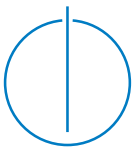
Matthias Kahl

Machine Learning for Non-Intrusive Load Monitoring

Technische
Universität
München



Technische Universität München

Fakultät für Informatik

Lehrstuhl für Wirtschaftsinformatik

Machine Learning for Non-Intrusive Load Monitoring

Matthias Kahl

Vollständiger Abdruck der von der Fakultät für Informatik der Technische Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: Prof. Dr. Hans Michael Gerndt

Prüfer der Dissertation:

1. Prof. Dr. Hans-Arno Jacobsen
2. Prof. Dr. Alexander Horsch

Die Dissertation wurde am 03.06.2019 bei der Technische Universität München eingereicht und durch die Fakultät für Informatik am 10.09.2019 angenommen.

Photo by Matthias Kahl – Bukit Lawang 2014

„Some people talk to animals. Not many listen though. That's the problem."

– Alan Alexander Milne

Abstract

Saving electrical energy is one important way of tackling the anthropogenic climate change. Non-intrusive load monitoring (NILM) is an information retrieval process for electrical appliances and their energy consumption without any significant intervention into the electric circuit. NILM may help reduce energy consumption by providing detailed consumption reports on the appliance level to consumers. By using new approaches based on machine and representation learning, we show the possibilities and limitations of consumption feedback under real-world circumstances.

The energy consumption feedback with NILM follows four processing steps of which we address the two inner steps in detail: event detection and appliance recognition. We present an efficient, machine learning based appliance recognition system, that uses the best-performing selection from a comprehensive comparison of 36 appliance features. To extract and evaluate appliance specific features, the WHITED dataset with 1259 individual appliance measurements has been composed. The focus lies on isolated measurements of the first 5 s after the appliance start-up. To provide a suggestion for the most effective usage in households, three smart meter configuration setups for appliance recognition are evaluated with multiple datasets. Object of relevance is the database from which the appliance recognition system learns. Regarding the event detection, the number of falsely positive detected events from switch-mode power supply (SMPS) equipped appliances can be reduced with the help of our multivariate event detection and its adaptive training approach. The detection is based on learning from consumer relevant appliance events to distinguish them from non-relevant appliance transients. Furthermore, we show that the appliance recognition can be implemented with modern deep learning approaches, allowing to relinquish from hand-crafted appliance feature extraction. Depending on the volume of supervised labeled training data, we can gain the same classification performances, compared to the classical machine learning approach with hand-crafted feature extraction.

Zusammenfassung

Das Sparen elektrischer Energie ist ein wesentlicher Schritt bei der Bekämpfung des durch den Menschen beeinflussten Klimawandels. Non-intrusive load monitoring (NILM) ist ein Prozess zur Analyse von elektrischen Verbrauchern und deren individuellem Energieverbrauch ohne signifikantem Eingriff in den elektrischen Stromkreis. NILM kann durch die einfache Installation dazu beitragen, Stromkonsumenten einen detaillierten, gerätebezogenen Verbrauchsbericht zu liefern, um Ansatzpunkte zum Energiesparen aufzuzeigen. Wir zeigen unter Verwendung neuer Verfahren, basierend auf Machine und Representation Learning, was die Möglichkeiten und Grenzen derartiger System in realen Einsatzszenarien sind.

Verbrauchsanalysen mit NILM folgen typischerweise einer vierstufigen Bearbeitungskette von der wir die inneren zwei Stufen, die Einschalt- und Gerätetyperkennung vertieft bearbeiten. Mit einer umfangreichen Analyse zu 36 Gerätemerkmalen und deren Eignung für die Unterscheidung von Gerätetypen zeigen wir, wie eine effiziente Geräteerkennung mit Hilfe von Machine Learning und einer geeignetem Auswahl an Merkmalen möglich ist. Um gerätespezifische Merkmale von einer Vielzahl an elektrischen Verbrauchern extrahieren und evaluieren zu können, wurde zu diesem Zweck der WHITED Datensatz mit 1259 einzelnen Gerätemessungen zusammengestellt. Der Fokus lag dabei auf der isolierten Messung der ersten 5 s nach dem Einschaltvorgang. Drei verschiedene Konfigurationsszenarien für einen potentiellen Stromzähler mit Geräteerkennung wurden anhand mehrerer Datensätze überprüft, um eine Empfehlung zum effektiven Einsatz geben zu können. Entscheidend ist dabei der Ursprung der Datenbasis anhand derer die Geräteklassen erlernt werden. Mit Hilfe unserer multivariaten Einschalterkennung und insbesondere deren adaptivem Trainingsansatz kann die Anzahl falsch positiv erkannter Einschaltmomente von Geräten mit Schaltnetzteilen signifikant reduziert werden. Hierbei werden konsumentenrelevante Einschaltmomente semantisch von anderen Geräteereignissen und -transienten unterschieden und explizit anhand von Beispielen erlernt. Desweiteren zeigen wir dass es möglich ist, die Geräteerkennung mithilfe moderner Deep Learning Verfahren zu implementieren und damit auf die manuelle Extraktion von Gerätemerkmalen verzichten zu können. Abhängig vom Umfang der überwacht erhobenen Menge an Trainingsdaten ist es uns möglich die gleichen Erkennungsleistungen zu erreichen wie sie mit manueller Merkmalsextraktion im klassischen Machine Learning möglich ist.

