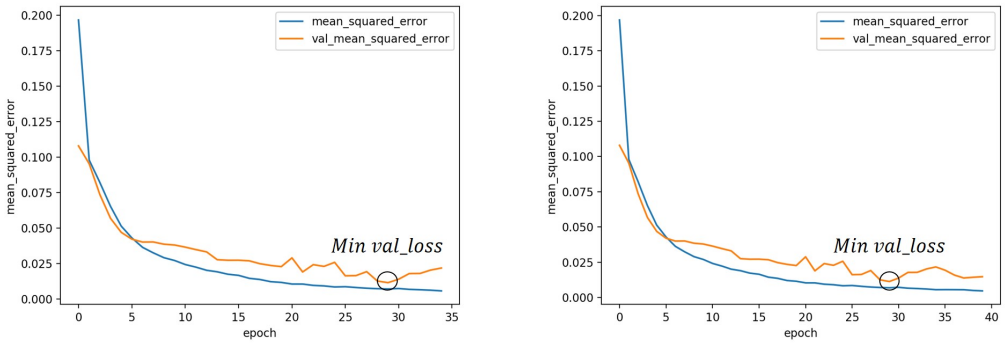
$lr \leq 10^{-2}$. The error escalates rapidly when the lr value is above this threshold, caused by the fact that the model does not converge due to the overshooting effect when lr is too high (Section 1.3.4.3). With a high lr value, the weight updates are too consequent to optimise the model's performance. Remembering that on-activations are more infrequent than off-activations (at least for many common appliances) is essential; a finer weight-updating might be required to capture those rare moments of on-activations. On the contrary, all three models perform better when the lr-variable is low ($lr \leq 10^{-3}$), except for the kettle power estimation with the DAE model where accuracy drops when $lr = 10^{-4}$, probably a higher patience value is needed for the model to converge.

- Patience

  The training phase is stopped whenever the model is satisfactory regarding the model performance on the validation set. Although DL models are black-box models and complex to comprehend, validation curves offer a convenient visualisation of the training process. The model is overfitting when it has high abilities to predict/estimate the training set, whereas it performs poorly on the validation set. In Figure 2.6, training is stopped when the model has not improved for a *patience*-number of epochs. Patience fine-tuning depicts a slight improvement for CNN-S2P and CNN-S2S when increasing the patience in Figure 2.2. For dishwasher disaggregation, DAE seems less sensitive to the patience hyperparameter, as respective curves in Figure 2.4(a) and Figure 2.2(a) are nearly overlapped totally. In the training phase for kettle detection, the DAE model is more sensitive to the patience value, as choosing $patience = 5$ under-performs in Figure 2.3(a) and Figure 2.5(a). The proximity between the fine-tuning curves for $patience = 15$ and those for $patience = 10$ suggests that further training of the models may not be necessary.

- Window length

  The authors of the work [48] have dedicated to fine-tuning window length for CNN-S2S and CNN-S2P learning. Indeed, this hyperparameter is substantial as it is among the factors that determine the ability of the model to learn patterns. A high window length is likely to capture more noise, whereas a short sequence might not be sufficient to capture a discriminant pattern entirely. In the presented fine-tuning experimentation, a window-length fine-tuning is proposed on a long-running appliance (dishwasher) and a short-running one (kettle). In Figure 2.4(c), the CNN-S2S model shows improvement when increasing window size from 79 to 119; however, the performance increase is not



(a) Patience = 5 **(b)** Patience = 10

Figure 2.6. Validation curve of the S2P training for the dishwasher disaggregation, $Min\ val\_loss$ is the lowest error values on the validation set.

as sharp as the one for CNN-S2P, in Figure 2.4(b), which seems highly sensitive to the window length. Despite having higher errors than CNN-S2P and CNN-S2S, the DAE model also looks sensitive to the window length with a minimum at $wd = 139$. When considering the Figure 2.5 for fine-tuning for kettle detection, CNN-S2P and CNN-S2S models do not display a precise preferred sequence length. This is likely due to the kettle running briefly; the imbalance between on and off periods is more substantial with more off periods. Moreover, in this work, $60s$-sampled experimentation (see Table 2.1) is performed to assess the low-frequency capabilities of NILM approaches, devices like kettle generally run during a period close to or shorter than $1min$. Consequently, patterns are smoothed and likely to be less discriminant. Even though single-target NILM models (see Figure 1.11(b)) are applied, a unique model should be capable of detecting when the device is running but also when the device is not running. Hence, single-target models indirectly learn patterns from other appliances. In the work [22], the authors exploit this indirect detection ability to apply transfer learning to NILM. The proposed kettle models are likely to be tuned to detect better when the kettle is off than when it is on due to the imbalance.

2.5 DL-based vs non-DL-based models

The models have been presented in Section 1.3.3 and Section 1.3.4. In the conducted experimentation, DL-based methods demonstrate higher accuracy according to Figure 2.7, where CO and Hart85 barely reach $NDE = 1.0$. In the work [17], artificial aggregation is applied to test the models, i.e. the models are trained and tested on a synthetic load generated by aggregating the available individual loads. According to the result of Batra et al.'s work [17], CO and Hart85 models seem to perform well, nearly as well as AI-based models, on artificial aggregate conditions. However, the performances drop sharply on a real aggregate assessment. Models, whether DL or non-DL-based, built with artificial aggregate, have low noise awareness as the training set contains only pure signals. In Batra et al. [17], when comparing an actual aggregate condition with an artificial one, the mean absolute error (MAE) drops by an average of $86W$ for CO. However, the CNN-S2P model demonstrates a much lower drop of $9W$ in MAE. Consequently, DL models have higher capabilities to cope with noises in the data. Non-DL models often require manual feature identification like Hart's method [13, 14]; it is necessary to spot every power edge and power level, which rapidly becomes a challenge in some households where the main load is highly noised. Using CNN allows the capture of the most discriminant features from the training set automatically, making the DL models more scalable than non-DL models. In the presented experiment, CNN-S2P reaches a promising 0.67 NDE value for the *Seen* configuration and 0.78 for the $Unseen5$.

2.6 "Seen" vs "unseen"

Table 2.2. Parameters for Seen vs Unseen experimentation

Param	Seen	Unseen1	Unseen2	Unseen3	Unseen4	Unseen5
Testing houses	5	5	5	5	5	5
Training houses	5	2	2 3	2 3 6	2 3 6 13	2 3 6 13 19
Total testing days	28	28	28	28	28	28
Total training days	150	150	150	150	150	150
Sampling period (s)	60	60	60	60	60	60
Validation split	15%	15%	15%	15%	15%	15%
Batch size	1000	1000	1000	1000	1000	1000
Maximum epoch	100	100	100	100	100	100
Patience	10	10	10	10	10	10
Learning rate	0.001	0.001	0.001	0.001	0.001	0.001
Window length	119	119	119	119	119	119
Appliances	dishwasher, kettle, microwave, washing machine					

Some NILM researchers [13, 14, 18, 42] train and evaluate their model on a unique house; evaluation is processed on houses seen during training. One possible implementation, aligning with this evaluation method, is to install an energy meter on each appliance temporarily. Once enough data is collected, the meters are uninstalled (and installed on another house), and a NILM model built with the measured data is embedded into the household for disaggregation purposes. The energy operator in charge of the ALM tasks can opt for a manual setup; instead of having a long ILM period, the operator can install the sensor on each appliance and run appliances on purpose to create possible combinations of each device. However, manual processing increases the time the operator spends in the dwelling depending on the number of appliances and their standard operating duration. It is more cost-effective for the energy operator to monitor intrusively for a short period than for an extended period to reduce the requirement of buying many metering devices for other houses left to monitor. Even though a shorter period is more cost-effective, a too-short data length causes poor model performances due to an inherent seasonality on residential load profiles; winter data are less correlated to summer data, and holidays demonstrate different demand patterns. Unseen experimentation consists of assessing the models on a house unseen in the training phase [22]. This evaluation is critical for NILM researchers as it directly assesses the generalisation capabilities of the models. There are two levels of generalisation: on the first level, the model's ability to correctly disaggregate a load of a house from the same region, and on the second level, the capacity to disaggregate a load from a completely different region from the ones considered in training. In this work, only the first level of generalisation is experimented with. To operationally implement a NILM model to detect individual appliance loads on an unseen house, first, it is necessary to train the model on fully-known houses, and then the built model is implemented to predict the individual loads of the house to disaggregate. In this section, two configurations are confronted quantitatively: *Unseen* and *Seen* scenarios. Table 2.2 gives the fixed parameters for comparing both configurations. Reasonable values of the hyperparameters, window length, learning rate and patience are determined based on the fine-tuning process in Section 2.4. A unique test set of 28 days from refit5 is shared among all experiments. To some significant degree, the length of the training set infringes on NILM performances. Therefore, a fixed training set length of 150 days is applied. For the NILM model to generalise well, the training set needs to have diversity. To assess the generalisation capabilities of unseen models, different numbers of houses (from 1 to 5) in the training set are tested for a fixed training length (150 days). The differences between seen and unseen configurations are blatant in Figure 2.7 for the DL models, i.e. CNN-S2P, CNN-S2S and DAE. When experimenting with the CNN-S2P model,

the NDE value for the *Seen*-scenario is about 0.66, while for the $Unseen5$-scenarios, NDE equals 0.78, representing an 18% error increase. A possible way to mitigate the gap between *Seen* and *Unseen* is to increase the number of houses within the training set. According to Figure 2.7, the accuracy of the DL models increases as more houses are included in the training. Note that for all experiments, the length of the training set is kept constant (see Table 2.2). Adding a new house to the training set increases diversity by including new combinations of appliances, schedules, and power levels. In work [60], it is stated that 3 to 6 examples of activations for a given appliance are sufficient to build a generalised model, to give an at-first-sight extrapolation of this result, 3 to 6 different houses are needed to come up with a generalised single-target NILM model in the condition that each house is composed of a unique appliance type version. Nevertheless, further experimentation is required to determine the number of houses for which the models have the best generalisation abilities. This section provides an overview of the impact of the number of houses and, thus, the importance of data variability on NILM performance.

2.7 Sample period

Table 2.3. Parameters for sample period tuning experimentation

Testing houses	5
Training houses	5
Total testing days	28
Total training days	150
Validation split	15%
Batch size	1000
Maximum epoch	100
Patience	10
Learning rate	0.001
Window length	119
Appliances	dishwasher, kettle, microwave, washing machine

In France, since 2016, the Linky smart meter has been massively installed in households to replace electromechanical meters. The main meter is legally bound to measure at specific sampling periods [61], 1 hour, 30 minutes and 10 minutes for grid management and electricity billing tasks. In all modern dwellings connected to the grid, a meter is used by the grid operator to bill the electricity. Ideally, the

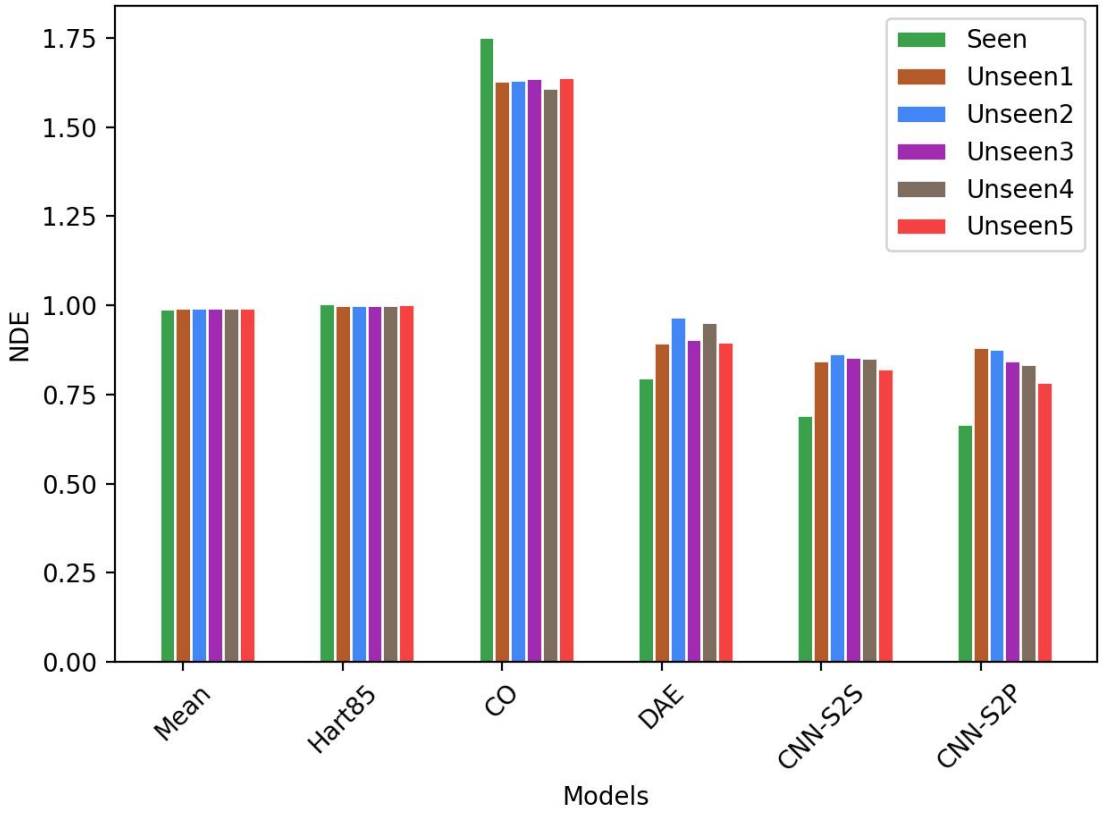


Figure 2.7. Average NDE for Unseen and Seen scenarios

NILM operator can use the initially present meter to minimise the costs. However, in NILM, only a few works [33] consider very long sampling periods superior to 1 minute. Indeed, the higher the sampling period, the more faded the load patterns are. Thus, detecting individual appliances becomes very challenging, particularly for short-running appliances, such as the kettle or the microwave, that generally run for a few minutes or seconds. Figure 2.9 provides a graphical representation of the effect of the sampling period on individual and main load curves. When applying a high-time step NILM, the number of monitoring points required to describe the patterns decreases. Consequently, the patterns are visually smoothed, and short-running device patterns (kettle, microwave) tend to disappear. Similarly, DL models capture fewer features and discriminant patterns for low-sampling rate data. Practically, the sampling period is crucial for the NILM operator; the lower the sampling period, the higher the installation and operational costs. Figure 2.8 visualises the NDE increase

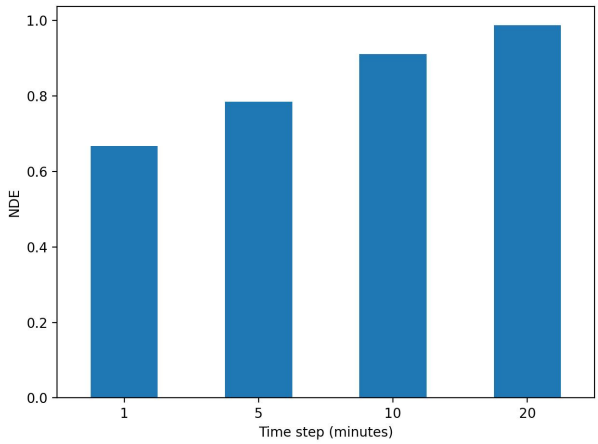
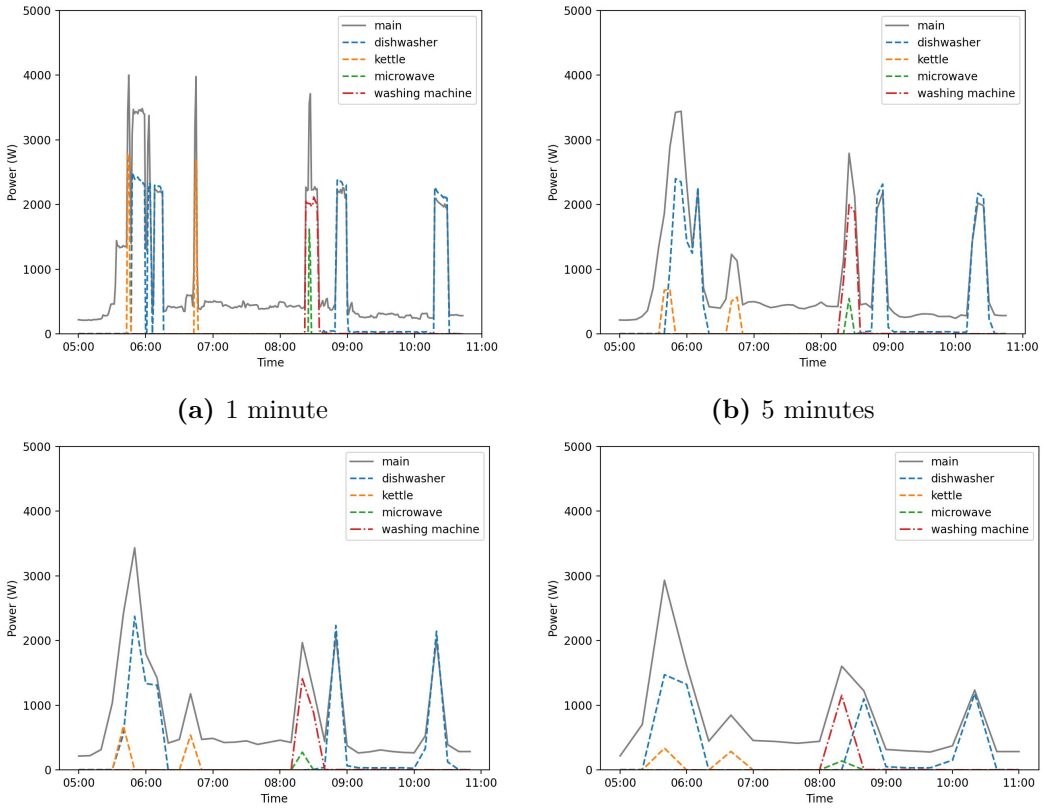


Figure 2.8. Average NDE function of the data sampling period using the CNN-S2P model

for sampling period increase. Experiments are held according to the parameters in Table 2.3 and came up with an NDE value that increases from 0.67 to 0.99 when shifting the time step from 1 to 20 minutes. A dilemma arises: whether to have a highly accurate and not cost-efficient NILM or to have a less precise and cost-efficient NILM. Depending on the NILM application, the NILM operator can install an additional meter beside the grid meter to collect finer data at a higher sampling rate.