Acknowledgments

This dissertation and necessary work towards it took place at the Department of Informatics of the Technische Universität München under the supervision of Prof. Hans-Arno Jacobsen.

First, I want to thank Prof. Hans-Arno Jacobsen for accepting me as a doctoral candidate, his valuable feedback, support and the great freedom I was allowed to live out in all aspects of the research activity.

I would like to thank Prof. Alexander Horsch for his feedback, interesting talks and his agreement to be the second examiner of this work. Further, I thank Prof. Dr. Hans Michael Gerndt for acting as chair of the committee.

A great "Thank you" goes to my colleagues that shared her time with me at the chair. Without their help, creativity, ideas, time for kicker-playing and motivational discussions during the years, the thesis would not be possible. Thanks go to Thomas Kriechbaumer as an always supportive officemate, as well as Anwar Ul Haq and Daniel Jorde for sharing ideas and knowledge as an integral part of our NILM research group. Further, I thank Elias Stehle, Martin Jergler and Christoph Doblander for their motivational input and friendship we could develop over the last years.

Besides academia, I want to say thank you to my parents Roswitha and Ekkehart for their patients, trust in me and their optimism. A special thank goes to my wife for her love, serenity, patients and for being the captain. I want to thank Dr. Bernd Flügel and Prof. Martin Golz for implanting the idea of doing a Ph.D. in natural sciences and machine learning. I also want to thank Linus and Luana for being human enough to be a great comfort in heavy times.

Contents

1

Introduction

Non-intrusive load monitoring (NILM) is a modern technique for observing voltage and current signals to retrieve detailed energy consumption and appliance state information of a rather small electric circuit such as a residential home, floor or small industrial environment. NILM combines several techniques to gain insights into consumption amount, consumption pattern and state of appliances [1]. NILM can be used for appliance start-up identification [2], appliance identification [3], demand response [4], predictive maintenance [5] and sensor-net simplification [6]. Similar techniques can be applied on water [7], gas or any other measurements on flowing matter.

One purpose of NILM is to represent an alternative to intrusive load monitoring (ILM). ILM considers measurement units on each appliance of interest [8], while NILM aims for one intelligent sensor at the aggregated signal, usually at the electric cabinet, see Figure 1.0.1 and 1.0.2. The intelligent sensor is equipped with state-of-the-art artificial intelligence to identify appliance states, class and consumption in real or near-real time on the basis of appliance-specific characteristics in the current and voltage signals. NILM can also be seen as reporting system for automated demand response to support consumers saving energy without lowering their comfort.

Several small companies provide services based on NILM [9]. Main drawbacks for NILM systems lie in their imperfect appliance prediction, as well as general privacy concerns.

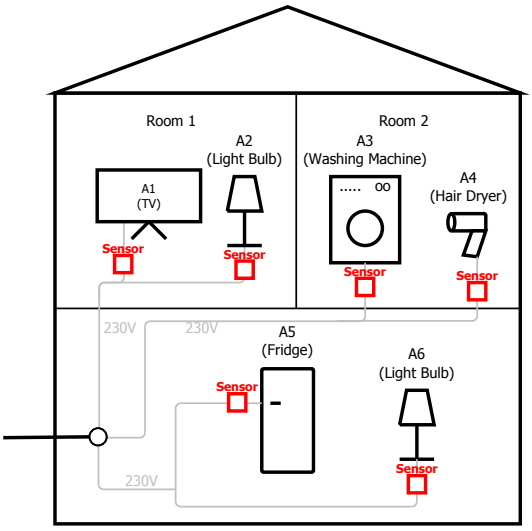


Figure 1.0.1: Intrusive Load Monitoring

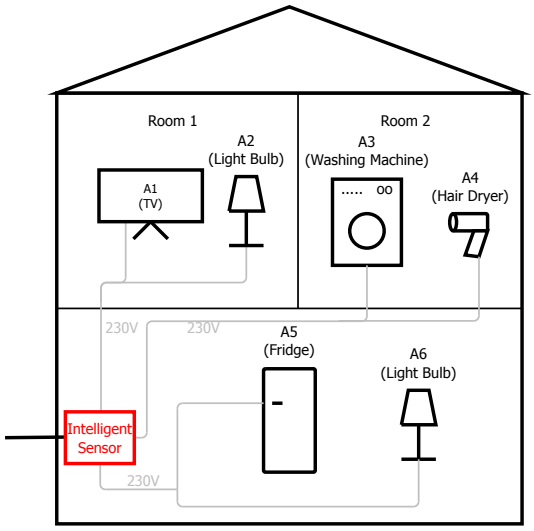


Figure 1.0.2: Non-Intrusive Load Monitoring

1.1 Motivation

Most climate scientists agree about the anthropogenic climate change [10]. The concerns focus on the global temperature increase in the upcoming decades. The main reason for the human caused temperature increase can be explained by the release of huge amounts of CO_2 into the atmosphere. Heating and electricity are the sectors with the largest CO_2 emissions [11]. More than 65% of the worldwide electrical energy is produced by non-carbon-neutral fossil fuels [11] which releases the greenhouse gas CO_2 into the atmosphere. A higher carbon concentration causes temperature increase of the atmosphere [12] due to the greenhouse effect.

Around 27% of the worldwide electrical energy consumption goes to the residential sector [11]. Unfortunately, residential consumers are usually not aware of their appliance-wise energy consumption due to a poor consumption report provided within the energy billing procedure. The meta study of Kelly and Knottenbelt [13] found little evidence that consumption feedback helps reducing electricity consumption by 0.7-4.5% [13]. Although there is no clear evidence, but several studies support the statement that disaggregated consumption feedback leads to a reduction of energy consumption [13]. A perfectly accurate disaggregation seems to be not necessary, making the NILM approach an even more convenient option for residential consumers. One reason for the small interest of

the consumer in consumption feedback is that energy prices are not seen as a significant cost factor and even an extravagant energy consumption is still affordable for most people. This might change in the future.

1.2 Problem Statement

The NILM process for energy consumption feedback can be divided into four steps (see Figure 2.1.1). In this work, we target the two inner challenges: event detection and appliance classification.

Many appliance event detection approaches are based on thresholds and simple rules, implicating a simple appliance event definition. These approaches are usually able to cover the vast number of appliance events. In the context of energy consumption feedback, not all appliance events are of interest. Furthermore, switch-mode power supply (SMPS) driven appliances draw very heterogeneous current signals that show event-like transients, challenging any event detection algorithm.

Smart meters are usually equipped with a rather small processing unit, limiting the appliance recognition algorithms to baseline approaches. Building an efficient appliance recognition system that is able to perform on such baseline hardware, a well-chosen selection of appliance features and classification algorithm are necessary. Many features, with different - usually unknown - discriminative potential for appliance recognition can be found in the literature from different research fields. Since appliances have their individual differentiability across their classes, the discriminative potential of each appliance class is also of high interest.

Appliance recognition is a discipline that still faces a significant error margin. Depending on the appliances and their usage pattern, the environment and the applied algorithms, an average classification F-Score between 0.75 and 1.0 is possible with machine learning approaches [3, 14, 15, 16]. Deep neural networks can be used for representation learning, an approach for classification that identifies features from raw data in an automated way, forming an worthy alternative to hand-crafted, expert-driven feature extraction - the conventional machine learning paradigm.

The large number of individual appliances of a certain type makes a challenge to build a uniform appliance model. It is impossible to collect representative measurements of each appliance model from each manufacturer. Since there is no such large database of appliance measurements, other strategies need to be followed to identify appliances successfully. Those strategies may include the integration of the consumer for initial appliance feedback.

Electrical appliances can be distinguished based on individual characteristics in the current and voltage signals with general features. The extraction of appliance-specific features may improve the classification performance. To retrieve appliance-specific features, the manual observation of appliance measurements is an elementary step. Appliances of many types need to be measured consistently with high quality and redundantly in an isolated environment to enable the extraction of appliance-specific features.

1.3 Approach

Appliance Events play an important role in the NILM process. They define amongst other things, the relevant time segments where an appliance has a transition from ON to OFF or vice versa. To properly identify appliance classes, a reasonable event detection and therefore a distinct event definition are necessary. Usually, hard-coded thresholds are used to define appliance events [17, 18, 19], making them easy to identify with a simple rule set. In the context of consumption feedback as well as from the consumer perspective, appliance ON / OFF events that have a causal origin (i.e., from user interaction or physical appliance state changes) are more relevant than transients that simply satisfy simple rules or pass thresholds. Therefore, for our event detection, we replace hand-crafted rules with a supervised, multivariate, binary classification to distinguish between unrelated event-like transients and actual user relevant appliance events. Our classification system learns from consumer-labeled appliance events to distinguish between consumer relevant appliance ON / OFF events and irrelevant event-like transients. A common challenge of event detection is to reduce the amount of false positives. Our two-step adaptive learning approach that is based on the boosting algorithm, ensures a relevant selection

of training samples for the event and non-event class by learning from false positives. Our experiments show that the algorithm can reduce the number of unrelated event-like transients (false positives) significantly.

One challenge in power disaggregation lies in finding an efficient set of features with a high discriminative potential on appliances with respect to the environment they are used in. A smart meter may not be equipped with high-performing processing unit. Therefore, it is necessary to equip the disaggregation system in a modular fashion with only the most relevant features. We present a wide set of features and show which appliance characteristics can be covered with each individual feature. We evaluated all features and several of their combinations on four common, publicly available high-frequency sampled energy datasets to get a priori results on isolated and aggregated household environments. To streamline the appliance recognition process, we composed an extensive list of 36 different appliance features derived from the current and voltage measurements of appliance start-up events.

A smart meter setup that includes appliance recognition and power disaggregation needs a well-chosen classification model configuration. We see three configuration scenarios, a smart meter could be rolled out with: a fully pre-delivery-trained appliance model with a high generalization potential, an exclusively consumer post-delivery-trained appliance model with a high specialization on locally existing appliances, and a mix of both appliance models. We present experimental results that consider four publicly available high-frequency sampled datasets: BLUED [20], UK-DALE [21], PLAID [22], and WHITED [23] to obtain reliable results for appliance recognition on high-frequency sampled measurements. Three experiments, based on these appliance events, are conducted to benchmark the future in-house smart meter configurations. With these results a suggestion on the smart meter configuration is given to ensure a high appliance classification performance.