2.8 Hours of day and days of week

In this section, experiments are held to quantify the correlation between human activity and NILM performance. The hypothesis to verify is that the more running appliances composing the aggregate load, the harder it is for NILM models to disaggregate correctly. To a certain extent, energy consumption is correlated to human occupancy [62, 63]. In periods of high household occupancy, it is more likely to have higher energy consumption with several electrical appliances running simultaneously. NILM is a complex task, particularly for the uncountable potentially existing appliance combinations (see Section 1.3.2). Those appliance combinations are emphasised during high-consuming periods when rare combinations can appear. As far as the author knows, there is no existing NILM dataset with real-time occupancy information. To overcome this barrier, the energy data are thoroughly analysed to spot the potential high-occupancy period. In a standard household, high occupancy is gener-



(a) 1 minute

(b) 5 minutes

(c) 10 minutes

(d) 20 minutes

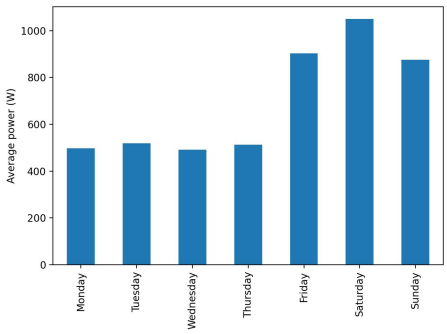
Figure 2.9. Observation of main and submeter measurements under different sampling rate conditions. Note in Figure 2.9(a), the dishwasher load can sometimes be higher than the main load due to measurement errors in the REFIT dataset.

ally in the evening after work, when people prepare dinner (oven, stove ...) and do all sorts of chores (vacuum, washing machine ...). Even if there seems to be a common trend, all households are different, and various behaviours might be observed. Some dwellers are used to staying home during the weekend, while others prefer to visit family, travel, do groceries, or do diverse outside activities. The experimental configuration shown in Table 2.4 examines the time of day and weekday impact. The case study of the REFIT5 household shows higher consumption during the weekend. One peak is observed in the morning and another in the evening. In the considered house, the dishwasher is mainly used at night; after midnight, the user probably schedules the dishwasher activation to take advantage of low electricity prices. The energy share of the dishwasher is consequent during those hours (between 00:00 and

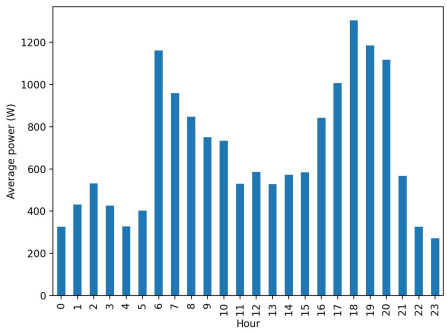
Table 2.4. Parameters for time of day and weekday impact observation

Testing houses	5
Training houses	5
Total testing days	28
Total training days	150
Validation split	15%
Batch size	1000
Maximum epoch	100
Patience	10
Learning rate	0.001
Window length	119
Appliances	dishwasher

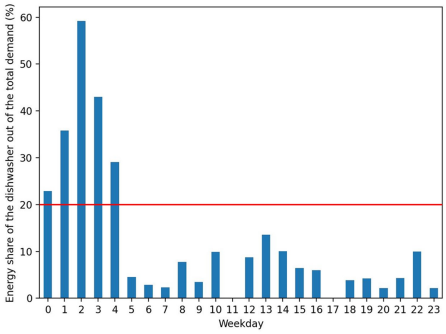
05:00 a.m.), reaching 60%. Dwellers will likely be asleep at these hours, so only a few appliances are activated. Seeing the scatter diagrams Figure 2.11, for high energy shares, the dishwasher detection reaches a very accurate result with a $NDE = 0.15$ while $NDE = 0.83$ when considering low energy shares. The simplicity of the signal causes this large gap during these after-midnight hours when nearly no noise appliances are activated. It can be seen that the points in Figure 2.11(b) are not far from the bisection, revealing a very convenient NILM except for a cluster at coordinates $\approx (1000, 2500)$ where the power level is under-estimated by the model. During the daytime, human activities lead to a more complex main load profile with many more devices that can be confused with the targeted one. When filtering in weekends and weekdays, a smaller NDE gap is visible for weekend days on average $NDE = 0.68$ while $NDE = 0.40$ during weekdays. The difference can be explained partially by Figure 2.10(a), showcasing a more significant energy demand during weekends. Similarly to the above hourly NDE comparison, a higher weekend consumption means more electrical appliance usage, leading to a more complex signal. This short experiment allows observing the performance variation under different aggregate load shape complexity. Nevertheless, given the possible existing load shapes, further investigation is necessary to give a general result on the performance gap between complex and non-complex periods.



(a) Average main power for each day of the week



(b) Average main power for each hour of the day

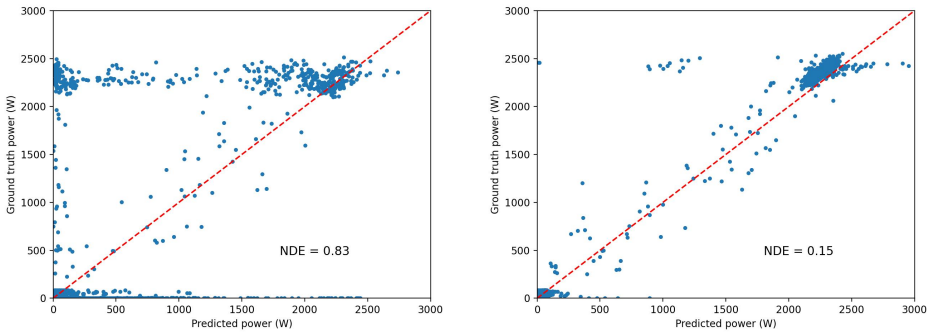


(c) Dishwasher energy share for each hour of the day

Figure 2.10. Observation of the REFIT5 household behaviour during the NILM testing period

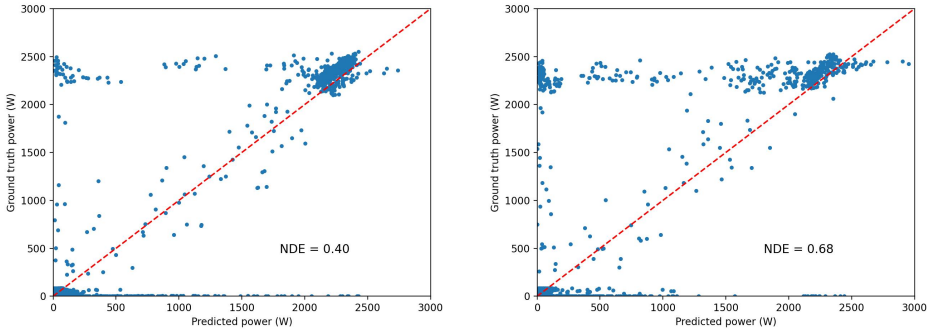
2.9 Training dataset length

Having enough data is essential to build a relevant NILM model; authors [27, 28, 64] highlight the data scarcity for DL NILM models. In this section, 5 different lengths of the training dataset are experimented with to assess the effect on model performances. The parameters are presented in Table 2.5. The training set comprises a unique household, REFIT2, to assess only the effect of dataset length. What one may expect is that the more information is given to the model, the higher the generalisation capabilities. However, the information must have quality to ward off overfitting risks. This section demonstrates the consequences of a NILM model trained on a unique house with different dataset lengths. In Figure 2.13 , only the DAE model improves when the training set length is increased, whereas other models do not show apparent



(a) Low ($< 20\%$) share hours **(b)** High ($\geq 20\%$) share hours

Figure 2.11. Scatter diagrams for high and low dishwasher energy shares, see Figure 2.10(c) for hourly energy share values.



(a) Weekdays **(b)** Weekend

Figure 2.12. Scatter diagrams for week and weekend days for dishwasher detection using CNN-S2P model.

and significant enhancement. On the contrary, CNN-S2P model performances are downgraded, suggesting that the CNN-S2P has already learnt enough with a one-month training set to predict the unseen house REFIT5. Considering the CNN-S2S, the best result, $NDE = 0.85$, is obtained with a three-month length and knows slight declines for a more extended training dataset. With a one-month dataset in training, the CNN-S2P, with a performance of $NDE = 0.86$, is more accurate than the DAE trained with a five-month dataset reaching $NDE = 0.89$. Errors for DAE decreased from 0.95 for a one-month length to 0.89 for a five-month length and the CNN-S2S from 0.87 to only 0.85. Therefore, the improvement margin in varying only the training set lengths is thin. This experimentation provides scope for future NILM datasets; recording a long period does not necessarily lead to powerful NILM

models. This investigation has limitations, as only 5 months were considered, from March to July, when the temperature is felt hotter in the UK (REFIT dataset is recorded in the UK). Hence, the experiment does not consider season change; the models were assessed and trained on data from more or less the same season.

Table 2.5. Parameters to evaluate data length effect

Testing houses	2
Training houses	5
Total testing days	28
Validation split	15%
Batch size	1000
Maximum epoch	100
Patience	10
Learning rate	0.001
Window length	119
Appliances	dishwasher, kettle, microwave, washing machine

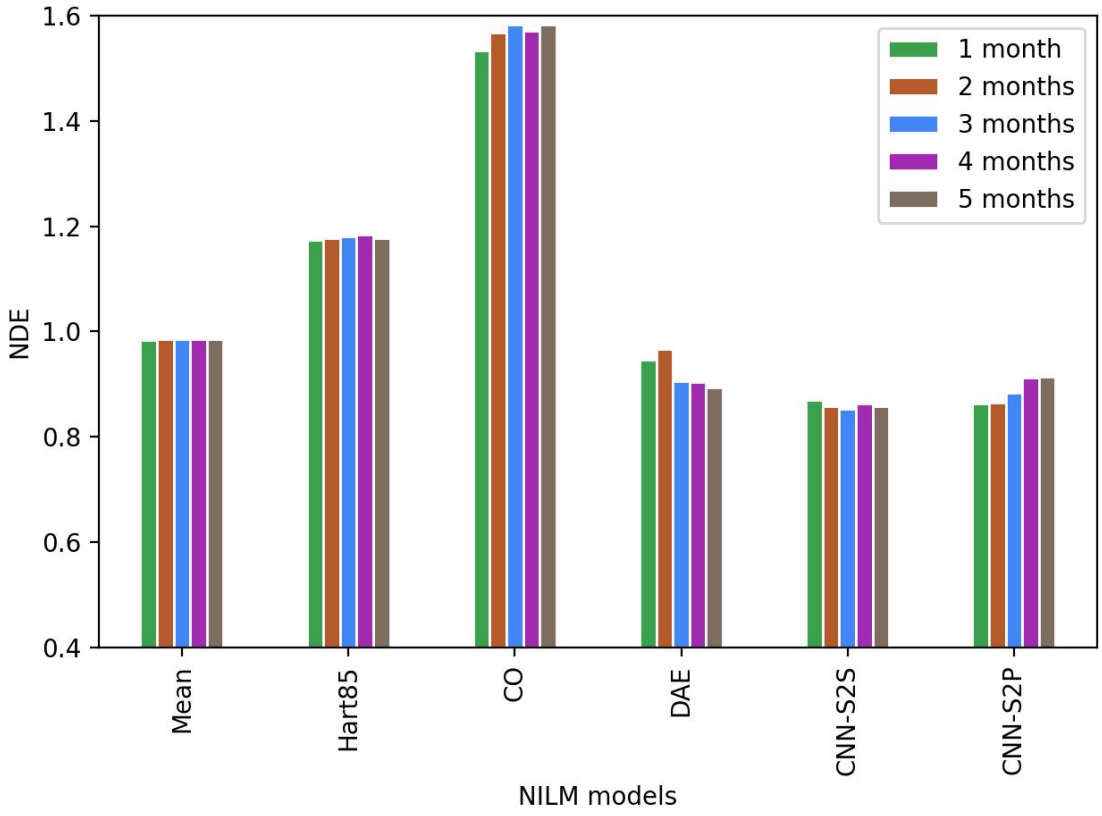


Figure 2.13. Impact of the training data length on the model performances

2.10 Conclusion

In this chapter, a series of experiments was conducted to understand the significant factors impacting the performances of NILM models. The scholarship thoroughly reviewed model type choice [17]; our experiment results align with this research, confirming DL models' superiority. In Section 2.5, results show better performances when using DL models (CNN-S2P, CNN-S2S and DAE). DL methods need prior fine-tuning to optimise the models; in this chapter, the influence of three hyperparameters was observed: the learning rate, the patience for early stopping and the window width. The latter is a crucial adjustable param-

eter in the NILM domain, as it has an incidence on both model complexity and model accuracy. Through fine-tuning Section 2.4, acceptable hyperparameter values were found to process further NILM experimentation. In Section 2.6, two configurations were compared to assess the accuracy decline when experimenting under generalised (*Unseen*) conditions. To determine the levers to unlock generalisation improvements, many points were investigated :

- The DL architecture (DAE, CNN-S2S, CNN-S2P)
- The number of houses comprised in the training set
- The sampling period
- Load complexity throughout the day and the week
- Training dataset length

The analysis revealed that load complexity has the highest impact on NILM performances with an NDE difference of about 0.68 (Figure 2.11(b)) between a disaggregation during a complex aggregated load and a simpler one. The CNN-S2P model showed an interesting improvement when adding variability (adding new houses) to the training set, reaching the best performances in *Unseen* conditions among all other models. Increasing the training dataset length does not bring apparent improvements. In some circumstances, this setting can negatively impact NILM quality. Considering all the analyses, low-frequency NILM in *Unseen* conditions can hardly replace a whole sensor network for the poor accuracy found. The *Seen* conditions gave better results but still showed poor results when the load became more complex. Three levers were identified to be relevant for NILM quality improvements :

- Adding diversity (new houses, new power levels of appliances, new patterns) in the training set
- Enriching the training set with data from periods when the aggregate load is more complex, corresponding to adding more appliance combinations
- Reducing the sampling period

The NILM cost-efficiency limits the latter lever. Still, the next chapter assesses to what extent data augmentation can activate other levers to enrich the training dataset for better NILM model performances.

CHAPTER 3

Data augmentation:Expanding the variety of NILM training data

This section is focused on data augmentation (DA) contribution to enhance NILM performances. The core idea behind DA is to enrich the training dataset to tackle data scarcity and quality issues. By introducing more data to the existing training set, a more extended dataset is available for training. The training set quality is aligned with the variety and reliability of the information encompassed. In NILM, only a few papers [26–28, 64] tackle DA, with promising results; nevertheless, there appear to be opportunities to strengthen the research on DA for NILM. This chapter revolves around the following research questions:

- Is DA an efficient method to enhance NILM generalisability?
- How effective is the OFFSETAUG method compared to existing DA techniques?
- Is DA relevant for very low-frequency data?

Indeed, thorough experiments must be conducted to demonstrate any positive effect of DA on NILM performances. In this section, a particular interest is given to the generalisability of NILM, which is supposed to be potentially enhanced with DA. Hence, *Unseen* experiments on two different datasets are carried out to compare configurations with and without DA, a methodology used in [26, 27] where authors proposed DA algorithms. A new DA algorithm is presented and compared with

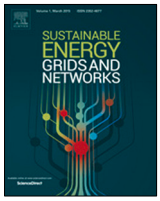
previously published methods in the present work. The paper entitled "Expanding the variety of non-intrusive load monitoring training data: Introducing and benchmarking a novel data augmentation technique", referenced [29], from the journal Sustainable Energy, Grids and Networks (SEGAN), is included into this chapter.

Sustainable Energy, Grids and Networks 35 (2023) 101142

Contents lists available at ScienceDirect

Sustainable Energy, Grids and Networks

journal homepage: www.elsevier.com/locate/segan



Expanding variety of non-intrusive load monitoring training data: Introducing and benchmarking a novel data augmentation technique

J. Francou *, D. Calogine, O. Chau, M. David, P. Lauret

PIMENT, University of La Réunion, Saint-Denis, 97715, Réunion

ARTICLE INFO

Article history:
Received 13 March 2023
Received in revised form 25 July 2023
Accepted 6 August 2023
Available online 9 August 2023

Keywords:
Non-intrusive load monitoring
Energy disaggregation
Deep learning
Data augmentation

ABSTRACT

Energy consumption monitoring is an important asset for demand side management systems. Although smart meters provide high-sample-rated and accurate measurements of various power variables measured on diverse appliances, their high prices restrain them from a large-scale deployment. Non-Intrusive Load Monitoring (NILM) reduces monitoring cost by disaggregating computationally a main measurement into appliance-level load measurements. Deep neural networks NILM techniques have shown higher performances when they are trained and tested on the same building. However, NILM approaches suffer from low generalization capabilities due to the challenging task of identifying a wide variety of load profiles for a few available supervised training data. To address the issues of data scarcity and low generalization capabilities, data augmentation has shown promising results. Data augmentation refers to technique of artificially increasing the size of the training data by creating transformed versions of the existing data. In this study, a new and easy-to-understand transformation technique was extensively tested to introduce variability into the training set. This technique is used for data augmentation and was compared with two other methods. The results of the experiments demonstrated improvements in both F1-score and NDE metrics for all data augmentation techniques when compared to the baseline case without any augmentation. Notably, the proposed data augmentation method (OFFSETAUG) showed even higher improvements than the other two algorithms. Based on these findings, it can be concluded that data augmentation is a straightforward and valuable resource to enhance the performance of low-frequency NILM tasks.

1. Introduction

The current energy crisis highlights the significant vulnerability of our fossil-fuel-dependent energy systems. In response, several governments have tightened energy policies and implemented awareness-raising programs to reduce energy use. Accelerating energy efficiency and improving building thermal performance are among the major actions to achieve a green and secure energy system. In the study [1], the author demonstrated that providing feedback on energy demand can result in a 15% decrease in energy consumption. A breakdown of the main electrical load can help users identify over-consuming items and target energy savings. However, demand monitoring incurs additional costs from sensor device purchasing, installation, and data processing. Monitoring every electrical appliance in a building is challenging, intrusive, and non-robust. In particular, for residential settings, intrusive monitoring is likely to generate reluctance among residents. Non-intrusive Load Monitoring (NILM) systems can help overcome traditional monitoring barriers [2]. NILM measures only the main load of a building and computationally disaggregates it to appliance-level profiles. NILM was first introduced by Hart [3] using an event-based approach that involves spotting on and off events with the help of transient-based analysis. An effective transient state identification requires a high sampling frequency. Although high frequency data are convenient to retain a more extensive feature range, high sampling rate smart meters are expensive. Consequently, numerous researchers have taken an interest in low-frequency NILM. There is no common agreement on the low-frequency threshold value, but most studies consider it lower than 1 Hz. With the advent of deep learning (DL) techniques, NILM has achieved promising performance on low-sampling data [4–10].