Representation Learning with deep neural networks define the state of the art in several disciplines of learning from large datasets. Their main advantage lies in replacing a hand-crafted feature extraction with learning from raw data. We show how NILM can benefit from deep learning. We implemented multiple representation learning approaches and performed a broad comparison to several prevalent machine learning approaches on

2 publicly available datasets (UK-DALE [21] and BLOND [24]) for appliance recognition. The evaluations include an expert-aided, 212-dimensional hand-crafted feature extraction model, three baseline raw data processing models, four different classifiers and three deep neural network architectures with their network parameter configuration for household consumption data.

1.4 Contribution

Our contributions mainly divide into two steps of the NILM process: appliance event detection and appliance classification as well as the dataset WHITED. The following enumerations sum up the main contributions to these research fields.

The main contributions of our work regarding appliance event detection are:

i. We propose a supervised event detection that eschews hard-coded event definitions by learning from relevant samples of the consumer perspective. The adaptive training algorithm reduces false positive events from switched-mode power supply driven appliances.

ii. We critically discuss the high number of appliance event definitions in the literature, making event detection algorithms hardly comparable.

The main contributions of our work regarding appliance classification are:

i. We perform a comprehensive evaluation of a wide set of electronics-, audio-, general signal processing- newly developed appliance-specific features and multiple combinations to their discriminative potential for electrical appliances on four publicly available datasets.

ii. We perform a comprehensive evaluation of a wide set of existing and newly developed appliance-specific features and multiple combinations to their discriminative potential for electrical appliances on four publicly available datasets.

iii. We map three smart meter use cases to three different dataset-cross-evaluation strategies to gain insights on real-world appliance classification performances and show the individual performances of appliance features and appliance types in each use case.

iv. We introduce representation learning approaches for appliance classification including suitable parameter settings and show a broad performance comparison between prevalent machine learning approaches and modern representation learning approaches.

The main contributions of our work regarding energy consumption datasets are:

i. We propose a novel, high-frequency sampled dataset with focus on household appliance transients.

ii. We show the dataset integrity with experiments on the discriminative potential of appliance measurements.

Parts of the content and contributions of this work have been published in:

- M. Kahl, A. U. Haq, T. Kriechbaumer, and H.-A. Jacobsen. "WHITED - A Worldwide Household and Industry Transient Energy Data Set." In: *3rd International Workshop on Non-Intrusive Load Monitoring*. 2016
- M. Kahl, A. U. Haq, T. Kriechbaumer, and H.-A. Jacobsen. "A Comprehensive Feature Study for Appliance Recognition on High Frequency Energy Data." In: *Proceedings of the 2017 ACM 8th International Conference on Future Energy Systems* (May 18, 2017). e-Energy '17. Hong Kong, Hong Kong: ACM, May 18, 2017. ISBN: 978-1-4503-5036-5. DOI: 10.1145/3077839.3077845
- M. Kahl, T. Kriechbaumer, A. U. Haq, and H.-A. Jacobsen. "Appliance Classification Across Multiple High Frequency Energy Datasets." In: *2017 IEEE International Conference on Smart Grid Communications (SmartGridComm)*. 2017. DOI: 10.1109/smartgridcomm.2017.8340664

- M. Kahl, T. Kriechbaumer, D. Jorde, A. U. Haq, and H.-A. Jacobsen "Appliance Event Detection - A Multivariate, Supervised Classification Approach" (unpublished, submitted to ACM e-Energy 2019)
- M. Kahl, D. Jorde, and H.-A. Jacobsen "Representation Learning for Appliance Recognition: A Comparison to Classical Machine Learning" (unpublished, submitted to Transactions on Smart Grids 2019)

1.5 Organization

The rest of the document is organized as follows. Chapter 2 provides the background to relevant topics. Chapter 3 gives an overview of related work to the individual research fields. In Chapter 4, we introduce the WHITED dataset for appliance-specific feature extraction. Chapter 5 presents the multivariate appliance event detection algorithm. In Chapter 6, we present a comprehensive feature study for appliance recognition. Chapter 7 presents a comparison between classical machine learning and representation learning for appliance recognition. In Chapter 8, we evaluate three smart meter configuration setups for appliance recognition and finally conclude in Chapter 9.

2

Background

In this chapter, we give insights into the NILM process and the most relevant, high-frequency sampled energy datasets that are used for this work. Furthermore, we describe the relevant performance metrics and give a short introduction to machine and representation learning as well as the applied algorithms.

2.1 NILM Process

The event-based NILM process describes the whole retrieval process from unobserved energy consumption to disaggregated per-appliance statistics. According to K. D. Anderson, Bergés, Ocneanu, Benitez, and Moura [17], event-based NILM comprises four elementary steps: (1) Data Acquisition, (2) Event Detection, (3) Appliance Classification and (4) Energy Disaggregation (see Figure 2.1.1).

Each NILM step is a relevant subject of research with individual approaches and studies [9, 26, 27]. Studies regarded to the first mainly focus on hardware setup, sampling frequency, measurement environment and usually belong to measurement engineering. Studies on event detection and appliance recognition usually focus on algorithms that belong to anomaly detection, pattern recognition, and classification. Further integral parts include

the usage of signal processing as well as empirically and heuristically driven feature extraction. The energy disaggregation step - depending on the output of the previous steps - summarizes, calculates and orders all retrieved data to generate information for the consumer.

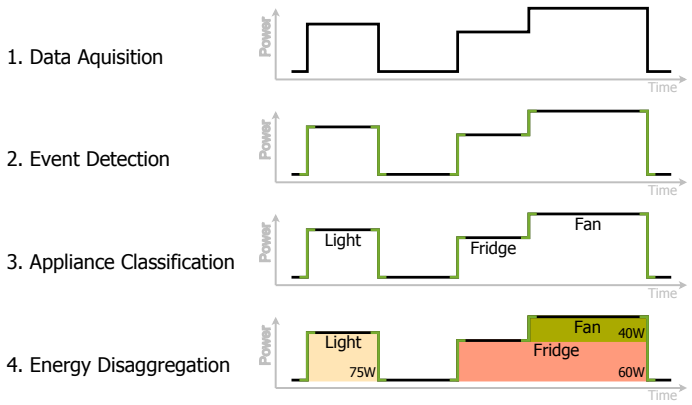


Figure 2.1.1: The general NILM process in four steps

2.2 Appliance Datasets

UK-DALE The UK Domestic Appliance-Level Electricity (UK-DALE) dataset consists of more than four years of energy consumption measurements for a residential building (house-1) with a high number of appliances of many different types. For our experiments, we considered measurements from 2013-04-22 to 2015-01-05. The dataset comprises low-frequency, non-equidistant sampled smart plug measurements (1/6 Hz) for each observed appliance (per-appliance signals) and high-frequency sampled measurements (16 kHz) from a custom sound card meter at the electric cabinet (aggregated signal). The per-appliance measurements allow a coarse determination of appliance events and power consumption to extract the relevant segments from the aggregated signal.

BLOND The Building-Level Office eNvironment Dataset (BLOND) comprises energy consumption measurements from an office building with a high number of appliances of only a few different types, mostly SMPS of office appliances. This appliance and appliance type distribution is the main difference to the UK-DALE dataset. The BLOND-50 subset

comprises 213 days of recording with 50 kHz sampling frequency for the aggregated signal at the electric cabinet and 90 individually observed sockets for the per-appliance measurements with 6.4 kHz sampling frequency.

BLUED The Building-Level fUlly-labeled dataset for Electricity Disaggregation (BLUED) is being introduced by K. Anderson, Ocneanu, Benitez, et al. [20]. The dataset contains continuous voltage and current measurements of around one week from a single-family household. The aggregated consumption signal is measured in a high amplitude (16-bit) and temporal resolution (12 kHz). Significant appliance state transients are labeled with timestamps and appliance information, to enable event detection research. The transient event ground truth stems from additional sensings such as light sensors and visual observation of humans.

PLAID The Plug Load Appliance Identification Dataset (PLAID) provides multiple appliance models of 11 appliance types. All appliances and events are used in our experiments (high inner-class diversity with 1074 events of 11 appliance types). PLAID is considered as a laboratory environment since the measurements are conducted in an isolated environment with the focus on the appliance events. The appliance events are fully labeled and sampled at 30 kHz with 16-bit amplitude resolution.

2.3 Performance Metrics

To evaluate the classification performance of multiple classes, three metrics can be stated as relevant: Precision (PR), Recall (RE) and F-Score. All three metrics can be retrieved from the classification confusion matrix. The metrics are calculated using the unweighted macro-average of all per-class results. To evaluate the classification performance for each class in a multi-class problem, PR, RE and F-Score are implemented using the class-wise True Positives (TP), False Positives (FP) and False Negatives (FN) as follows:

$$\textbf{F-Score} = 2 \cdot \frac{\mathrm{PR} \cdot \mathrm{RE}}{\mathrm{PR} + \mathrm{RE}} \qquad \textbf{PR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \qquad \textbf{RE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

2.4 Machine-, Representation- and Deep Learning

Machine learning is a research field and part of the general term artificial intelligence to retrieve knowledge from experience. In practice, machine learning describes a collection of algorithms that allow forming an internal, generalizable model from observations. With the transfer of learning, machine learning is a superior alternative to rule-based systems due to the ability to find patterns and principles in data allowing it to predict from unknown and noisy data. Features with discriminative potential are retrieved from given observations to learn from data. This process is called feature extraction and is usually conducted by humans with specific domain knowledge. Representation learning is a further subfield of machine learning with the advantage of automated feature extraction from given observations. In some contexts, a hierarchical decomposition of the entity is a further strategy to retrieve features from different arbitrary perspectives to the entity. This hierarchical inspection of represented observations characterizes deep learning, which is currently state of the art in speech and visual object recognition [28, 29].

Autoencoder (AE) are feedforward nets with a different number of neurons in the inner coding layer. The goal of an AE is to reach the same output as the input by propagating the input through the different dimensional coding layer in the middle. In our case, the AE's target is to reduce the number of representative neurons in the inner coding layer. The AE is built of an encoding section in which the input data is reduced and a decoding section in which the reduced codings are upscaled to reproduce the input. The output of the encoding section can be seen as a lower dimensional representation of the input [30] that went through a bottleneck, keeping only the essence of the data.