In many DL experiments involving NILM for residential, models are often tested on the same house used for training. However, in a real-world scenario, training data from the target house may not be available, making it essential to evaluate the generalization capabilities of NILM. Previous studies [4,6,9] have attempted to address this issue through generalized experiments. However, the results obtained from these studies have been found to vary depending on factors such as the length of the testing and training datasets, the selected appliances, and the chosen houses. In [4],

* Corresponding author.
E-mail address: yvon.francou@univ-reunion.fr (J. Francou).

https://doi.org/10.1016/j.segan.2023.101142
2352-4677/

a comprehensive benchmark was conducted, which included a generalization assessment using the DATAPORT dataset [11]. Classical deep learning NILM models such as the sequence-to-point (S2P) approach presented in [5] and reused in [10] resulted in a mean absolute error less than an averaged 30 W for the refrigerator at a 1-min sampling rate. The CNN proposed by [6] nearly achieved 8.00 W when trained on all houses of REFIT [12] at 8 s sampling rate. Although results are encouraging, it seems some experimental cases results can be further improved. So far deep -learning-based NILM methods have been very promising however they need a lot of data with variability to generalize well on unseen houses. [13,14] pointed out the lack of supervised data. To address this issue contribution of data augmentation (DA) is underlined. DA is a technique used in deep learning to artificially increase the size of the training set by creating transformed versions of the existing data to improve model generalization. In the realm of DL NILM, the core idea is to generate new synthetic datasets derived from available data. A relevant DA should be realistic (representative of the initial training data) while being a source of variety to provide new information to the models. For NILM purposes, various DA algorithms have been proposed. The basic process consists of artificially combining known data chunks to generate a synthetic dataset. For example, [8] proposed generating a synthetic aggregate data by randomly adding target appliance and distractor appliance (all appliances different from the target appliance) load profiles to the main load, while [15] suggested successively adding off and on profiles to an actual aggregate load. However, these methods may not consider the realistic aspect of the synthetic load as they certainly do not depict the behavior features. To address this issue, [13] proposed generating synthetic aggregate data by adding activation profiles based on real consumption scenarios from real survey reports. Similarly, the SynD dataset [16] was synthesized based on real consumption patterns. A generative adversarial network approach is presented in [14] to generate new data that preserves major features. To assess DA contribution behavioral information should still be derived only from the available data. During the NILM process, the available datasets are the aggregated and sub-metered training set, as well as the aggregated testing set. [17] proposed randomly adding activation profiles of the target appliance to the aggregated testing set to generate a new synthetic training set. The idea is to extract information from the main testing set to improve the learning process. This practical methodology gives interesting improvements in generalization capabilities. Although the previously presented works achieve good results, some of them diverge from the common definition of data augmentation by adding additional information, such as behavioral information [13] or information extracted from the test base history [17]. In [9], the importance of power levels and power shift features is underlined. Hence a training set containing the same power levels/shifts as the testing set is expected to build better models. However, the power levels from the main testing load are unlabeled and then barely exploitable without techniques from unsupervised NILM. DA is beneficial to reduce the well-known class-imbalance effect, seen in [8,18] in the NILM domain. Generally, appliances, except for refrigerators and freezers, are usually more off-state than on-state. To tackle this issue a classical approach is to oversample appliance on-activation. In this work, a random adaptation of the power levels is proposed. The goal is to enhance the diversity of the training set by introducing additional samples that encompass a broad range of power levels for the target appliances. It is anticipated that some of these new samples will closely resemble the tested load. This research focuses on evaluating the impact of the proposed DA technique on the generalization of a DL NILM system. To assess its contribution, the proposed DA method is compared with two other DA techniques found in existing literature, and with the baseline case where no DA is used. The study's major contributions include the following:

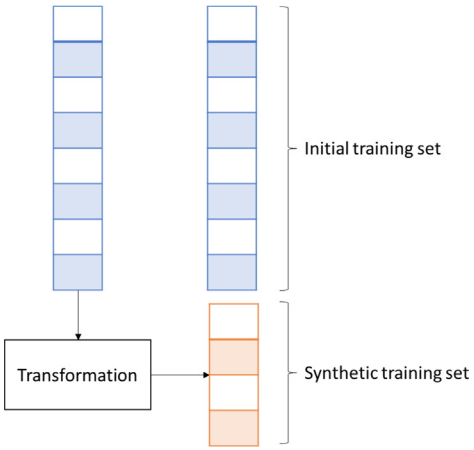


Fig. 1. Core principle of data augmentation.

- Benchmarking and comparing 3 straightforward DA techniques in the context of NILM.
- Evaluating the impact of DA on very low-frequency NILM tasks (intervals of at least 1 min).
- Developing and thoroughly assessing an easy-to-implement and straightforward DA algorithm.

By conducting these comparisons, this work aims to provide valuable insights into the effectiveness and practicality of DA for enhancing DL NILM performance and generalization.

The paper is organized as follows, Section 2 presents DL applications for NILM. Section 3 introduces the augmentation algorithms used. Finally 4 describes experimental configurations and results are exposed in 5.

2. Deep learning models for NILM

NILM consists in disaggregating a main load measurement, so a time series let it be $Y = (Y_1, Y_2, \ldots, Y_T)$ into appliance-level loads such as $X^k = (x_1^k, x_2^k, \ldots, x_T^k)$ for the appliance k. At each timestamp, $t \in [1, T]$, the aggregated signal can be defined by the equation,

$$Y_t = \sum_{k=1}^{K} x_t^k + \epsilon_t. \tag{1}$$

While the principle of NILM is defined by the equation,

$$\hat{X}^k = f^k(Y_t). \tag{2}$$

NILM models, noted f^k, aims at evaluating X^k values knowing only the main load time series Y. K is the set of all measured appliances and ϵ_t is a noise value. Deep learning models are very common for NILM, due to their promising performances. To train f^k to identify the load of an appliance k the supervised training is processed on a dataset where Y and X^k are fully known. Developing a specific model for each appliance is a frequent practice [6,8], as in available training datasets labeling every single appliance within the household is very challenging. A promising approach is the sequence-to-point (S2P) which is designed to predict the window midpoint. S2P is defined as,

$$f_{s2p}^k(Y_{t:t+W_{in}-1}, \theta) = \hat{X}_\tau^k, \tag{3}$$

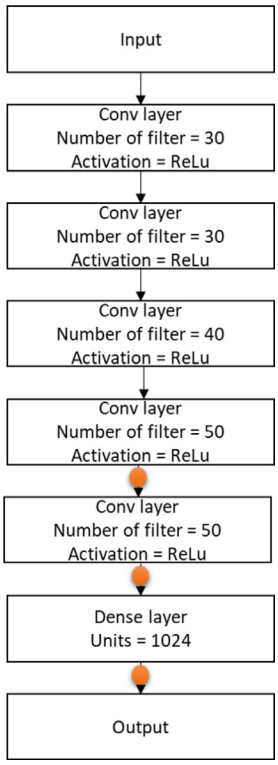


Fig. 2. Architecture of S2P approach, orange circles represent the dropouts.

where $\tau = W_{out}/2$. W_{out} and W_{in} being the output and input window lengths. The idea behind the S2P configuration is that the midpoint of the output window can be explained by the values before and after. The S2P architecture used for the study is presented in Fig. 2.

3. Data augmentation algorithms

The principle of DA is presented in Fig. 1, which consists of generating a new synthetic dataset from a transformation of the actual training set. Some transformations found in the domain consist in adding activation profiles to the actual dataset [8], combining on-period and off-period chunks [15] or generating a new dataset from a generative adversarial network [14]. In this work, a straightforward DA algorithm approach, noted OFFSETAUG, is proposed to generate a rich synthetic dataset. The DA algorithm is described in Fig. 3. STEP 1 - When training NILM models, the only known data are the main and the sub-meter loads from the training set. An algorithm inspired from *get_activations*() function from the work [8], is used to spot on-periods on the appliance-level load. Hence a database composed of on-activations is generated. This process is common for DA in the NILM domain such as in [8] or [17]. STEP 2 - Isolated profiles are randomly selected within the previously generated database. The number of selections is user-defined. STEP 3 - The major novelty of this work is the high power level offsetting process of the isolated profiles. To process offsetting, a distinction must be made between high power and low power levels, see third step in Fig. 3. The threshold, noted *thr*, is user-defined according to the target appliance type; in this work, $thr = 300W$ is chosen based on a beforehand trial and error test. The next operation consists in offsetting the high power levels with Eq. (4) inspired by the noise definition in the work [19],

$$\begin{cases} x_t^k = x_t^k + \delta p \text{ if } x_t^k > thr \\ x_t^k = x_t^k \text{ otherwise} \end{cases} \quad (4)$$

$\Delta P = [\delta p, \delta p, \ldots \delta p]$ operates as an offset. A unique ΔP is randomly selected for each isolated profile. The selection process follows a Gaussian distribution with μ and σ respectively the mean and the standard deviation. The idea behind the method is to artificially emphasize variability over the isolated profile database. The power variation is applied only on high power levels to limit noise emphasizing and to avoid negative values. Commonly high power levels on on-activations correspond to resistor loads, the power level can vary according to the type of device, the brand, the weather or the operation program chosen by the dweller for complex appliances. Consequently, the method is expected to work well on appliances composed with resistor loads (dishwasher, washing machine, oven, heater ...) as randomly offsetting would mimic those power level variations. Values of μ and σ are defined by the user, they should be cautiously chosen in order to avoid unrealistic profiles. Note that offsetting preserves the pattern as it results in a vertical translation of the initial profile. Applying successive offsets to the isolated profiles is presumed to make the models less dependent on power values and more on patterns. Moreover, this method would operate as a regularizer for DL models reducing overfitting risk. STEP 4 - Once the isolated profiles are processed, they are added iteratively to the main and the appliance-level training set for off-period chunks, in Fig. 5 a new combination is created with an air conditioner activation and a main set chunk. During the preprocessing stage before model training, a subset of windows resulting from the sliding windows process is selected from the augmented data. This subset specifically contains new combinations of "on" and "off" states. However, the "off" windows in the augmented dataset are excluded from the training process since they already exist in the original training set. The focus is on incorporating novel variations of "on" and "off" combinations to enrich the training set and improve model generalization.

The DA algorithm is compared to existing methods. The first method (AUG) is similar to the one proposed in this work except no offsetting ($\Delta P = [0, 0, 0, \ldots, 0]$) is used. In NILM literature, AUG is close to DA applied in [8,15]. Adding noise to the training data is generally a classical data augmentation process, in [19], a noise is added to the main dataset following the equation,

$$Y_{noise} = Y + n_f * N(0, 1) \quad (5)$$

Y_{noise} and Y are respectively the main training set with noise addition and its initial value. n_f is the noise factor defined in [19] to define the percentage of added noise. In this work, a $n_f = 1$ value is considered. The noise $N(0, 1)$ follows a Gaussian Distribution with $\mu_{NOISEAUG} = 0$ and $\sigma_{NOISEAUG} = 1$. In this work, this approach is noted NOISEAUG. All augmentation methods, see Table 1, are compared to the baseline case, NOAUG, in which the model is trained without augmenting the training data.

4. Experimentation

The DA technique is applied on a wide-spread NILM model, the S2P (3). The first experiment is to assess the efficiency of the proposed DA algorithm for NILM generalization, comparing the baseline NILM to the NILM enhanced with the DA technique. The second experiment consists of evaluating the ability of DA to improve models for higher sampling period data. In this section, the carried-out experiments are described.

4.1. Data preprocessing

Despite the availability of several Non-Intrusive Load Monitoring (NILM) datasets, they may not be sufficient to address the high variability of existing load profiles, which consist of a multitude of appliance types with countless potential appliance

Table 1
Overview table of the DA methods.

Designation	Details	References
NOAUG	No data augmentation	
AUG	DA by combining ON and OFF periods	[8] and [15]
OFFSETAUG	DA with the proposed method by offsetting ON periods and by combining ON and OFF periods	
NOISEAUG	DA by adding noise to the main training set	[19]

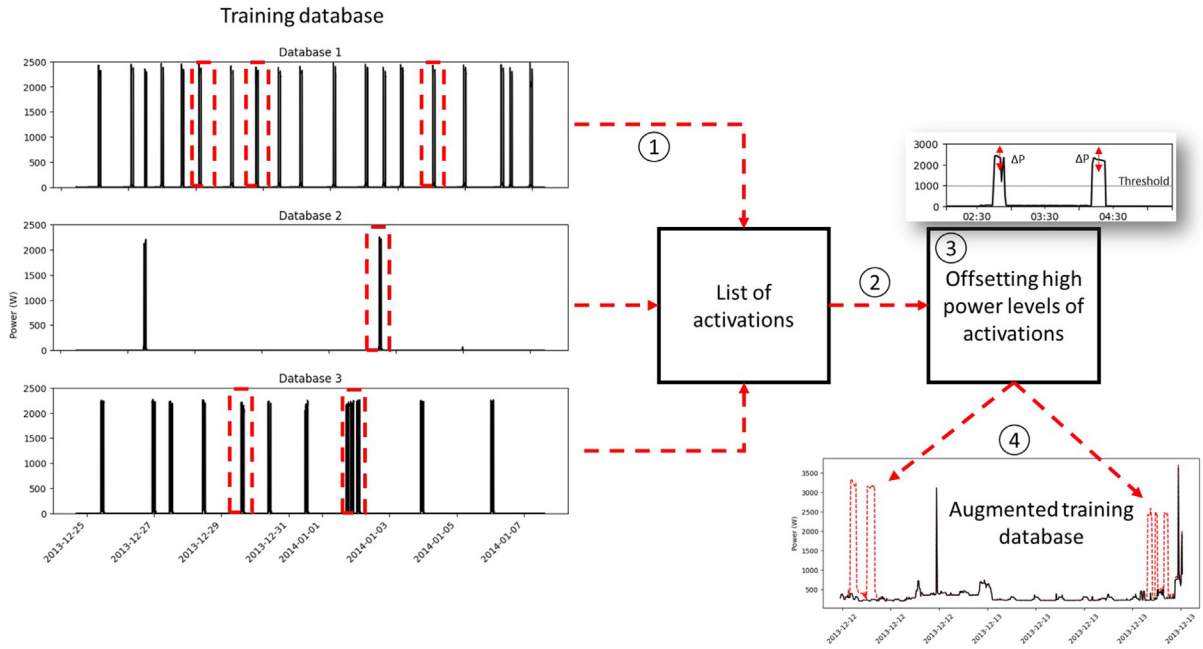


Fig. 3. Data augmentation through synthetic dataset generation using a high power offsetting block.

combinations. Monitoring residential loads is particularly challenging due to the intrusion of sensors capable of tracing personal data. REFIT dataset, described in [12] is composed of 21 houses in the United Kingdom with various occupancies and various numbers of appliances recorded. REFIT database has one of the longest monitoring durations of about 2 years with an 8 s sampling period. To achieve reliable deep learning NILM experiments, the data must be well inspected before processing. In the below experiments, the data are downsampled to 1 min like in the experiments in [4], to reduce noises, to alleviate computational resources, and to assess the generalization capabilities on very low-frequency NILM tasks. The input sequence W_{in}, chosen for the experiment is 99 samples for the washing machine and the dishwasher detection. Whereas a slightly wider window, 119 samples, is chosen for the electric furnace and the air conditioner disaggregation as those appliances generally run for a longer period. The latter two appliance profiles are extracted from the DATAPORT dataset [11]. Data standardization is processed as follows,

$$\bar{x}_i = \frac{x_i - \mu_x}{\sigma_x}. \tag{6}$$

The values of μ_x and σ_x are in Table 2. The DA processes have been incorporated into the NILMTK (Non-Intrusive Load Monitoring Toolkit) framework [20].

4.2. Experimentation settings

To comprehensively evaluate the generalization capabilities a leave-one-out looped experimentation is processed on the data set as illustrated in Fig. 4. For evaluating a model's performance

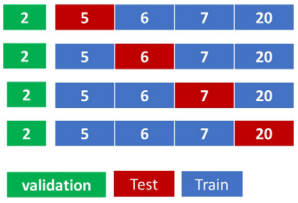


Fig. 4. Successive leave-one-out on some REFIT houses for a comprehensive assessment.

on a given house, the model is trained on all the other houses. Since the DA approaches are random, each experiment is repeated 3 times to obtain aggregate results. 0.5% of the whole training window samples results from DA are used for training, as seen in [14], the length of the augmented set is kept relatively low. The models are trained with 120 days from each training house, plus the synthetic set when DA is applied, while tested on a 30-day chunk of the tested household. The ADAM optimizer presented in [21] is used to train the model. Dropouts of 20% are used for training. Dropout layers are represented in Fig. 2. Main experimentation settings are in the Table 2. To reduce overfitting risks during the training phase, the model is monitored over a certain number of epochs. If there is no improvement in the model's performance over the last 10 epochs, the training is stopped and the best previous weights are retrieved regarding the validation loss. For the experimentation, a single-target NILM model approach is applied, hence a unique model is trained for each appliance type.

Table 2
Experimentation settings for performance assessment.

	Dishwasher	Washing machine	Air conditioner	Electric furnace
Mean for standardization (W)	700	400	1500	1500
Standard deviation for standardization (W)	1000	700	1800	1000
Input window size (min)	99	99	119	119
Sample period (s)	60	60	60	60
Maximum epoch	100	100	100	100
Batch size	1000	1000	1000	1000
Patience of early-stopping (epochs)	10	10	10	10
Learning rate	0.001	0.001	0.001	0.001
Dataset	REFIT	REFIT	DATAPORT	DATAPORT
Training and testing on houses	5,6,7,17,20	5,6,7,17,20	1642, 2335, 3039, 8386	661, 1642, 2335, 8386, 9160
Validation house	2	2	2818	2818
Training length (days/per house)	120	120	120	120
Testing length (days)	30	30	30	30
Validation length (days)	120	120	120	120
Offset random selection parameters (μ, σ)	(0,20)	(0,20)	(0,20)	(0,20)

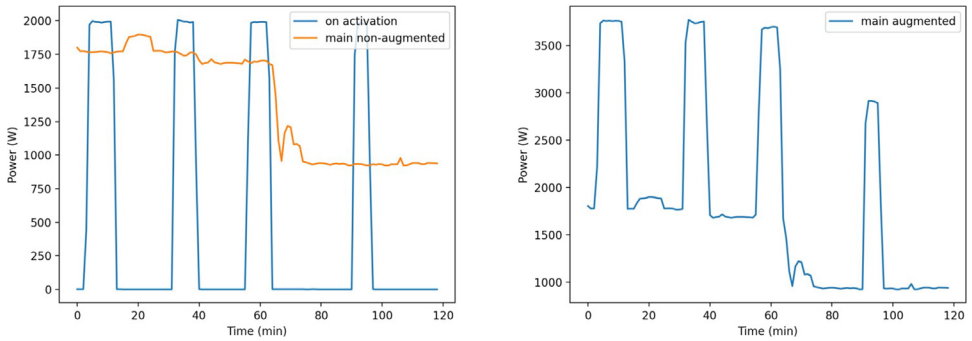


Fig. 5. Representation of AUG process for air conditioner detection. Figure (a) overlaps a randomly selected on-activation submeter and an off-activation main set chunk. Figure (b) is the addition of both latter, an augmented main load chunk.