Convolutional neural networks (CNN) are designed to process signals that follow the principle of locality. Natural signals can be efficiently processed by local connections, shared weights, pooling and the use of multiple layers [28]. CNNs benefit from the typical hierarchical composition of natural signals.

Convolutional Autoencoder (CAE) follow the same underlying architecture as a standard AE. The hidden layers are replaced by convolutional layers, inheriting the advantages of locality, shared weights and pooling.

3

Related Work

We begin with the publicly available high-frequency sampled energy datasets and what is missing to extract appliance specific features. We continue with recent approaches on appliance event detection and discuss existing discriminative appliance features. The Chapter ends with the presentation of studies to deep neural network on NILM and cross-dataset evaluation.

3.1 High-Frequency Sampled Appliance Datasets

Several public datasets covering appliance-level energy consumption already exist. The purpose of these datasets is to measure demand in private households through a non-intrusive single point measurement in either low or high frequency. Through constant observation of household energy demand, these datasets provide comprehensive longtime measurements to cover user behavior in the corresponding residence. These datasets are a good source for power disaggregation tasks as they indirectly provide transient start-up features at an appliance level. Real-world scenario datasets include REDD[31], UK DALE [21] and BLUED[20] among others.

When looking in more detail at appliance transients, it can be cumbersome to extract

them from these single measurements. Since the ground truth is mostly based on 1s to 6s data without explicit voltage or current waveforms, it might be possible that two start-ups fall in the same time window, thus, violating the assumption of the switch continuity principle (SCP) [32]. Therefore, it is helpful to take a closer look at transient-focused datasets such as PLAID [22] and HFED [33]. PLAID examines start-up transients at 30 kHz whereas in HFED short transient spectral traces of up to 5 MHz were observed but require high effort in terms of hardware and experimental setup to reproduce.

Table 3.1.1: Comparison of datasets with high-frequency sampled appliance traces

			Appliance		
Dataset	Bit	Fs	Classes	Variety	Purpose
REDD [31]	24	15 kHz	~ 20	10	house demand
BLUED [20]	16	12 kHz	~ 30	~ 1	house demand
UK DALE [21]	20	16 kHz	~ 40	~ 1 – 3	house demand
PLAID [22]	16	30 kHz	~ 12	~ 20	transients
HFED [33]	16	5 MHz	15	1	spectral traces
WHITED [23]	16	44 kHz	46	1 – 9	transients
BLOND [24]	16	50/250 kHz	16	1 – 17	office demand
COOLL [34]	16	100 kHz	12	1 – 8	transients
LILACD [35]	16	50 kHz	15	1	industrial transients

Table 3.1.1 gives a comparison between the above mentioned high-frequency sampled datasets in terms of resolution, purpose, amount of appliance types (classes) and quantity of appliances for each class (variety). The information about the appliance types and quantity are inferred from the available data. We believe that a high intra-class variety leads to a more reliable result in terms of appliance classification.

With WHITED - a Worldwide Household and Industry Transient Energy Dataset - we want to contribute to existing energy datasets in terms of higher sampling frequency and higher amount of appliance types and variety. In addition, we provide a region classification for each measurement to potentially enable the investigation of region specific research questions.

3.2 Appliance Event Detection

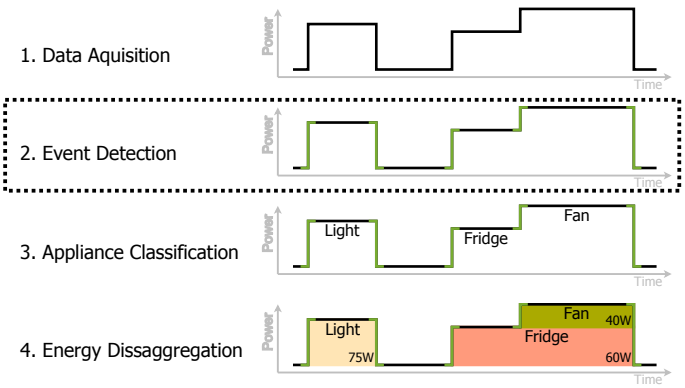


Figure 3.2.1: The focus here lies on event detection of the general NILM process

Appliance event detection can be seen as the second step in the NILM process, see Figure 3.2.1. Multi-state and SMPS-driven appliances often show unrelated event-like transients due to appliance state changes. These transients can be caused, amongst others, by computers that switch spontaneously from idle to full processor load (see Figure 3.2.2). Organic-LED-driven monitors have an image dependent energy consumption that can switch from minimum load to maximum load in between milliseconds just by changing from black to white in the displaying image. These undesired or unrelated transients affect the appliance classification and power disaggregation performance and make the event detection a challenging part. Rule-based event detection algorithms would need a complex rule set that is hardly feasible and sensitive to environment changes or appliance set changes due to their inflexibility.