5. Results and discussions

5.1. Metrics

As per [22,23], the evaluation of load disaggregation can be broken down into two tasks, event detection and energy estimation. For event detection, the main metric considered in this paper is the f1-score defined in Eq. (7) calculated with the True Positive Rate (TPR), see Eq. (9) and the False Positive Rate (FPR), see Eq. (10). As NILM is originally a regression problem, a user-defined threshold separates on-events from off-events. In this paper, three different thresholds are selected to evaluate the F1Score metric, at 25 W, 100 W and 500 W. The thresholds are used to compute the confusion matrix variables in Fig. 6. The Mean Absolute Error (MAE) is used to assess an average power deviation between ground truth and prediction, although MAE metric comparison can be biased by low-amplitude noises when the device is not operating and by the operating power level. Then, the Normalized Disaggregation Error (NDE), see Eq. (12), is applied in this work.

- F1-score

$$\text{F1Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{7}$$

- Precision

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

- Recall or sensitivity or True Positive Rate (TPR)

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

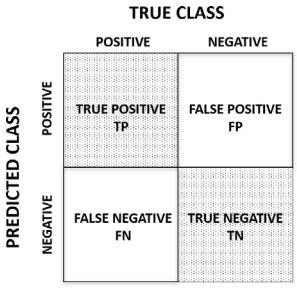


Fig. 6. Confusion matrix.

- False Positive Rate (FPR) or fall-out

$$\text{FPR} = \frac{FP}{FP + TN} \tag{10}$$

- Mean Absolute Error

$$\text{MAE} = \frac{1}{T} \sum_{t=1}^{T} |\hat{x}_t^k - x_t^k| \tag{11}$$

- Normalized Disaggregation Error

$$\text{NDE} = \frac{\sum_{t=1}^{T} (x_t^k - \hat{x}_t^k)^2}{\sum_{t=1}^{T} (x_t^k)^2}, \tag{12}$$

- ΔNDE

$$\Delta\text{NDE} = \text{NDE}_{NOAUG} - \text{NDE} \tag{13}$$

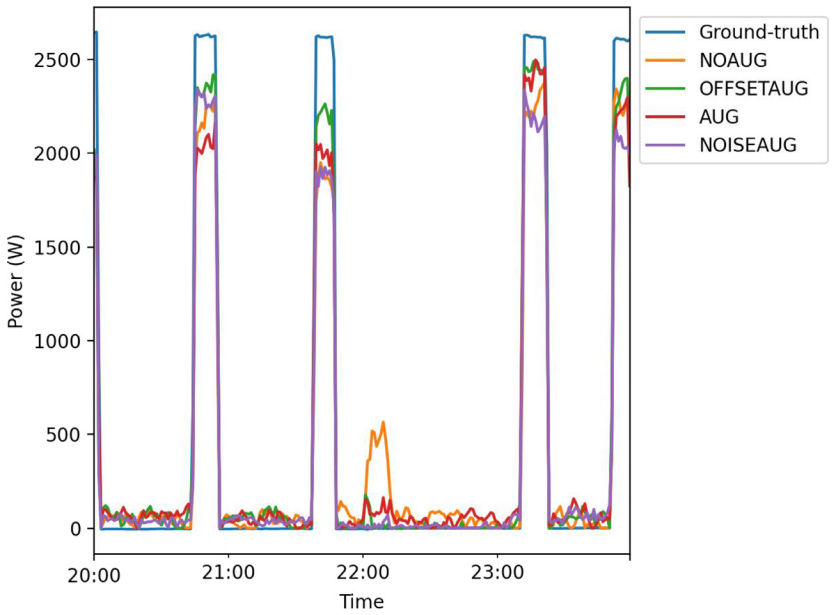


Fig. 7. Output curves for air conditioner power estimation compared to the ground truth on the house 9160 from DATAPORT.

Table 3
Cross validation results per appliance, F1score label is followed by the considered on-off threshold, MAE in W.

Air conditioner	F1score500	F1score100	F1score25	MAE	NDE
AUG	0.7003	0.6685	0.4645	217.1415	0.6423
OFFSETAUG	**0.7219**	**0.6930**	0.5107	**203.9826**	**0.6139**
NOAUG	0.6707	0.6356	**0.5337**	218.2975	0.6370
NOISEAUG	0.6844	0.6480	0.4920	204.6370	0.6389
Dish washer	F1score500	F1score100	F1score25	MAE	NDE
AUG	0.7398	0.5713	0.3463	51.0791	**0.5870**
OFFSETAUG	**0.7792**	**0.5749**	0.2978	48.0004	0.6067
NOAUG	0.6726	0.5030	0.2329	54.4085	0.6542
NOISEAUG	0.6977	0.5331	**0.3472**	**46.1146**	0.6519
Electric furnace	F1score500	F1score100	F1score25	MAE	NDE
AUG	0.2515	0.4247	0.5111	196.4844	0.9325
OFFSETAUG	0.2640	**0.4482**	**0.5518**	193.0571	**0.8959**
NOAUG	**0.2913**	0.4205	0.5171	**186.4006**	0.9313
NOISEAUG	0.2904	0.4198	0.5263	192.2699	0.9096
Washing machine	F1score500	F1score100	F1score25	MAE	NDE
AUG	0.3357	0.2605	0.2324	66.2479	0.9548
OFFSETAUG	**0.3806**	**0.2734**	0.2172	56.4337	**0.9319**
NOAUG	0.3182	0.2179	**0.2752**	**42.9283**	0.9368
NOISEAUG	0.2719	0.1920	0.1887	53.9879	0.9995

y_t^k and $\hat{y}_t^k$ are respectively the ground truth and the predicted load for the appliance k. TN, TP, FP and FN are defined in the confusion matrix in Fig. 6.

5.2. DA contributions

In Table 4, the results demonstrate the effectiveness of 2 DA techniques, AUG and OFFSETAUG. For the proposed OFFSETAUG, on average, there is a 10% increase in F1-score500 metric, a 12% increase in F1-score100 metric and a slight 1% increase in F1-score25 value. Although the results for AUG experiments have shown fewer improvements, its contribution worth to be underlined, a 4% and a 8% increases are respectively observed on F1-score500 and F1-score100. The improvements regarding the F1-score brought about by both techniques are likely due to the extra information provided by the transformation, as AUG and OFFSETAUG consist in injecting new combinations that were not seen in the actual training set. Seeing only the overall results in Table 4, NOISEAUG seems under performing comparing to the benchmark (NOAUG) in terms of F1-score and NDE. However, when considering the details in Table 3, it can be seen NOISEAUG is able to perform on air conditioner, dish washer and electric furnace detections. Serious underperformances are observed on the washing machine power regression for NOISEAUG. This is likely due to the load shape, when the rotor is operating an inherent noise is visible on the profile. This noise is believed to be discriminant for washing machine. Thus transforming the data with a Gaussian noise addition to create a synthetic training set, as done for NOISEAUG, can be a source of confusion for washing machine identification. In the field of NILM, MAE and NDE are standard metrics. AUG and OFFSETAUG has shown promising improvements in NDE metrics, as a NDE decrease is observed for both, a sharper NDE decline is seen for OFFSETAUG. Observation of the estimated power loads in Fig. 7 supports the performance of AUG and OFFSET configurations. When the appliance is operating, OFFSETAUG is often the closest to the actual power level. As may seen around 22:00 in Fig. 7, a false detection is observed for NOAUG configuration. This false detection is not retrieved

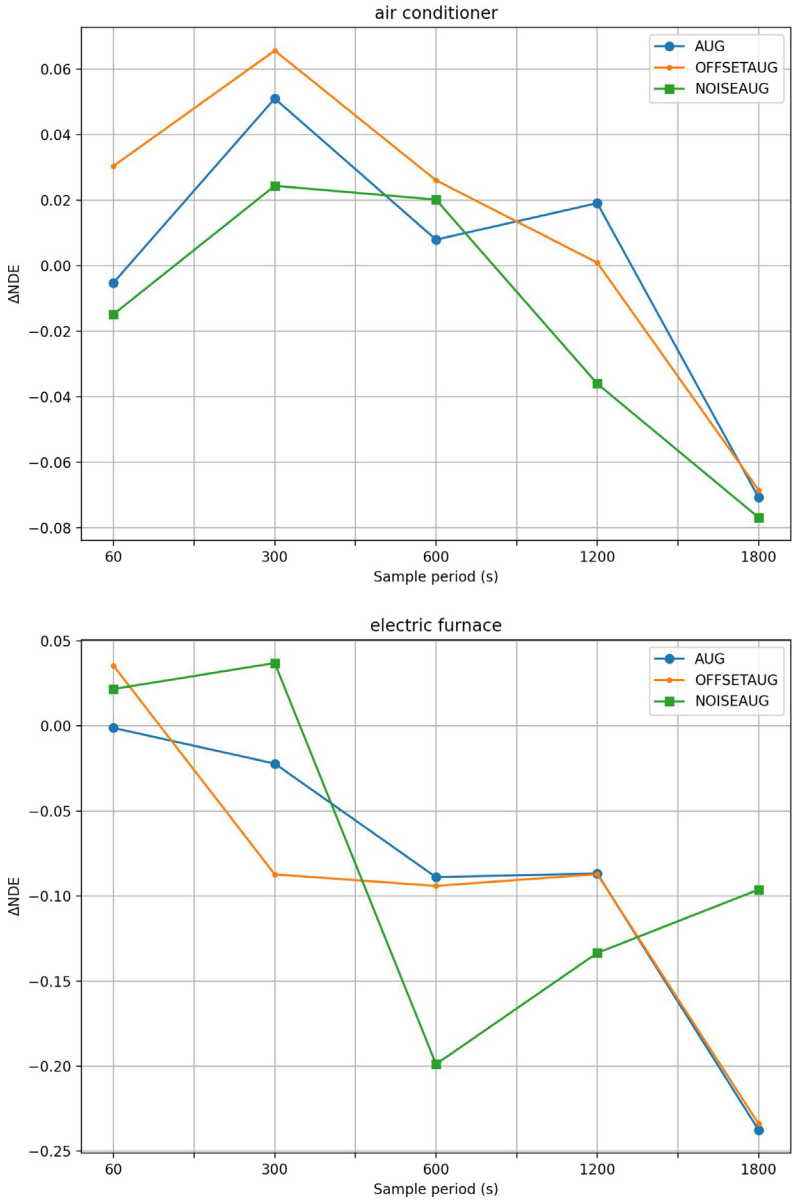


Fig. 8. Sample period effect on DA improvement. ΔNDE is the NDE difference when the model is trained without augmentation and when trained with augmentation. A ΔNDE > 0 indicates an improvement compared to the baseline.

Table 4
Overall results.

	F1score500	F1score100	F1score25	MAE	NDE
AUG	0.5068	0.4812	0.3886	132.7382	0.7791
OFFSETAUG	**0.5364**	**0.4974**	**0.3944**	125.3685	**0.7621**
NOAUG	0.4882	0.4442	0.3897	125.5087	0.7898
NOISEAUG	0.4861	0.4482	0.3886	**124.2524**	0.8000

on the 3 DA experiments. In general, using DA in data preprocessing phase permits a training data enriching. For operational implementation, DA augmentation can be beneficial when NILM is used on the edge or on the cloud during the training-phase of the model to deploy. The synthetic expanding of the training set does involve insignificant preprocessing surplus times, making it a lightweight add-on to optimize NILM model training.

5.3. DA contribution on higher sampling periods

NILM researchers and manufacturers are particularly interested in the ability of NILM to perform at lower sampling periods. Greater granularity of data means higher operational implementation and maintenance costs [24]. DA for NILM has been assessed for high frequency data [25], although in this study, the contribution of DA for lower sampling period has been explored. To quantify the contribution, the ΔNDE metric defined in Eq. (13) is used as it reveals a direct improvement whenever ΔNDE> 0. The Figs. 8 represent the evolution of ΔNDE in relation to the sampling rate considering the air conditioner and the electric furnace. All three DA algorithms have brought an enhancement on air conditioner until the 10-min-sample-period. AUG and OFFSETAUG still contribute until 20 min. However, for the electric

furnace, only the NOISEAUG can perform at a higher sample period until 5 min. Reducing sampling rate causes a pattern fading, hence it is supposed the electric furnace discriminant patterns are discernible up to 5-min-sampling-period.

6. Conclusion and perspectives

The emergence of DL NILM models has highlighted the issue of data scarcity in this field [25,26]. While one solution would be to monitor as many buildings as possible, this is difficult to achieve comprehensively. Instead, some researchers propose to enhance DL NILM model training with DA. In this work, a new DA algorithm, OFFSETAUG, is proposed. It consists in generating new data and data combinations. This involves isolating the operating periods of the target device from the known training set, adding an offset to the high power levels, and randomly adding them to a duplicate chunk of the actual training dataset. The novelty of this work lies in the offsetting process, which involves applying a straightforward offset to each isolated profile's high power levels to maintain realistic patterns for NILM tasks. To ensure power level variability, offset values are randomly selected according to a Gaussian distribution. In this study, the proposed DA approach has been thoroughly evaluated, and it has shown clear improvements in f1-score and NDE on 4 common appliances when compared to two existing DA algorithms and to the baseline case. Varying power levels during the training phase has resulted in better power estimation and improved event detection capabilities. This paper evaluates the contribution of DA in relation to the sampling period. Simulations demonstrated DA is valuable for low-frequency data up to a specific sampling period dependent on the appliance profile. Although DA techniques for NILM are becoming a significant research topic, this is currently, as far as the author knows, the first benchmarking study for comparing DA techniques in NILM. Therefore, future research should address this lack of benchmarking to seek to answer the question what is the more convenient DA method for NILM?

CRediT authorship contribution statement

J. Francou: Carried out the experiment, Wrote the manuscript. **D. Calogine:** Helped supervise the project. **O. Chau:** Helped supervise the project. **M. David:** Helped supervise the project. **P. Lauret:** Helped supervise the project.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] Corinna Fischer, Feedback on household electricity consumption: A tool for saving energy? Energy Efficiency 1 (1) (2008) 79–104.

[2] Yu Hsiu Lin, Men Shen Tsai, An advanced home energy management system facilitated by nonintrusive load monitoring with automated multiobjective power scheduling, IEEE Trans. Smart Grid 6 (4) (2015) 1839–1851.

[3] G.W. Hart, Non-intrusive appliance load monitoring, Proc. IEEE 80 (12) (1992) 1870–1891.

[4] Nipun Batra, Rithwik Kukunuri, Ayush Pandey, Raktim Malakar, Rajat Kumar, Odysseas Krystalakos, Mingjun Zhong, Paulo Meira, Oliver Parson, Towards reproducible state-of-the-art energy disaggregation, 2020, 2019, pp. 193–202.

[5] Chaoyun Zhang, Mingjun Zhong, Zongzuo Wang, Nigel Goddard, Charles Sutton, Sequence-to-point learning with neural networks for non-intrusive load monitoring, vol. 2018, in: 32nd AAAI Conference on Artificial Intelligence, AAAI, 2018, pp. 2604–2611.

[6] David Murray, Lina Stankovic, Vladimir Stankovic, Srdjan Lulic, Srdjan Sladokevic, Transferability of neural network approaches for low-rate energy disaggregation, in: ICASSP 2019, Vol. 1, no. 1, 2019, pp. 8330–8334.

[7] Xiao Zhou, Shujian Li, Chengxi Liu, Haojun Zhu, Nan Dong, Tianying Xiao, Non-intrusive load monitoring using a CNN-LSTM-RF model considering label correlation and class-imbalance, IEEE Access 9 (2021) 1.

[8] Jack Kelly, William Knottenbelt, Neural NILM : Deep neural networks applied to energy disaggregation, 2012, pp. 55–64.

[9] Michele D'Incecco, Stefano Squartini, Mingjun Zhong, Transfer learning for non-intrusive load monitoring, IEEE Trans. Smart Grid 11 (2) (2020) 1419–1429.

[10] Wenpeng Luan, Ruiqi Zhang, Bo Liu, Bochao Zhao, Yixin Yu, Leveraging sequence-to-sequence learning for online non-intrusive load monitoring in edge device, Int. J. Electr. Power Energy Syst. 148 (92) (2023) 108910.

[11] Oliver Parson, Grant Fisher, April Hersey, Nipun Batra, Jack Kelly, Amarjeet Singh, William Knottenbelt, Alex 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, IEEE, 2016, pp. 210–214.

[12] David Murray, Lina Stankovic, Vladimir Stankovic, An electrical load measurements dataset of United Kingdom households from a two-year longitudinal study, Sci. Data 4 (2017) 1–12.

[13] Aurélien Delfosse, Georges Hebrail, Aimen Zerroug, Deep learning applied to nilm: Is data augmentation deep learning applied to nilm: Is data augmentation, Front. Artif. Intell. Appl. 325 (2020) 2972–2977.

[14] Alon Harell, Richard Jones, Stephen Makonin, Ivan V. Bajic, TraceGAN: Synthesizing appliance power signatures using generative adversarial networks, IEEE Trans. Smart Grid 12 (5) (2021) 4553–4563.

[15] Hasan Rafiq, Xiaohan Shi, Hengxu Zhang, Huimin Li, Manesh Kumar Ochani, Aamer Abbas Shah, Generalizability improvement of deep learning-based non-intrusive load monitoring system using data augmentation, IEEE Trans. Smart Grid 12 (4) (2021) 3265–3277.

[16] Christoph Klemenjak, Christoph Kovatsch, Manuel Herold, Wilfried Elmenreich, A synthetic energy dataset for non-intrusive load monitoring in households, Sci. Data 7 (1) (2020) 1–17.

[17] Weicong Kong, Zhao Yang Dong, Bo Wang, Junhua Zhao, Jie Huang, A practical solution for non-intrusive type II load monitoring based on deep learning and post-processing, IEEE Trans. Smart Grid 11 (1) (2020) 148–160.

[18] Haiping Wu, Hui Liu, Non-intrusive load transient identification based on multivariate LSTM neural network and time series data augmentation, Sustain. Energy Grids Netw. 27 (2021) 100490.

[19] Nikolaos Virtsionis Gkalinikis, Christoforos Nalmpantis, Dimitris Vrakas, Torch-NILM: An effective deep learning toolkit for non-intrusive load monitoring in Pytorch, Energies 15 (7) (2022).

[20] Nipun Batra, Jack Kelly, Oliver Parson, Haimonti Dutta, William Knottenbelt, Alex Rogers, Amarjeet Singh, Mani Srivastava, NILMTK: An open source toolkit for non-intrusive load monitoring, in: E-Energy 2014 - Proceedings of the 5th ACM International Conference on Future Energy Systems, 2014, pp. 265–276.

[21] Thomas Moreau, Mathurin Massias, Alexandre Gramfort, Pierre Ablin, Pierre-Antoine Bannier, Benjamin Charlier, Mathieu Dagréou, Dupré.la Tour, Ghislain Durif, Cassio F. Dantas, Quentin Klopfenstein, Johan Larsson, En Lai, Tanguy Lefort, Benoit Malézieux, Badr Moufad, Binh T. Nguyen, Alain Rakotomamonjy, Zaccharie Ramzi, Joseph Salmon, Samuel Vaiter, Benchopt: reproducible, in: Efficient and Collaborative Optimization Benchmarks, Vol. 2022, NeurIPS, 2022, pp. 1–43.

[22] Lucas Pereira, Nuno Nunes, Performance evaluation in non-intrusive load monitoring: Datasets, metrics, and tools—a review, Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 8 (6) (2018) 1–17.

[23] Ebony T. Mayhorn, Joseph Petersenand Ryan S. Butner, Erica M. Johnson, Load disaggregation technologies: Real world and laboratory performance, in: ACEEE Summer Study on Energy Efficiency in Buildings, 2016, pp. 1–13.

[24] Christos L. Athanasiadis, Theofilos A. Papadopoulos, Dimitrios I. Doukas, Real-time non-intrusive load monitoring: A light-weight and scalable approach, Energy Build. 253 (2021) 111523.

[25] Mohamad Nour, et al., Data augmentation strategies for high frequency NILM datasets, IEEE Trans. Instrum. Meas. 72 (2023) 1–9.

[26] Tai Le Quy, Sergej Zerr, Eirini Ntoutsi, Wolfgang Nejdl, Data augmentation for dealing with low sampling rates in NILM, 2021.

CHAPTER 4

NILM contribution to HEMS

A deep feeling of energy supply vulnerability has become apparent in society. In residential areas, some households (mainly middle-class households) are under uncomfortable pressure due to a constant rise in energy bills. To face the inevitable price increase, some actions an user may apply to cut his electricity bill are to reduce energy use, limit energy waste, or avoid consuming during peak periods when the cost of electricity is high. All these solutions belong to the Demand-Side Management (DSM) strategies. Generally, DSM is implemented by the grid operator to manage the energy demand. The goal is to adjust energy usage and then modify the load shape to enhance grid stability, increase cost-effectiveness for utility companies, or lower household electricity bills. The most common load shaping techniques involve [65, 66]:

- Load building: increasing the load throughout the day by increasing the demand. It could be surprising to present a means to increase consumption, but it may be needed for the cases where the grid is overproducing; augmenting the demand is a lever to keep grid stability.
- Load shifting: during peak demand, grid load is reduced while load building is applied during off-peak periods.
- Load shedding: an intentional process of temporarily switching off certain energy consumer regions, buildings or appliances.
- Peak clipping or peak shaving: any actions to reduce the peak demand.
- Valley filling: consists of increasing the load during periods of low energy demand to obtain a flat demand profile to ease grid management.
- Conservation: reduces general consumption by replacing existing devices with more energy-efficient ones.

DSM has two main temporalities: long and short terms [67]. Long-term programs refer to improving general energy efficiency, which refers to using technology to

provide the same or better energy services by consuming less energy [68]. For instance, using LED lighting instead of incandescent lamps is an energy-efficient alternative [69] involved in conservation strategies. The government's energy policies in France encourage DSM, for instance, through the CEE (Certificats d'économie d'énergie) mechanism, easing the funding of more efficient devices [70]. The idea is to alleviate global energy consumption and to restrain the energy demand increase. DSM can also be applied on a short-term horizon basis by implementing Demand Response (DR) techniques. The actionable levers for DR implementation are mainly financial, whether for load curtailment strategies or dynamic pricing programs [66]. The purpose of load curtailment is for the grid operator to curtail specific loads in exchange for payment from offloaded customers. Load curtailment is generally applied for high-consuming customers, such as factories. An example of the need for curtailment is when the load is nearly above the limit where an additional power plant needs to be started; instead, curtailment can be applied to reduce the global grid operation cost. For residential applications, a dynamic tariff is preferred. The concept is to encourage customers to consume during off-peak hours by implementing lower electricity prices during these periods. It has been demonstrated incentive pricing is critical to keep the balance between production and consumption [71, 72]. Among the most common dynamic pricing mechanisms:

- Time of Use (TOU); the pricing structure is divided into two periods: a peak period and an off-peak period. The electricity price varies based on the time of day, with higher prices during peak periods and lower prices during off-peak periods.
- Real Time Pricing (RTP); the price changes dynamically hourly. The customers are notified regarding the rates generally one day ahead.
- Critical-Peak Pricing (CPP); this pricing structure fixes a very high electricity price when the grid encounters a critical state for grid stability. Generally, the customer is alerted quickly (day or hour) ahead of the critical event.
- Peak-Time Rebate (PTR); This pricing structure is designed to reward customers with rebates when they reduce their peak loads.

DR applies to industries, particularly heavy industries [10, 73] such as cement factories for significant peak shavings. In the commercial sector, electricity is used for heating/cooling, refrigerating, lighting or running any sort of electronic device (computers, printers, ...). The commercial sector allows the implementation of many DR techniques [10], which can be load shifting, i.e. delaying the loads when electricity

is cheaper or when renewable sources are available. Demand response (DR) mechanisms in industrial and commercial sectors should not disturb the existing work organisation with a risk of reducing productivity. The ultimate goal of DR should be to optimise energy consumption and reduce costs without negatively impacting the daily operations of businesses and industries. Similarly, in the residential sector, load shape modifications have to be aligned with the willingness and comfort of each individual. It is hardly possible to shed the television operation while the dweller is watching. Hence, DR has to be aligned with the dwellers' comfort [74–76]. The variety in schedules, with the highly personal comfort aspect, makes a standard DR less effective for an individual household, a customised DR is needed for each household. Significant progress has been made in the concept of Home Energy Management Systems (HEMS) [12]. HEMS refers to various strategies used to influence and manage the patterns of an individual household's electricity demand. This chapter explores how NILM can contribute to implementing HEMS and deals with the following research question:

- To what extent can NILM enhance HEMS?
- If it contributes, how far can a NILM trained with DA overpass a standard NILM?

To answer the question, the methodology for evaluating the effectiveness of NILM is addressed, and a series of experiments are undertaken to compare a NILM-equipped HEMS with both a HEMS without ALM system and a HEMS with ILM. The contribution of DA has been demonstrated in Chapter 3, particularly when it comes to improving NILM trained in a generalised configuration (i.e. deployed on a never-seen house). So, this chapter also evaluates the performance of an augmented NILM for a HEMS.

4.1 Case study: a day-ahead load scheduler

HEMS is a complex tool that automatically manages the energy demand within an individual household [12, 77]. This section defines this concept and provides a practical description of its application. The Figure 4.1 depicts a schematic view

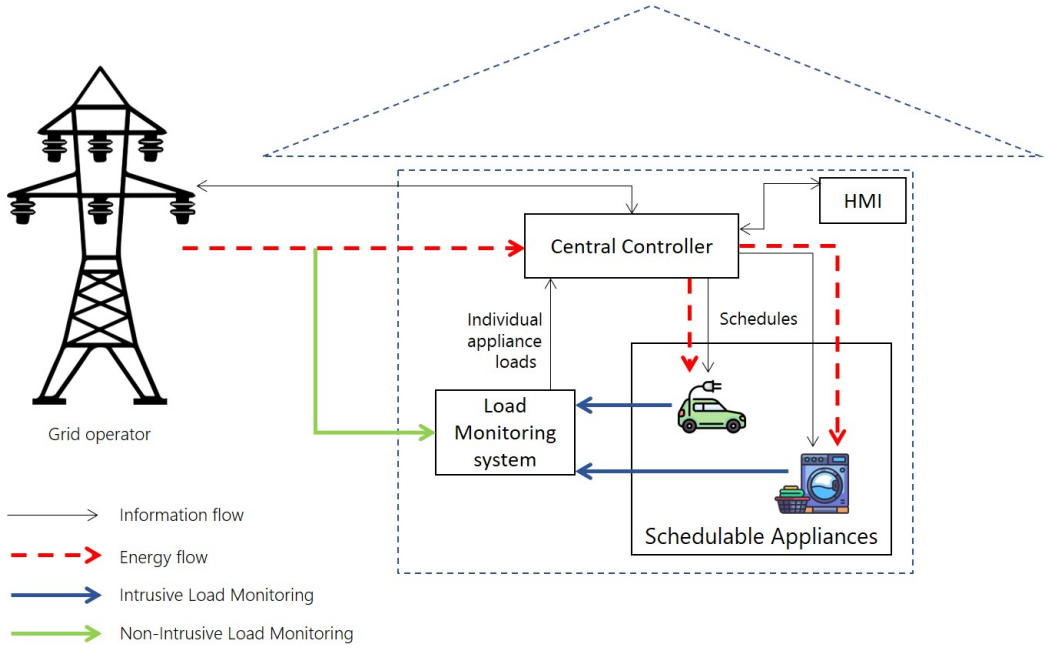


Figure 4.1. How does a HEMS work ?

of a HEMS operation. The following components and concepts are necessary to understand an HEMS operation:

- The central controller

  The HEMS main component is the central controller. Centralising and processing information from all other components, the central controller is the brain of the HEMS, producing the optimal schedules for appliance usage. The HEMS controller is the interface between the home and the grid operator. It provides feedback to the operator in case of any malfunction or power outage. It also serves as a communication interface between the user and the HEMS, which can be done using a Human Machine Interface (HMI). This interface gives the user control over the system, which is necessary to preserve their free will [12]. Moreover, it provides the user with helpful information to monitor the system's operation, such as the schedule for each smart appliance.

- Smart appliances

  In everyday life, at least for those who are not equipped with smart appliances, the appliances do not work without a physical intervention by a human be-

ing. This traditional way of utilising appliances has the advantage of making machines function as slaves, obeying human wishes. However, this approach severely limits the effectiveness of energy management systems because the dependence on the user is too significant [12, 78]. Besides, smart appliances embody communication and monitoring systems that automatically receive, process and apply schedules. Conventional appliances must be enhanced with communication units for HEMS applications.

- Load monitoring

  Different load monitoring configurations have been reviewed in previous sections (Section 1.1). Load monitoring is crucial for HEMS applications; in the works [32, 79, 80], the authors identified the power level, the average operation time or the number of uses per day with load monitoring. In the following, the contribution of 3 ALM methods is assessed: ILM, NILM and a data-augmented NILM.

- Objectives

  HEMS must have a specific objective. In the article [32], a HEMS is designed to reduce the household's energy bills while smoothing out the consumption peaks. On the other hand, in the work [79], the objective is to optimise the load profile based on the operating cost of the grid, specifically for a microgrid. A common aim in both works is to ensure the end-user's comfort as much as possible.

It is essential to experiment with a HEMS to evaluate the contribution of NILM properly. A HEMS experimental facility would have been ideal, but only simulations were exploitable in this work. One research target is developing a suitable HEMS simulation framework to compare the contributions with and without NILM. This chapter simulates the operation illustrated in Figure 4.1. This HEMS is a load scheduler, allowing specific schedulable loads, e.g. the washing machine and the dishwasher, to be day-ahead planned. The HMI will inform users of the scheduled operating times, enabling the user to decide whether or not to follow the schedule. Load scheduling is an optimisation problem involving complex load allocation tasks considering various factors and constraints [15, 81–83]. The proposed optimisation layout is detailed in the following subsections, in which the objective function (OF) and the constraints are introduced.

4.1.1 Variables and notations

This subsection is essential to understand the following mathematical formulations. The HEMSs are tested day by day at a 1-minute sampling period, so from here, we defined the time t_0 varying in the interval $[0, 1439]$ corresponding to the number of minutes in a day. It is crucial to discern the future from the past as the past is analysed to provide a schedule for the next day. So, at a present-day d, the set of past days is called D_{past} with $D_{past} = \{\dots, d-2, d-1\}$. Symmetrically, the set of future days is defined with $D_{future} = \{d+1, d+2, \dots\}$. The given schedules are the optimal starting times of each schedulable appliance represented by the variable $\hat{s}_{d,i}^{(k_s)}$; we make the hypothesis for each day, and for each appliance, only one schedule is provided per day; thus, the notation is simplified $\hat{s}_{d}^{(k_s)}$. The number of schedulable appliances is noted K_s, notably $K_s \leq K$. The variable i represents the number of activations ($N_d^{(k_s)}$) of the k_s^{th} appliance that started on the day d. We call activation a sequence of an appliance operation from start to end. An activation sequence is noted $a_{d,t_0,i}^{(k_s)}$; the operation starts at time $s_{d,i}^{(k_s)}$ and lasts $\delta_{d,i}^{(k_s)}$. Therefore, $a_{d,t_0,i}^{(k_s)}$ is defined only when t_0 belongs to the interval $[s_{d,i}^{(k_s)}, s_{d,i}^{(k_s)} + \delta_{d,i}^{(k_s)}]$. The main load ($y_{d,t_0}$) can be separated into schedulable ($\tilde{y}_{d,t_0}$) and non-schedulable ($\bar{y}_{d,t_0}$) loads, represented by the equation (4.1). We define the set of scheduled starting times for the day d by $\hat{S}_d$, such that $\hat{S}_d = \{\hat{s}_d^{(k_s)} \mid k_s \in [1, K_s]\}$. Similarly we define the set of actual starting times as $S_d = \{s_{d,i}^{(k_s)} \mid i \in [0, N_d^{(k_s)}] \mid k_s \in [1, K_s]\}$. The Figure 4.2 summarise the set of variables.

$$y_{d,t_0} = \bar{y}_{d,t_0} + \tilde{y}_{d,t_0} \tag{4.1}$$

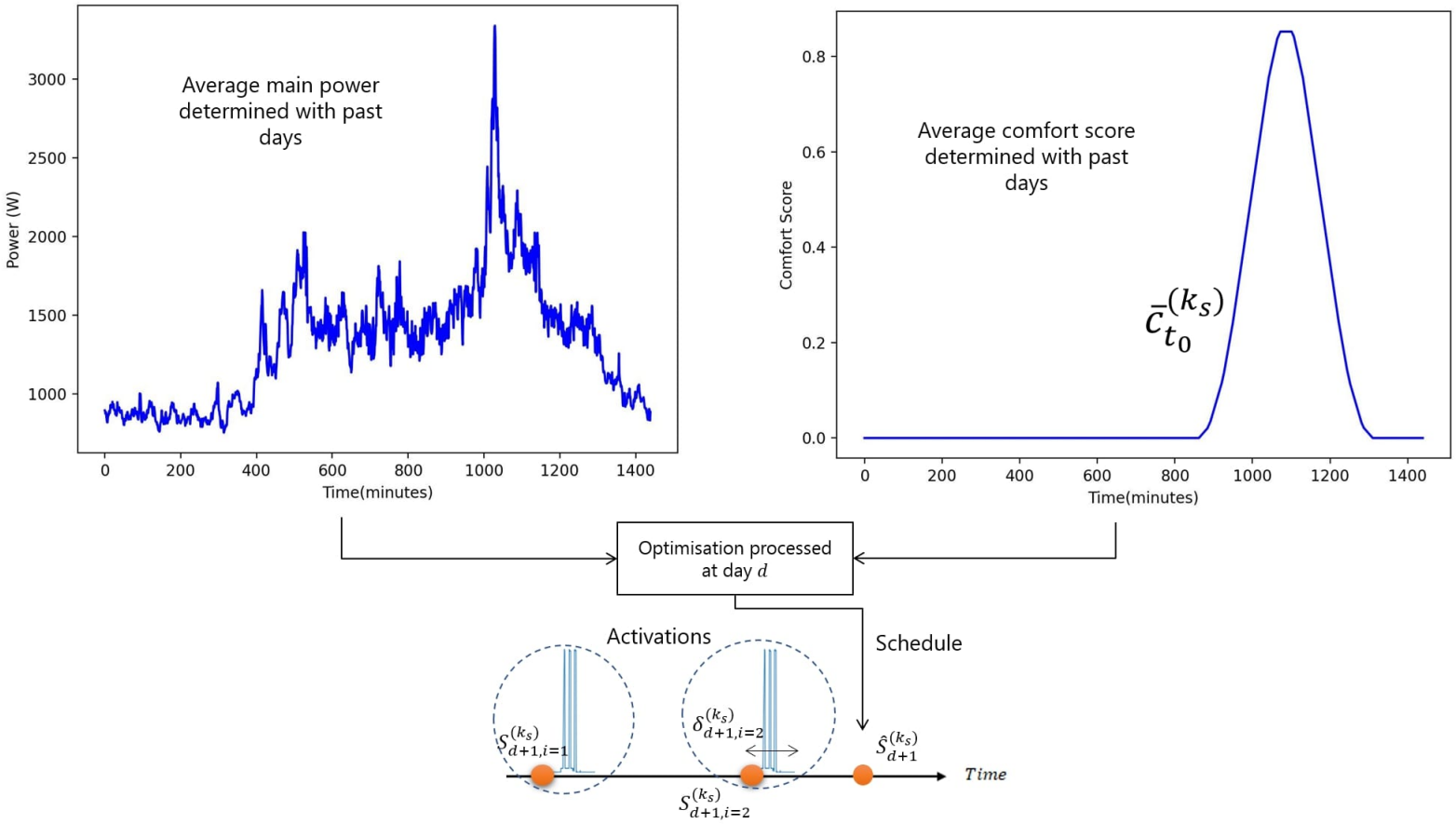


Figure 4.2. Schematic view of the schedule generation at day d for the day $d+1$ and definition of the variables.