Event Detection

NILM is commonly divided into event-based and state-based approaches. Event-based approaches rely on using detection algorithms in order to find electrical events such as switch-ON or switch-OFF of an individual appliance. State-based methods on the other hand, take into account every sample of the signal to perform the inference step. Event-based methods are generally more efficient in the inference step than state-based approaches. This efficiency is caused by pre-processing of the voltage and current signals with labeling and extracting the regions of interest of the signal after the events have

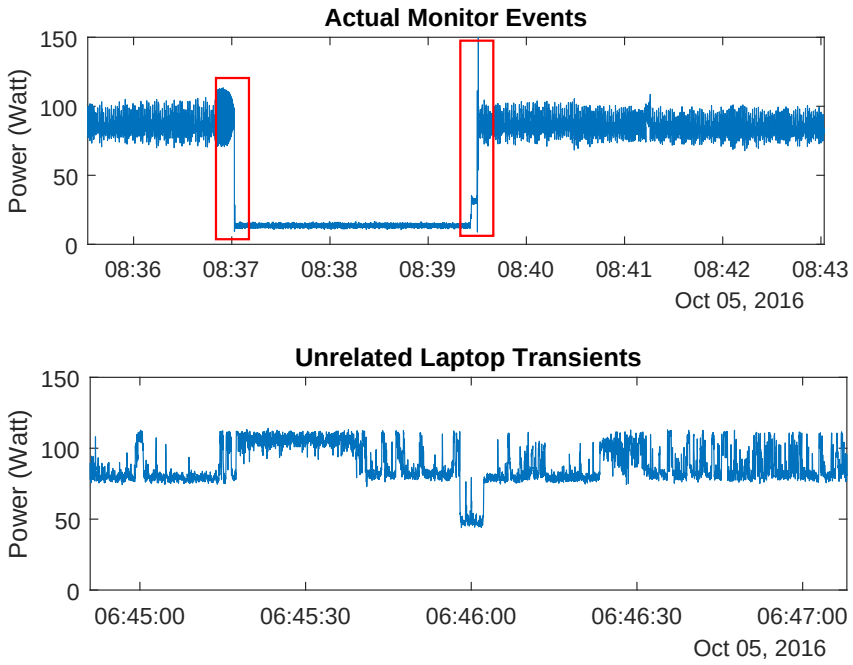


Figure 3.2.2: The first plot shows an actual OFF, followed by an ON event of a monitor. The second plot shows sudden laptop transients, that most likely stem from processor load changes. The goal is to differentiate between actual ON / OFF events and transients that are irrelevant to the user.

occurred. [36]. Most of the event-based methods rely on the switch continuity principle [32], which was initially introduced by Hart in 1992 [1]. It essentially states that there is only up to one event, i.e., not multiple ones, at a given point in time. Furthermore, it assumes that events are relatively rare when looking at the overall signal, allowing to see the event detection as anomaly detection. Sampling data at higher rates increases the validity of this principle. Employing this principle allows event-detection methods and other algorithms to treat electric events as being isolated from one another [32].

Three categories of event detection approaches are introduced by K. D. Anderson, Bergés, Ocneanu, Benitez, and Moura [17]. Expert heuristics describe mostly rule-based approaches that consider prior knowledge to define sets of parameters and thresholds [1, 37]. Probabilistic models consider statistical metrics, including variance and standard deviation, to estimate the probability of a change in a time series [38, 39]. Approaches of the matched-filter category try to find a universal event pattern in the signal by exceeding a likelihood threshold [40, 41]. The approach of K. D. Anderson, Bergés, Ocneanu, Benitez, and Moura [17] considers the usage of a modified general likelihood ratio detector to

compare four different evaluation metrics.

Baets, Ruyssinck, Deschrijver, and Dhaene [18] apply a cepstrum smoothing high-pass filter to the signal. This way, only very low frequency and step changes remain in the signal. The assumption is that in the case of an event, all remaining low frequencies lie above a certain threshold. The optimal parameter values were empirically evaluated. Baets, Ruyssinck, Deschrijver, and Dhaene [18] compare the results on the BLUED dataset with the chi-squared goodness-of-fit (X^2 GOF) approach by Jin, Tebekaemi, Berges, and Soibelman [42] and could reach comparable results.

Barsim, Streubel, and B. Yang [43] introduced an unsupervised event detection algorithm which creates the logarithm of the P, Q plane [1] to find steady states as clusters, while transients are represented as single scatters or outliers. The extraction of actual events was performed in three stages: a coarse search, followed by a fine search, and a final verification stage. The unsupervised way has the advantage that no learning from existing ground truth is necessary. The results show a very similar performance compared to Baets, Ruyssinck, Deschrijver, and Dhaene [18].

Wild, Barsim, and B. Yang [44] introduce a new event definition which gives events a dimension in time, they are not infinite anymore. This definition allows a Fisher discriminant analysis in combination with some constraints a robust unsupervised appliance event detection in the spectral domain.

Houidi, Auger, Sethom, et al. [45] investigate three commonly used techniques for the abrupt event detection that are typically used in other research fields: the Effective Residual algorithm [46], the Cumulative Sum (CUSUM) algorithm [47], and the Bayesian Information Criterion algorithm [48]. These algorithms are probabilistic event detection techniques. By comparing the algorithms in a real-world environment, Houidi, Auger, Sethom, et al. [45] conclude that the CUSUM algorithm outperforms the other two and achieves good results on their internal dataset.