4.1.2 Objective: Energy above a power reference

The proposed HEMS aims to reduce the daily energy above a power reference noted E_d by producing schedules of the operation starting times. The daily E_d is defined by equations (4.2), (4.3) and (4.4). $p_{d,t_0}(S_d)$ is the house's main load value, and this variable conveys the modifications of the main load when shifting the starting time of the k^{th} appliance. Δt is the time step, in hour, of the load monitoring.

$$E_d(S_d) = \sum_{t_0=0}^{1439} max(p_{d,t_0}(S_d) - p_{ref}, 0) \times \Delta t \tag{4.2}$$

with:

$$p_{d,t_0}(S_d) = \bar{y}_{d,t_0} + \sum_{k_s=1}^{K_s} \sum_{i=1}^{N^{(k_s)}} m_{d,t_0,i}^{(k_s)}(s_{d,i}^{(k_s)}, \delta_{d,i}^{(k_s)}) \tag{4.3}$$

and:

$$m_{d,t_0,i}^{(k_s)} = \begin{cases} a_{d,t_0,i}^{(k_s)} & \text{if } t_0 \in [s_{d,i}^{(k_s)}, s_{d,i}^{(k_s)} + \delta_{d,i}^{(k_s)}] \\ 0 & \text{otherwise.} \end{cases} \tag{4.4}$$

The concept of power reference in DSM has been used in [84]. The goal is to keep the power usage below a certain threshold noted $p_{ref}(t_0)$. This mechanism allows the preservation of the adequacy between energy production and load. $p_{ref}(t_0)$ is non-constant and varies with time. A basic example would be a solar-powered building; the idea would be to keep the load below the solar power production. In our case study, a constant power reference is considered, with $p_{ref} = 3000W$. This assumption applies perfectly to a house powered by a power generator; in this instance, the power reference p_{ref} is the generator power rating. By optimising the schedulable appliance starting times $\hat{s}_d^{(k_s)}$ the main load is expected to be improved. The OF is defined by equation (4.5):

$$\{\hat{s}_d^{(k_s)} \mid k_s \in [1, K_s]\} = \underset{\hat{s}_d^{(k_s)}}{argmin} |E_d(\hat{S}_d)| \tag{4.5}$$

4.1.3 Constraint 1: End-users' acceptance

User comfort is essential in the HEMS application [85] to ensure the acceptance of the technology. Indeed, the idea of HEMS is not to remove dwellers' free will but instead

to suggest a trade-off between household comfort and grid operation requirements. End-users' acceptance can highly hamper the broad application of HEMS [74–76]. The term comfort, in this whole manuscript, refers to the energy usage preferences of a household. The word "comfort" here must be dissociated from all other types of comfort, such as thermal or acoustic comfort, even if a correlation may exist. A basic example of an energy usage discomfort would be when a dweller feels cold and cannot run its electrical heater for any reason; the heater has broken down, or a HEMS has drastically delayed the device running. The work [76] gives relevant and concrete explanations of how unsuitable schedules may impact dwellers: It is crucial to launch the washing machine not far from the preferred schedules; otherwise, there is a risk of leaving wet laundry inside. Similarly, for the dishwasher, if the dishes are not cleaned on time, the dwellers may run out of available plates and forks for the next meal. For electric vehicle owners, it is a must to have enough power to be prepared for future trips. Quantifying energy preferences is challenging as they constantly change; during weekdays, working dwellers generally need to run their devices during the evening for cooking, entertaining, or simply doing the required chores. During holidays, the energy preferences may change as well. Those are quick observations and assumptions, but energy preferences are challenging to predict, making it a broad research topic [86, 87]. Within a HEMS scheduler, the comfort information has to be provided to the central controller before the processing. There are three ways to collect the comfort information:

- Manually: The dweller can inform the HEMS controller of their preferred device usage hours through an HMI on a mobile phone or a tablet [85]. Several drawbacks can be identified when using the manual method; firstly, it requires the end user's involvement so he does not forget to input the information. Furthermore, as mentioned earlier, comfort varies over time; with the manual approach, the dweller will have the annoying task of regularly accessing the HMI to have reliable schedules generated by the HEMS.

- Automatically: A forecasting unit is used to predict the energy usage comfort for the next day. The automatic method is less intrusive, but its effectiveness depends tightly on forecasting capabilities. With poor accuracy, the produced schedules can be expected to be far from the end-user's will.

- Hybrid: A mix between the manual and the automatic methods. The hybrid approach is a trade-off as the user's contribution is reduced to correcting the automatically delivered schedules if necessary.

Regarding the pros and cons of each comfort collection method, it is crucial to develop a HEMS capable of generating acceptable schedules automatically. In the HEMS community, the end-users comfort is considered an objective [15, 79] to maximise. However, in this study, comfort is included as a constraint. The philosophical idea behind the comfort constraint is that for a conventional house (without any HEMS installed), the introduction of the HEMS is similar to a robot asking the dwellers for a trade-off between their comfort and a second variable that improves the load shape. Hence, a certain margin of comfort reduction must be established for the HEMS to manage load effectively; an acceptable threshold on comfort value is set following the equation (4.7).

$$c_{d,t_0}^{(k_s)} \geq c_0 . \max_{t_0 \in [0,1439]} (\{\bar{c}_{t_0}^{(k_s)} |\ t_0 \in [0, 1439]\}) \tag{4.6}$$

$$c_{d}^{(k_s)} \geq c_0 . \max_{t_0 \in [0,1439]} (\{\bar{c}_{t_0}^{(k_s)}\}) \tag{4.7}$$

$\bar{c}_{t_0}^{(k_s)}$ being the average comfort score at time t_0. c_0 is a coefficient set in the interval $[0, 1]$; the higher the level of comfort requirement of the end-user, the higher the c value. There are 2 steps to determine the average comfort score:

- **Step 1**: *get_activations*

  This function was developed in the work [21] and allows the retrieval of activations from a sub-meter load. Following previously cited work, activation is defined as a device operation profile. The algorithm requires a power threshold to differentiate off and potentially on states to identify an activation. If the power is above the power threshold, the appliance power consumption has to be above at least for a set minimum duration parameter to be considered activated. We can find every past activation by applying the *get_activations* function on each past day. It is essential to underline that individual appliance load for the past days is known, or at least estimated with NILM, when the HEMS processes the schedules for the next day.

$$get\_activations(X_{d,t_0 \in [0:1439]}^{(k_s)}) = \{A_{d,i}^{(k_s)} \mid i \in [0 : N^{(k_s)}]\} \tag{4.8}$$

$$start\_time(A_{d,i}^{(k_s)}) = s_{d,i}^{(k_s)} \tag{4.9}$$

Each activation $A_{d,i}^{(k_s)}$ is a sequence composed of the $a_{d,t_0,i}^{(k_s)}$ values constrained to the time interval $[s_{d,i}^{(k_s)} : s_{d,i}^{(k_s)} + \delta_{d,i}^{(k_s)}]$, as summarised in equation (4.10).

$$A_{d,i}^{(k_s)} = \{a_{d,t_0,i}^{(k_s)} \mid t_0 \in [s_{d,i}^{(k_s)} : s_{d,i}^{(k_s)} + \delta_{d,i}^{(k_s)}]\} \tag{4.10}$$

- **Step 2**: Compute the comfort score

  The comfort score $c_{d,t_0}^{(k_s)}$ measures the willingness of the household to use the k^{th} appliance for each time of the day, based on the analysis of the set of past days. The concept of quantifying comfort for HEMS purposes is done in previous studies [32, 79], where it is presented as a probability density function or an appliance usage ratio. Contrary to the comfort scores in the literature, the proposed one is computed with activations found; only the activations detected in the previous step are considered. The first advantage of the proposed comfort score is to filter noises from actual activations; the second one is that the comfort score is computed with the start time of each activation using the function *start_time* (4.9) allowing to quantify the comfort depending on the exact time the user activates the appliance.

$$c_{d,t_0}^{(k_s)} = \frac{1}{N^{(k_s)}} \sum_{i=1}^{N^{(k_s)}} max(0, 1 - \frac{|s_{d,i}^{(k_s)} - t_0|}{AST}) \tag{4.11}$$

$$\bar{c}_{t_0}^{(k_s)} = \frac{1}{||D_{past}||} \sum_{d \in D_{past}} c_{d,t_0}^{(k_s)} \tag{4.12}$$

In the HEMS community, the authors [75, 76] have introduced the energy-usage-comfort through the Acceptable Delay Time (ADT) variable, which is the maximum delay of an appliance operation without disrupting consumers' comfort. The delay here means a postponement compared to the "normal" activation time. This work introduces the Acceptable Shifting Time (AST) for advancing or postponing the running up to a given time threshold. AST values in Table 4.1 are used for the following sections.

Table 4.1. AST values

	AST (h)
dishwasher	3
washing machine	3

4.1.4 Constraint 2: Constant daily energy

In this case study, the HEMS gives a schedule for the next day between midnight and 11:59 p.m. A constraint ensures the load is effectively scheduled for the desired day and does not spill over into the day after. For instance, if a washing machine operation lasts 2 hours, the HEMS cannot plan a start-up at 10:59 p.m. This convention is helpful to facilitate the daily basis assessment of the HEMS performances since the total energy consumed during the day is expected to be unchanged, and only the cleverness of allocating the schedulable load will be assessed. The constraint is summarised in equation (4.13), with λ_d an energy value targetted to stay constant when optimising the schedulable appliance starting times in S_d.

$$\sum_{t_0=0}^{1439} p_{d,t_0}(S_d) = \lambda_d \tag{4.13}$$

4.1.5 Solver: Genetic Algorithm

The Genetic Algorithm (GA) is a population-based meta-heuristic algorithm thoroughly explained in [88]. The utilisation of GA in our case study is predicated on the reason that the GA is excellent at performing global searches in solution spaces. They are proficient in exploring diverse potential solutions, making them highly suitable for discovering global or near-global optima in complex optimisation problems without being trapped in local optima. In our study, the search space complexity is $O(1440^{(k_s)})$; the more appliances, the more complex the problem. Also, GA can handle non-linear OF, as they do not require mathematical models or derivatives. This makes them particularly adapted to the OF of this work (4.5) that contains a non-linear max-function. To implement the GA, the optimisation problem is formalised within the *pymoo* library [89].
The Listing 4.1 gives the pymoo framework for the optimisation. A precedes the code comments. The comfort constraints are within the list $constr\_ieq$, which stands for inequality constraints. The constraint 2 Section 4.1.4 is an equality constraint but converted to an inequality. The objective function called $Eref\_fo$ in the code is added to the list $objs$. The *pymoo* functional problem will take both lists, $constr\_ieq$ and $objs$, inside the $FunctionalProblem$ module. Once the problem is formalised, the solver is initialised with the GA method, where all the required GA parameters are specified. The termination criteria are the number of generations. About the number of generations, the example of the fitness curve in Figure 4.3 shows a converging path. The convergence happens before 20 generations.

```
1
2     # DEFINE OBJECTIVE FUNCTION
3     # Y -> forecasted main load
4     # pref -> power reference
5     # Eref_fo -> Energy above power reference threshold
6     # A_0 -> activation {a_0_washing_machine, a_0_dishwasher}
7     objs = [lambda ts :  Eref_fo(ts,pref,Y,A_0,n)]
8
9     # DEFINE CONSTRAINTS
10    # Cf -> Comfort score ; A_0 -> list of activations to shift
11    constr_ieq = [
12        lambda x : min_comfort(x,Cf,c=0.5),
13        lambda x: constant_daily_energy(x,Y,A_0)]
14
15    # FORMALISE THE PROBLEM AS A FUNCTIONAL PROBLEM
16problem = FunctionalProblem(n_var,objs,constr_ieq=constr_ieq,xl=0,xu
    =1439)
17
18# INITIALISE THE GA ALGORITHM AND SET PARAMETER VALUES
19    # GA -> pymoo GA module
20    algorithm = GA(
21    pop_size=200,
22    n_offsprings=50,
23    sampling=IntegerRandomSampling(),crossover=SBX(prob=0.9, eta=15,
    repair=RoundingRepair()),mutation=PM(eta=20,repair=RoundingRepair())
    )
24
25    # DEFINE TERMINATION CRITERIA
26    termination = get_termination("n_gen", 30)
27
28    # MINIMISATION
29    res = minimize(problem,
30                algorithm,
31                termination)
```

Listing 4.1. Python code snippet for the optimisation problem formalised in the pymoo framework

4.2 HEMS assessment: Simulations

Simulations were conducted since no experimental facilities were available during the thesis work. Public datasets from real houses were used to preserve realistic condi-

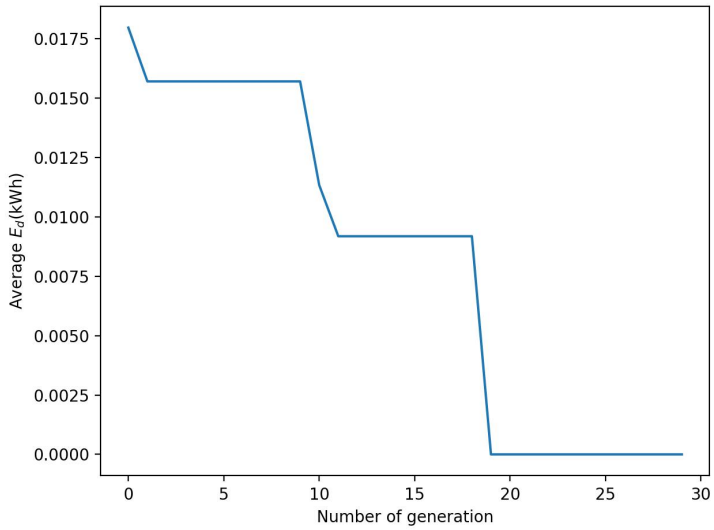


Figure 4.3. Fitness curve

tions. This section details the simulation method, the conditions and the parameters.

4.2.1 HEMS versions

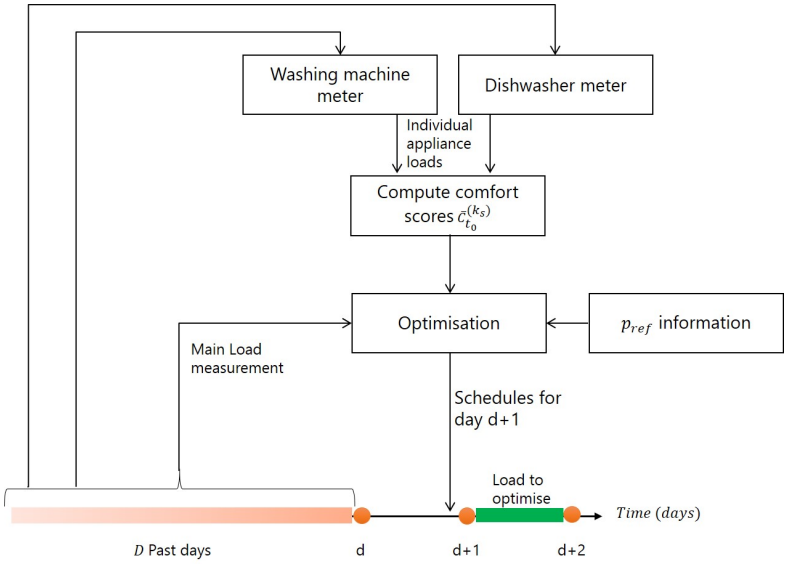
As mentioned earlier, different types of ALM systems are confronted, and the improvement brought by each of them is measured. Hence, various HEMS versions are tested here. In this case study, ALM is used in a load scheduler tool to give a behavioural analysis to determine each appliance's preferred time of use. To quantify the contribution of sub-metering, a comparison between the following configurations is proposed:

- *Hnoaug*: ALM tasks are performed by a generalised NILM. The integration of the NILM tool is depicted in Figure 4.4(b), allowing us to retrieve the comfort scores for each day of the past and compute an average comfort score.
- *Hoffsetaug*: The HEMS is equipped with a NILM trained with the OFFSETAUG data augmentation technique proposed in Chapter 3.
- *Hilm*: The HEMS recovers individual appliance measurements using intrusive sensors as seen in Figure 4.4(a). A nearly perfect measurement is expected in this case.

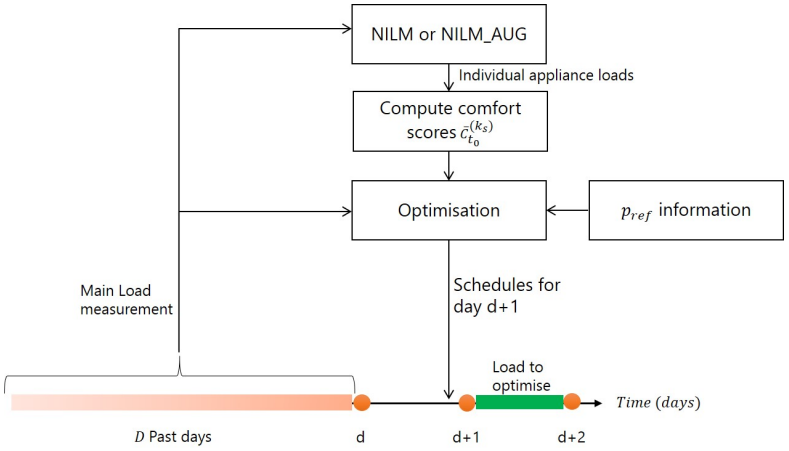
- *Hnoalm*: No sub-metering applied as depicted in Figure 4.4(c), only a past main load measurements are collected.
- *Hrandom*: A naive method providing a random schedule.
- *Hgrid*: This method recommends shifting loads to the grid's usual off-peak times, mainly at night. So, random schedules on off-peak times are generated. Figure 4.4(d) describes cases where no optimisation is undertaken; *Hrandom* and *Hgrid*.

Table 4.2. HEMS versions for simulations

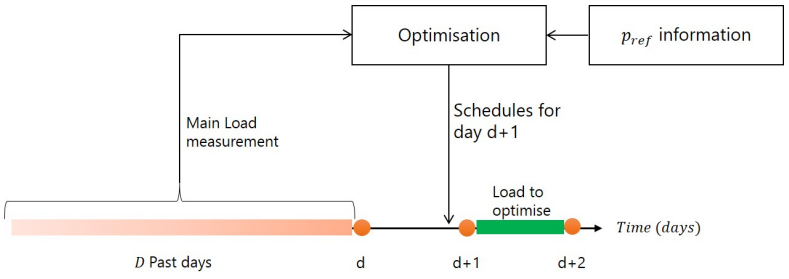
Configurations	Objective	Constraint 1	Constraint 2	ALM
Hilm	E_d	$c_0 = 0.5$	yes	ILM
Hnoaug	E_d	$c_0 = 0.5$	yes	NILM
Hoffsetaug	E_d	$c_0 = 0.5$	yes	NILM_OFFSETAUG
Hnoalm	E_d	-	yes	-
Hgrid	-	-	-	-
Hrandom	-	-	-	-



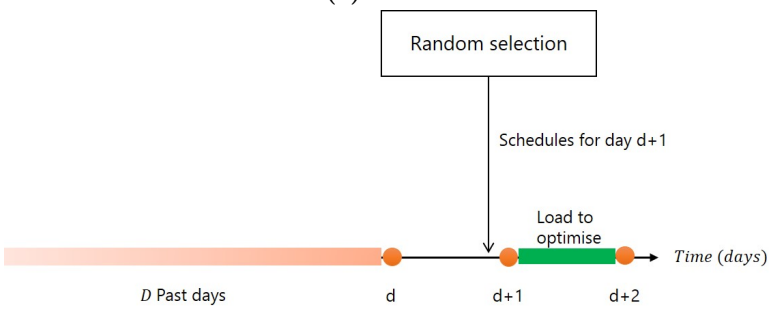
(a) *Hilm*



(b) *Hnoaug* and *Hoffsetaug*



(c) *Hnoalm*



(d) *Hrandom* and *Hgrid*

Figure 4.4. Daily ΔE_d for each HEMS configuration

4.2.2 Acceptability model

The question here is how to evaluate the relevance of the delivered schedules. They are evaluated for their potential to improve the load shape and distance with the user's preference. The significant advantage of a simulated assessment based on actual data is that we can assume the data were obtained by recording loads voluntarily activated by the end user. Consequently, we know beforehand the preferred starting times $s_{d,i}^{(k_s)}$ for each device. The idea will be to confront scheduled versus actual activation. For evaluating the schedules $\hat{s}_d^{(k_s)}$, the idea is to apply a basic appliance model based on values in Table 4.1. If a schedule is too far from the nearest actual activation starting time, the schedule is rejected, as defined in (4.14). A schedule rejection means no load shape modification caused by the k_s^{th} appliance. When the day to optimise does not contain actual activation of the k_s^{th} appliance, there is no shifting of this appliance.

$$\min_i \{|\hat{s_d}^{(k_s)} - s_{d,i}^{(k_s)}|, i \in [1, N^{(k_s)}]\} \leq AST^{(k_s)} \tag{4.14}$$

4.2.3 Hypotheses for simulation

To exemplify the hypotheses, we consider the present day d. The hypotheses of the simulation are listed below:

- The devices requiring human intervention, washing machines and dishwashers, are ready to be started at the end of the day d. Thus, the washing machine is loaded with the dirty laundry, and the dishwasher is loaded with dishes to be washed before 00:00 a.m. of day $d+1$. Consequently, the appliance just needs the top from the HEMS to be activated. This hypothesis is essential to give sense to the load scheduling process.
- At the end of the day d, the dweller receives the schedules from the central controller. He is fully free to reject the proposed schedule. Otherwise, the schedule remains.
- The resident never fails to read the HEMS information before the first scheduled activation so he can reject it.
- The only reason for schedule rejection is because the timing does not follow the user's wishes.

- The limit between acceptable and not acceptable is identified, and quantified by the AST values in Table 4.1.
- The only schedulable devices are the dishwasher and the washing machine.
- The central controller never fails to provide a schedule.

4.2.4 Simulations

To evaluate the performances of each HEMS presented in Section 4.2.1, the algorithms are incorporated into a real house from databases. A method for simulating the operation of these HEMS with actual data taken from the REFIT database is presented here. The house 17 from REFIT is used for simulations. The simulations are carried out iteratively throughout 365 days. The HEMS aims to predict the best allocation of schedulable loads, in this case, the washing machine and the dishwasher, for the following day. As depicted in Figure 4.4(a), Figure 4.4(b) and Figure 4.4(c), to optimise future day $d+1$, information from past days is necessary, except for naive HEMSs Figure 4.4(d), where schedules are generated randomly. The simulation framework is described in Figure 4.5 and follows the steps:

- **Step 0:** The input database is separated into past, present, and future days. The past days D_{past} are needed for the HEMS to generate the schedules. For the present simulation, D_{past} is taken as the last 30 days.
- **Step 1**: When the HEMS is equipped, historical data are retrieved with an ALM system. *Hoffsetaug, Hnoaug* and *Hilm* have the ALM ability contrary to *Hrandom* and *Hgrid*.
- **Step 2**: The HEMS is processing to provide schedules for the next day, $d+1$.
- **Step 3**: A reference is needed to assess the schedules' relevance. The actual load profile for the day $d+1$ is assumed to be the maximum comfort condition. Knowing this, schedules are compared with actual activations from actual data.
- **Step 4**: If the schedule is acceptable, in other words, respecting the conditions (4.14), the load is shifted. More precisely, the schedule is not rejected only if the nearest activation from the schedule, is at a distance inferior to the AST value.

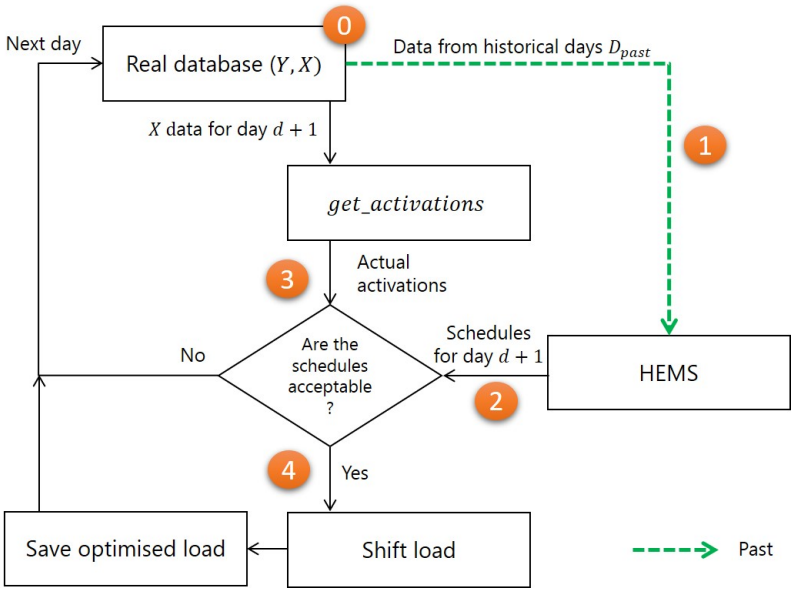


Figure 4.5. Framework of the simulation to assess HEMSs, using an existing database with a main load Y and sub-meter loads X

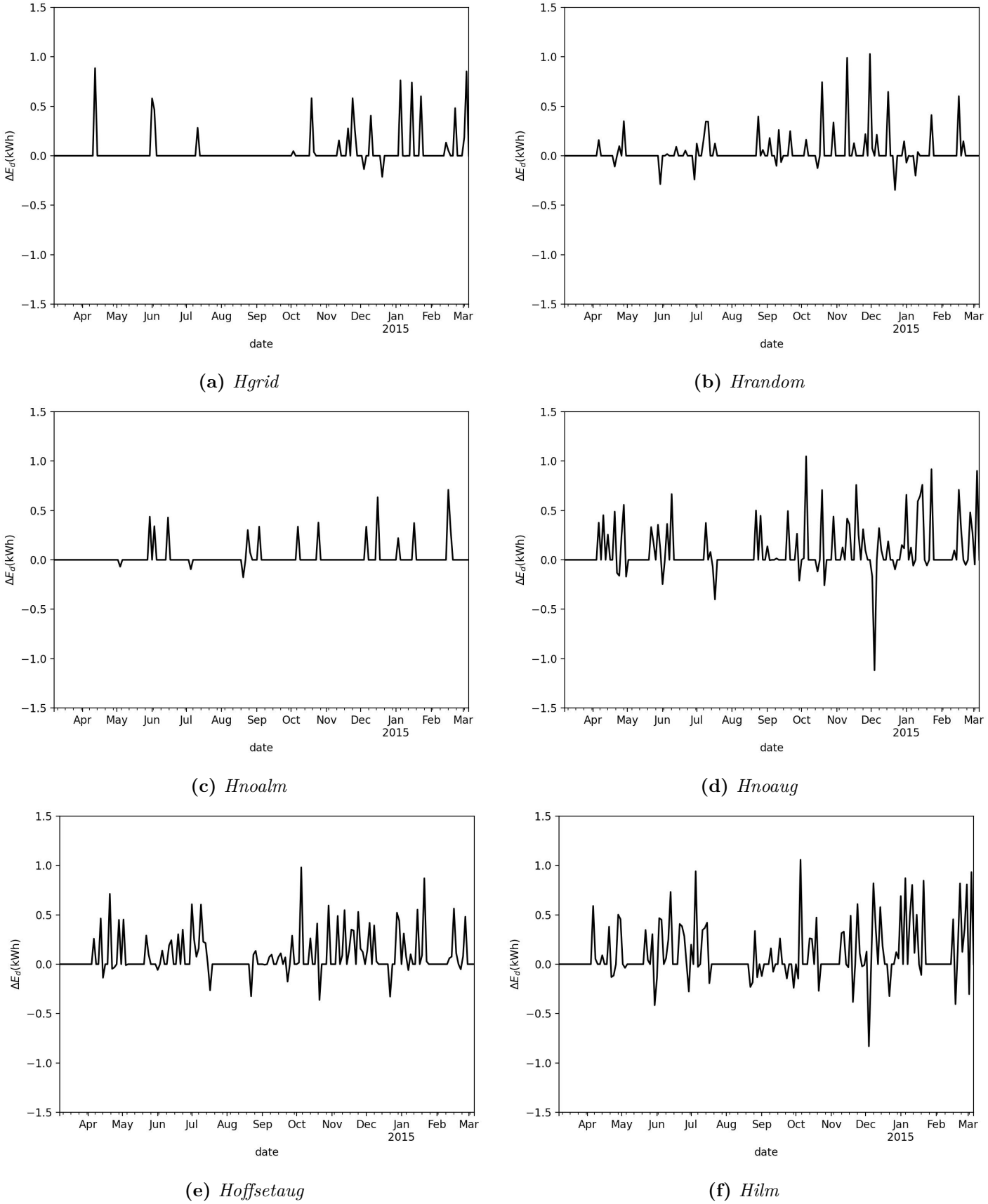
4.2.5 Metrics

Defining the metrics used to assess the HEMS before presenting the results is necessary. The chosen metrics aim to observe how effectively the HEMS can enhance the load shape, with the metric ΔE_d (4.15). When the HEMS provides improving schedules, we have $\Delta E_d > 0$. When the schedule is acceptable, but the starting time actually wanted by the user is more favourable than the HEMS proposed schedule, we have $\Delta E_d < 0$. Third case, the given schedules are not acceptable so no load scheduling could be followed; thus, for those days, $\Delta E_d = 0$.

$$\Delta E_d = E_d(S_d) - E_d(\hat{S}_d) \tag{4.15}$$

The cumulative delta, noted $C_{\Delta E_d}$ metric, provides a yearly aggregated ΔE_d as an estimate of the potential gain from using an HEMS for 1 year. This metric can be directly aligned with the size of the power supply units, for instance in a micro-grid. The implementation of a HEMS would decrease the energy above the generator power rating by a $C_{\Delta E_d}$ energy value.

$$C_{\Delta E_d} = \sum_{d \in D} \Delta E_d \tag{4.16}$$



(a) *Hgrid*

(b) *Hrandom*

(c) *Hnoalm*

(d) *Hnoaug*

(e) *Hoffsetaug*

(f) *Hilm*

Figure 4.6. Daily ΔE_d for each HEMS configuration

4.3 Results and discussions

To generate the schedules, the HEMSs need historical data analysis to find an average comfort score and main load. We can observe an average comfort score in Section 4.3. The first point to note on those curves is that house 17 from the REFIT database only uses its dishwasher in the evening. The NILM models also estimate a comfortable operation in the evening. However, the NILMs detect an earlier comfort peak. This is due to two things: firstly, the S2P model is based on the technique of centred sliding windows, which predicts the present point (central point) from future and past information (see Chapter 1). Therefore, the future influences the present. So it could be understandable to have a slight gap in starting times. The second reason is that the databases have imperfections, one of which was pointed out in Chapter 1: the timestamps of the sub-metered databases and the main measurements are sometimes different, with observable time lags. The Section 4.3 depicts a comfort discrepancy for the *Hnoaug* configuration, between about 400 to 600 minutes. A discrepancy that does not appear for *Hoffsetaug*. The NILM with DA has a comfort peak closer to the *Hilm*'s one. At this point, it would be desirable to define a metric to evaluate the accuracy in comfort. Still, the slight offset between the real start time and NILM predicted starting time makes the metric definition more challenging. Hence, these graphical observations are exploited.

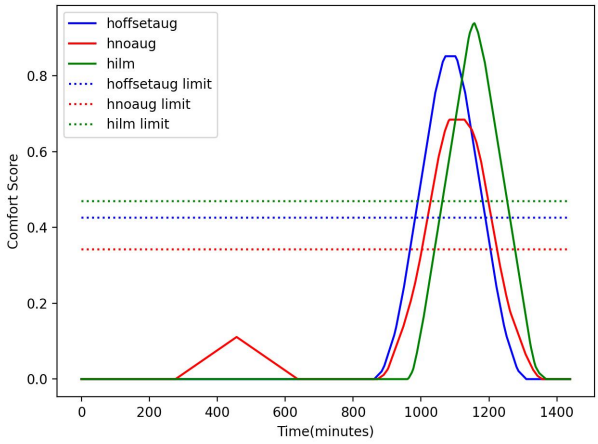


Figure 4.7. The average comfort score for the dishwasher (solid lines) and the acceptability limit with $c_0 = 0.5$ (dotted lines)

This section describes the simulation results. As stated by the previously mentioned

metrics, the analysis aims to compare the gains when the HEMS operates in the house. The comparison is done daily, by the seen of Figure 4.6. This figure shows the evolution of the ΔE_d value, which seems bounded between -1 and 1 kWh a day. This relatively low amplitude is understandable, as only the washing machine and the dishwasher operations were shifted. The positive ΔE_d values indicate improvements brought by the HEMS, and it seems all HEMSs, even the naive ones, can improve the load shape because positive peaks may be observed on all the curves. This is because the house under scrutiny has a poor load shape; to overstate it on purpose, we can say every load shape modification brings improvement. The question is, which option yields more significant enhancement? In Table 4.3, we can see HEMS equipped with an ALM system is more convenient for the end-user as all three cases *Hilm, Hnoaug*, and *Hoffsetaug* show fewer cases where $\Delta E_d = 0$ and have more cases where $\Delta E_d > 0$ compared to other HEMS versions. Thus, this demonstrates HEMS with ALM can facilitate acceptability. It is interesting to know whether, at the end of the simulation year, the HEMS has made a visible difference or not. Graphically, the answer seems evident in Figure 4.6. Table 4.3 shows the number of times that the HEMS brought about an improvement. However, the extent of the improvement is not quantified. For this, we also evaluated $C_{\Delta E_d}$, which is the algebraic area of the curves (see Figure 4.6). The results are shown in Figure 4.9. The higher the $C_{\Delta E_d}$ the more the HEMS has enhanced load management. The first observation is that the HEMS without sub-metering have less performance, as the user does not follow the schedules as depicted on the few numbers of times the ΔE_d have been modified for those configurations (see Table 4.3). The *Hilm* outperforms the others, followed by the *Hoffsetaug*. Indeed, *Hoffsetaug* performs better than *Hnoaug*, with a 2% increase in $C_{\Delta E_d}$. The difference seems thin and the contribution of OFFSETAUG to HEMS might be questioned, however, those 2% differences can be very valuable when scheduling high energy-consuming appliances such as electric vehicles.

Table 4.3. The number of times ΔE_d improved, decreased or did not change.

	$\Delta E_d > 0$	$\Delta E_d < 0$	$\Delta E_d = 0$
Hilm	55	27	101
Hnoaug	48	17	118
Hoffsetaug	60	13	110
Hnoalm	14	3	166
Hgrid	20	3	160
Hrandom	32	10	141

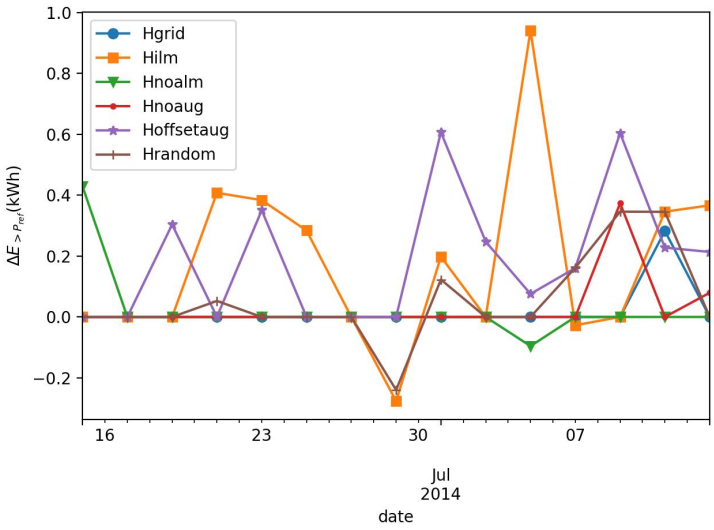


Figure 4.8. Zoomed in observation of the ΔE_d

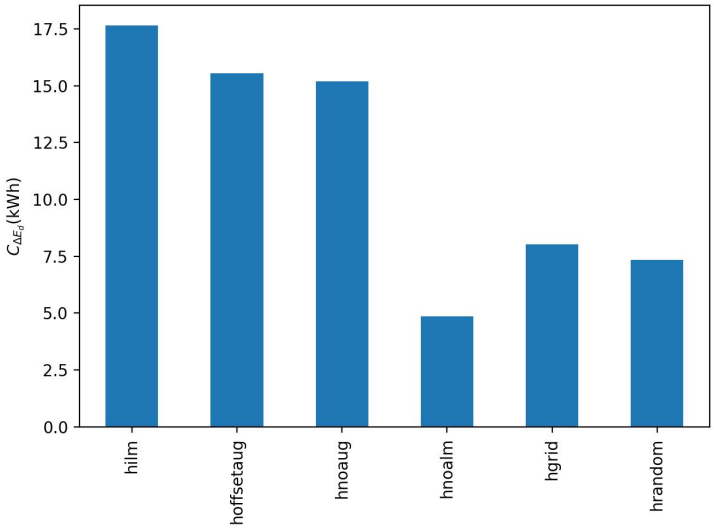


Figure 4.9. Comparison of the $C_{\Delta E_d}$ values

4.4 Conclusion

This chapter ends the manuscript with a case study application aiming to demonstrate a contribution of NILM to HEMS. The application targets load scheduling HEMS. Reproducing a relevant load scheduling application is challenging without an experimental facility. In this work, a method to simulate a HEMS is developed, based on typical database habitually used for NILM experimentation. Those data contains individual appliance loads and main load set. The idea is to simulate load scheduling to compare the contribution of HEMS with a configuration without HEMS. A very straightforward acceptability model is applied, where the users reject the schedule if it is too far from the actual preferred operation. Based on the observations, the results are more favourable for HEMS equipped with ALM, whether with ILM or with NILM. The contributions of OFFSETAUG NILM and NILM without augmentation are compared; the first demonstrates higher performances for HEMS application.

Conclusion

This work centres on NILM, a non-intrusive system to disaggregate the main load. Intrusive systems exist to monitor individual appliance loads, but those measurement solutions suffer from high installation complexity and low cost-effectiveness when monitoring several appliances. NILM is a promising cost-effective solution to unlock widely deployed ALM projects due to its ease of installation and lower costs.

Nevertheless, it is crucial to remember NILM's cost increases for lower sampling period data. To keep a certain congruity level, this work proposes to experiment with low-frequency NILM, i.e., NILM, which uses data with a high sampling period.

DL is promising among all NILM solutions; a CNN-based model, the $S2P$, is used in this work. It is well-known that DL models require a high number of data to exhibit a generalisation ability. In NILM, one of the main targets is to disaggregate any house, even houses that were never seen in the training phase.

In this work, the influence of the training database length is observed; interestingly, data length seems to have an impact when there is variability in the dataset. Length and variability are identified to be essential to unlocking generalisation capabilities. Indeed, a DA technique called OFFSETAUG has been presented, consisting of creating new power profiles and then new data based on offsetting operations. OFFSETAUG is benchmarked against some of the few known DA techniques in the scholarship and has shown promising results. Improving NILM accuracy has been a popular research topic; many algorithms have been developed recently, and each author claims higher accuracy. On the other hand, the NILM application is less prevalent in the literature. In this work, we proposed to assess the contribution of NILM to a HEMS solution. The HEMS optimises the load shape in line with a power reference value. The simulations were held on a particular house for 2 schedulable appliances, the washing machine and the dishwasher. Besides, further work should focus on simulating the HEMS with more energy-consuming appliances such as Electric Vehicles. Nevertheless, the results demonstrated a case where ALM has a significant role in determining energy usage preferences automatically. Moreover, even a low-frequency NILM can help find the user's energy comfort from historical data. Although this work highlights the technical advancements of AI, NILM and

HEMS, it also emphasises the importance of flexibility in our daily energy usage to allow for load delay, shifting or shedding. This paves the way for the conclusion of this work, where technical advancements without consumer involvement are inefficient. A collective effort is needed to strive to address recent globalised challenges.

Bibliography

[1] C. Klemenjak, C. Kovatsch, M. Herold, and W. Elmenreich, "A synthetic energy dataset for non-intrusive load monitoring in households," *Scientific Data*, vol. 7, no. 1, pp. 1–17, 2020.

[2] I. E. Agency, "World energy outlook 2021," tech. rep., IEA, 2021.

[3] "Fossil fuel energy consumption in European countries," *Energy Procedia*, vol. 153, pp. 107–111, 2018.

[4] Z. Y. Zhao, Y. X. Hao, R. D. Chang, and Q. C. Wang, "Assessing the vulnerability of energy supply chains: Influencing factors and countermeasures," *Sustainable Energy Technologies and Assessments*, vol. 56, no. October 2022, p. 103018, 2023.

[5] I. P. on Climate Change, *Special Report on Global Warming of 1.5°C*. 2018.

[6] I. S.-P. P. on Biodiversity and E. S. (IPBES), *IPBES Global Assessment Report on Biodiversity and Ecosystem Services*. 2019.

[7] I. E. A. (IEA), "Renewables 2021: Analysis and forecast to 2026," tech. rep., 2021.

[8] I. R. E. A. (IRENA), "Renewable energy integration in power grids," tech. rep., 2020.

[9] H. A. Muqeet, R. Liaqat, and M. Jamil, "A State-of-the-Art Review of Smart Energy Systems and Their Management in a Smart Grid Environment," 2023.

[10] B. Williams, D. Bishop, P. Gallardo, and J. G. Chase, "Demand Side Management in Industrial, Commercial, and Residential Sectors: A Review of Constraints and Considerations," *Energies*, vol. 16, no. 13, 2023.

[11] C. Fischer, "Feedback on household electricity consumption: A tool for saving energy?," *Energy Efficiency*, vol. 1, pp. 79–104, 2008.

[12] B. Han, Y. Zahraoui, M. Mubin, S. Mekhilef, M. Seyedmahmoudian, and A. Stojcevski, "Home Energy Management Systems: A Review of the Concept, Architecture, and Scheduling Strategies," *IEEE Access*, vol. 11, no. February, pp. 19999–20025, 2023.

[13] G. W. Hart, "Prototype nonintrusive appliance load monitor," pp. 1–170, 1985.

[14] G. W. Hart, "Nonintrusive Appliance Load Monitoring," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870–1891, 1992.

[15] X. Wang, X. Mao, and H. Khodaei, "A multi-objective home energy management system based on internet of things and optimization algorithms," *Journal of Building Engineering*, vol. 33, no. July 2019, p. 101603, 2021.

[16] C. Dinesh, B. W. Nettasinghe, R. I. Godaliyadda, M. P. B. Ekanayake, J. Ekanayake, and J. V. Wijayakulasooriya, "Residential Appliance Identification Based on Spectral Information of Low Frequency Smart Meter Measurements," *IEEE Transactions on Smart Grid*, vol. 7, no. 6, pp. 2781–2792, 2016.

[17] N. Batra, R. Kukunuri, A. Pandey, R. Malakar, R. Kumar, O. Krystalakos, M. Zhong, P. Meira, and O. Parson, "Towards reproducible state-of-the-art energy disaggregation," no. February 2020, pp. 193–202, 2019.

[18] C. Zhang, M. Zhong, Z. Wang, N. Goddard, and C. Sutton, "Sequence-to-point learning with neural networks for non-intrusive load monitoring," *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pp. 2604–2611, 2018.

[19] D. Murray, L. Stankovic, V. Stankovic, S. Lulic, and S. Sladokevic, "Transferability of Neural Network Approaches for Low-Rate Energy Disaggregation," *ICASSP 2019*, vol. 1, no. 1, pp. 8330–8334, 2019.

[20] X. Zhou, S. Li, C. Liu, H. Zhu, N. Dong, and T. Xiao, "Non-Intrusive Load Monitoring Using a CNN-LSTM-RF Model Considering Label Correlation and Class-Imbalance," *IEEE Access*, vol. 9, pp. 1–1, 2021.

[21] J. Kelly and W. Knottenbelt, "Neural NILM: Deep Neural Networks Applied to Energy Disaggregation," 2015.

[22] M. D'Incecco, S. Squartini, and M. Zhong, "Transfer Learning for Non-Intrusive Load Monitoring," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1419–1429, 2020.

[23] W. Luan, R. Zhang, B. Liu, B. Zhao, and Y. Yu, "Leveraging sequence-to-sequence learning for online non-intrusive load monitoring in edge device," *International Journal of Electrical Power and Energy Systems*, vol. 148, no. 92, p. 108910, 2023.

[24] D. Murray, L. Stankovic, and V. Stankovic, "An electrical load measurements dataset of United Kingdom households from a two-year longitudinal study," *Scientific Data*, vol. 4, pp. 1–12, 2017.

[25] A. Harell, R. Jones, S. Makonin, and I. V. Bajic, "Tracegan: Synthesizing appliance power signatures using generative adversarial networks," *IEEE Transactions on Smart Grid*, vol. 12, pp. 4553–4563, 2021.

[26] A. Delfosse, G. Hebrail, and A. Zerroug, "Deep learning applied to nilm: Is data augmentation deep learning applied to nilm: Is data augmentation," *Frontiers in Artificial Intelligence and Applications*, vol. 325, pp. 2972–2977, 2020. Multi-Agent Simulatorsample rate : 2s.

[27] H. Rafiq, X. Shi, H. Zhang, H. Li, M. K. Ochani, and A. A. Shah, "Generalizability improvement of deep learning-based non-intrusive load monitoring system using data augmentation," *IEEE Transactions on Smart Grid*, vol. 12, no. 4, pp. 3265–3277, 2021.

[28] M. Nour, J. C. Le Bunetel, P. Ravier, and Y. Raingeaud, "Data Augmentation Strategies for High-Frequency NILM Datasets," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–9, 2023.

[29] J. Francou, D. Calogine, O. Chau, M. David, and P. Lauret, "Expanding variety of non-intrusive load monitoring training data: Introducing and benchmarking a novel data augmentation technique," *Sustainable Energy, Grids and Networks*, vol. 35, p. 101142, 2023.

[30] Y. Pan, K. Liu, Z. Shen, X. Cai, and Z. Jia, "Sequence-to-Subsequence Learning with Conditional GAN for Power Disaggregation," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3202–3206, 2020.

[31] D. He, W. Lin, N. Liu, R. G. Harley, and T. G. Habetler, "Incorporating non-intrusive load monitoring into building level demand response," *IEEE Transactions on Smart Grid*, vol. 4, no. 4, pp. 1870–1877, 2013.

[32] Y. H. Lin and M. S. Tsai, "An Advanced Home Energy Management System Facilitated by Nonintrusive Load Monitoring with Automated Multiobjective Power Scheduling," *IEEE Transactions on Smart Grid*, vol. 6, no. 4, pp. 1839–1851, 2015.

[33] M. Amayri, C. S. Silva, H. Pombeiro, and S. Ploix, "Flexibility characterization of residential electricity consumption: A machine learning approach," *Sustainable Energy, Grids and Networks*, vol. 32, p. 100801, 2022.

[34] L. W. Tokam and S. S. Ouro-Djobo, "Comparative Study on Load Monitoring Approaches," *Appl. Sci. 2023*, vol. 13, no. 9, pp. 124–134, 2023.

[35] S. Henrique, J. Correia, A. J. Gano, A. M. de Campos, and I. Teixeira, "Domestic Power Consumption Measurement and Automatic Home Appliance Detection," *Proceedings of the 4th IEEE International Workshop on Intelligent Signal Processing (WISP)*, pp. 128–132, 2005.

[36] M. Marcu and A. Stancovici, "Systems classification based on power signatures," *IEEE PES Innovative Smart Grid Technologies Conference Europe*, no. 1, pp. 1–6, 2011.

[37] A. Ridi, C. Gisler, and J. Hennebert, "A survey on intrusive load monitoring for appliance recognition," *Proceedings - International Conference on Pattern Recognition*, pp. 3702–3707, 2014.

[38] E. J. Aladesanmi and K. A. Folly, "Overview of non-intrusive load monitoring and identification techniques," *IFAC-PapersOnLine*, vol. 48, no. 30, pp. 415–420, 2015.

[39] L. Diamond, J. Schrammel, P. Fröhlich, G. Regal, and M. Tscheligi, "Privacy in the smart grid: End-user concerns and requirements," *MobileHCI 2018 - Beyond Mobile: The Next 20 Years - 20th International Conference on Human-Computer Interaction with Mobile Devices and Services, Conference Proceedings Adjunct*, no. March 2019, pp. 189–196, 2018.

[40] L. Alabdulkarim and Z. Lukszo, "Impact of privacy concerns on consumers' acceptance of smart metering in the Netherlands," *2011 International Conference on Networking, Sensing and Control, ICNSC 2011*, no. April, pp. 287–292, 2011.

[41] S. Heuninckx, M. Meitern, G. te Boveldt, and T. Coosemans, "Practical problems before privacy concerns: How European energy community initiatives strug-

gle with data collection," *Energy Research and Social Science*, vol. 98, no. March, p. 103040, 2023.

[42] A. Faustine, L. Pereira, H. Bousbiat, and S. Kulkarni, "UNet-NILM: A Deep Neural Network for Multi-tasks Appliances State Detection and Power Estimation in NILM," *NILM 2020 - Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring*, no. November, pp. 84–88, 2020.

[43] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[44] H. Bousbiat, Y. Himeur, I. Varlamis, F. Bensaali, and A. Amira, "Neural Load Disaggregation: Meta-Analysis, Federated Learning and Beyond," *Energies*, vol. 16, no. 2, pp. 1–22, 2023.

[45] Y. Lecun and Y. Bengio, "Convolutional Networks for Images, Speech, and Time-Series," *The handbook of brain theory and neural networks*, pp. 255–258, 1995.

[46] D. E. Rumelhart and G. E. Hintont, "Learning Representations by Back-Propagating Errors," *Nature*, vol. 323, no. 9, 1986.

[47] O. Krystalakos, C. Nalmpantis, and D. Vrakas, "Sliding window approach for online energy disaggregation using artificial neural networks," *ACM International Conference Proceeding Series*, pp. 1–6, 2018.

[48] A. Reinhardt and M. Bouchur, "On the Impact of the Sequence Length on Sequence-to-Sequence and Sequence-to-Point Learning for NILM," *NILM 2020 - Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring*, pp. 75–78, 2020.

[49] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[50] N. Batra, J. Kelly, O. Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, and M. Srivastava, "NILMTK: An open source toolkit for non-intrusive load monitoring," *e-Energy 2014 - Proceedings of the 5th ACM International Conference on Future Energy Systems*, pp. 265–276, 2014.

[51] J. Kelly, N. Batra, O. Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, and M. Srivastava, "NILMTK v0.2: A non-intrusive load monitoring toolkit for

large scale data sets," *BuildSys 2014 - Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*, pp. 182–183, 2014.

[52] A. Collette and contributors, "Hdf5 for python," 2014.

[53] Z. Kolter and M. J.Johnson, "REDD Dataset," 2011.

[54] J. Kelly and W. Knottenbelt, "The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes," *Scientific Data*, vol. 2, pp. 1–14, 2015.

[55] O. Parson, G. Fisher, A. Hersey, N. Batra, J. Kelly, A. Singh, W. Knottenbelt, and A. Rogers, "Dataport and NILMTK: A building data set designed for non-intrusive load monitoring," *2015 IEEE Global Conference on Signal and Information Processing, GlobalSIP 2015*, pp. 210–214, 2016.

[56] C. Klemenjak, C. Kovatsch, M. Herold, and W. Elmenreich, "A synthetic energy dataset for non-intrusive load monitoring in households," *Scientific Data*, vol. 7, no. 1, pp. 1–17, 2020.

[57] S. Makonin and F. Popowich, "Nonintrusive load monitoring (NILM) performance evaluation: A unified approach for accuracy reporting," *Energy Efficiency*, vol. 8, no. 4, pp. 809–814, 2015.

[58] E. T. Mayhorn, J. Petersen, R. S. Butner, and E. M. Johnson, "Load Disaggregation Technologies: Real World and Laboratory Performance," *ACEEE Summer Study on Energy Efficiency in Buildings*, pp. 1–13, 2016.

[59] L. Pereira and N. Nunes, "Performance evaluation in non-intrusive load monitoring: Datasets, metrics, and tools—A review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 6, pp. 1–17, 2018.

[60] O. Parson, S. Ghosh, M. Weal, and A. Rogers, "An unsupervised training method for non-intrusive appliance load monitoring," *Artificial Intelligence*, vol. 217, pp. 1–19, 2014.

[61] Legifrance, "Arrêté du 4 janvier 2012 pris en application de l'article 4 du décret n° 2010-1022 du 31 août 2010 relatif aux dispositifs de comptage sur les réseaux publics d'électricité," 2012.

[62] O. Guerra Santin, L. Itard, and H. Visscher, "The effect of occupancy and building characteristics on energy use for space and water heating in Dutch residential stock," *Energy and Buildings*, vol. 41, no. 11, pp. 1223–1232, 2009.

[63] Z. Wang and Y. Ding, "An occupant-based energy consumption prediction model for office equipment," *Energy and Buildings*, vol. 109, pp. 12–22, 2015.

[64] T. L. Quy, S. Zerr, E. Ntoutsi, and W. Nejdl, "Data augmentation for dealing with low sampling rates in NILM," 2021.

[65] P. R. I. Electric, "Demand-side management: Utility options for the future," *EPRI Reports*, vol. Cu. 3028.10.89, 2006.

[66] M. AboGaleela, M. El-Sobki, and M. El-Marsafawy, "A two-level optimal dsm load shifting formulation using genetics algorithm case study: Residential loads," *IEEE Power and Energy Society Conference and Exposition in Africa: Intelligent Grid Integration of Renewable Energy Resources, PowerAfrica 2012*, pp. 9–13, 2012.

[67] P. Siano, "Demand response and smart grids - A survey," *Renewable and Sustainable Energy Reviews*, vol. 30, pp. 461–478, 2014.

[68] D. York and M. Kushler, "Exploring the relationship between demand response and energy efficiency: a review of experience and discussion of key issues.," *American Council for an Energy-Efficient Economy*, vol. Report no. U052, 2005.

[69] B. Jacob, "Lamps for improving the energy efficiency of domestic lighting," *Lighiting Research Technology*, vol. 41, 2009.

[70] https://www.ecologie.gouv.fr/dispositif-des-certificats-deconomies-denergie. Accessed: 2023-10-07.

[71] W. Y. Chiu, J. T. Hsieh, and C. M. Chen, "Pareto optimal demand response based on energy costs and load factor in smart grid," *IEEE Transactions on Industrial Informatics*, vol. 16, pp. 1811–1822, 2020.

[72] W.-Y. Chiu, H. Sun, and H. V. Poor, "Energy imbalance management using a robust pricing scheme," *IEEE Trans. Smart Grid*, vol. 4, p. 896–904, 2013.

[73] H. Golmohamadi, "Demand-side management in industrial sector: A review of heavy industries," *Renewable and Sustainable Energy Reviews*, vol. 156, no. June 2021, 2022.

[74] A. Safdarian, M. Fotuhi-Firuzabad, and M. Lehtonen, "Benefits of demand response on operation of distribution networks: A case study," *IEEE Systems Journal*, vol. 10, pp. 189–197, 2016. Definition of Appliance Delay TimeRef

16 - Conditional analysis for disaggregatingRef 19 - MATPOWER program for power flow analysis.

[75] R. Stamminger, "Synergy potential of smart appliances," *D2.3 of WP2 from the Smart-A project*, p. 237, 2008.

[76] W. Mert, "Consumer acceptance of smart appliances," *D 5.5 of WP 5 report from Smart-A project*, 2008.

[77] U. Zafar, S. Bayhan, and A. Sanfilippo, "Home Energy Management System Concepts, Configurations, and Technologies for the Smart Grid," *IEEE Access*, pp. 119271 – 119286, 2020.

[78] M. Kuzlu, M. Pipattanasomporn, and S. Rahman, "Review of communication technologies for smart homes/building applications," *IEEE Innov. Smart Grid Technol.*, pp. 1–6, 2015.

[79] Cimen, Halil and Cetinkaya, Nurettin and Vasquez, Juan C. and Guerrero, Josep M., "A Microgrid Energy Management System Based on Non-Intrusive Load Monitoring via Multitask Learning," *IEEE Transactions on Smart Grid*, vol. 12, no. 2, pp. 977–987, 2021.

[80] N. Zhang, L. F. Ochoa, and D. S. Kirschen, "Investigating the impact of demand side management on residential customers," *IEEE PES Innovative Smart Grid Technologies Conference Europe*, pp. 1–6, 2011.

[81] S. Gottwalt, W. Ketter, C. Block, J. Collins, and C. Weinhardt, "Demand side management-A simulation of household behavior under variable prices," *Energy Policy*, vol. 39, no. 12, pp. 8163–8174, 2011.

[82] T. Logenthiran, D. Srinivasan, and T. Z. Shun, "Demand side management in smart grid using heuristic optimization," *IEEE Transactions on Smart Grid*, vol. 3, no. 3, pp. 1244–1252, 2012.

[83] Y. F. Du, L. Jiang, Y. Li, and Q. Wu, "A robust optimization approach for demand side scheduling considering uncertainty of manually operated appliances," *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 743–755, 2018.

[84] N. Zhang, L. F. Ochoa, and D. S. Kirschen, "Investigating the impact of demand side management on residential customers," *IEEE PES Innovative Smart Grid Technologies Conference Europe*, pp. 1–6, 2011.

[85] J. Leitao, P. Gil, B. Ribeiro, and A. Cardoso, "A survey on home energy management," *IEEE Access*, pp. 5699–5722, 2020.

[86] T. Chaudhuri, Y. C. Soh, H. Li, and L. Xie, "A feedforward neural network based indoor-climate control framework for thermal comfort and energy saving in buildings," *Appl. Energy*, vol. 248, pp. 44–53, 2019.

[87] S. M. A. Kazmi and A. M. Khattak, "ApplianceNet: A neural network based framework to recognize daily life activities and behavior in smart home using smart plugs," *Neural Comput. Appl.*, pp. 12749–12763, 2022.

[88] M. Srinivas and L. M. Patnaik, "Genetic Algorithms: A Survey," *Computer*, vol. 27, no. 6, pp. 17–26, 1994.

[89] https://pymoo.org/. Accessed: 2023-10-29.