Azzini, Torquato, and Silva [2] introduce the "window with margin" method. This threshold-based algorithm uses a sliding window and a subset of the samples within the window, i.e., samples from the beginning and the end of the window, to calculate two averages of the active power consumption. Azzini, Torquato, and Silva [2] then use

heuristically defined thresholds to check if the difference between the averages exceeds a certain limit in order to detect events in the signal.

The event detection methods above are developed for residential settings, whereas Leeb and Kirtley [49] propose a multi-scale transient event detector for industrial settings. To tolerate overlapping events, the author's algorithm searches for time patterns of segments in the signal that exhibit significant variation instead of searching for complete transient shapes. The algorithm detects such segments by using a change-of-mean detector. The transient changes in the signal are then detected by using sets of the previously computed segments as features for particular events and a pattern matching algorithm.

In contrast to the majority of the event detection approaches, R. Cox, Leeb, Shaw, and Norford [50] do not use current signals and analyze only aggregated voltage measurements. By using a spectral decomposition of the voltage signal to compute the harmonic voltage distortion, they are able to detect residential appliance events reliably. They further show that the voltage signal exhibits sufficient information to identify events.

All mentioned approaches have in common that all significant transients are interpreted as events. Every approach considers another event definition making it hard to compare their results. They do not allow to distinguish between different kinds of events or ignore undesired events.

Table 3.2.1: Event detection results on BLUED, using different event definitions making the results hardly comparable

Work of…	F-Score
Baets, Ruyssinck, Deschrijver, and Dhaene [18]	80.04
Jin, Telebakemi, and Berges [39]	81.01
Wild, Barsim, and B. Yang [44]	89.15

Event Definition

Regarding the event definition itself, multiple different interpretations of events can be found in the literature. Wild, Barsim, and B. Yang [44] present a classical and an extended

event definition. A classical event is a "transient from one steady state to another steady state which definitely differs from the previous one" [44], while an extended event describes a "so-called active section where the signal is somehow deviating from the previous steady state" [44], which provides a higher resilience against peaks and short pulses. K. D. Anderson, Bergés, Ocneanu, Benitez, and Moura [17] define an event with a state change of 30 W for a certain amount of time in a concrete value-based way, while Jin, Tebekaemi, Berges, and Soibelman [51] see event detection as a way to find ON and OFF transients of appliances. Girmay and Camarda [19] see an event as an active region from any appliance activation in which the power consumption is "well above" the background power.

The list of definitions above shows that there is no common agreement on what an appliance event can be. The event detection performance depends strongly on the event definition itself. A simple definition that includes a significant change of power for a certain amount of time, regardless of the cause, can simply be put into a rule-based system that may allow for a perfect detection performance. From the consumer perspective, appliance ON / OFF events that have a causal origin (i.e., from user interaction or physical appliance state changes) are more relevant than transients that simply satisfy the rule set. In practice, the consumer might be interested in the fridge or washing machine spin cycles. The temporarily increased energy consumption from a laptop during an irregular 5 minute lasting operating system update or the suddenly content dependent energy consumption of an organic-LED-driven TV is only of minor interest to the consumer.

Our approach avoids a distinct, hard-coded appliance event definition by learning from individual consumer-configured appliance event segments to build a tailored event model. This way we step back from a distinct event definition in favor of a user-definable event model. Since events from different appliances show individual characteristics, a rule-based approach with thresholds may not be sufficient to find ON / OFF switches. Our system is able to learn from different event features in the time and spectral domain which are fed as features into a supervised binary classification system. To improve the classification performance, we introduce an adaptive training technique that learns from previously wrong detected transients that lie on the border between events and non-events.

3.3 Discriminative Appliance Features

Machine learning and pattern recognition is an efficient and widely used approach for many NILM research questions, especially for appliance identification. The appliance identification problem is often tackled with a classification task, using supervised training on existing, with ground truth-labeled appliance measurements. These types of machine learning approaches were already used in 1994 by Roos, Lane, Botha, and Hancke [52] with neural networks and are still commonly used in more recent papers [14, 31, 53, 54].

Since appliances have individual characteristics, electrical power quantities including *Active Power*, *Reactive Power*, and *Apparent Power* may not be sufficient for all kinds of appliances. In 1992, NILM pioneer George Hart [1] described further signatures including harmonics and transient features for different appliance types. Several signal processing metrics of the temporal and spectral domain have also been applied to NILM over the years.

Further approaches, like spectral analysis using a wavelet transform, were implemented in 1995 by Leeb, Shaw, and Kirtley [40] to build a prototype detector that "performs remarkably well" and was able to identify four appliance types. The *V-I Trajectory* was first used in 2005 by Ting, Lucente, G. S. Fung, W. Lee, and Hui [55] and in 2007 by Lam, G. Fung, and W. Lee [56] for appliance classification and taxonomy purposes. Waveform-based and general signal processing metrics, including *Total Harmonic Distortion* and *Crest Factor*, which "provide a tremendously improved recognition capacity" were used in 2007 by H.-T. Yang, H.-H. Chang, and C.-L. Lin [57]. In 2011, Y. H. Lin, M. S. Tsai, and Chen [58] used the *Crest Factor* feature among others to reach identification rates of "higher than 93%" for three different appliances. S. Gupta, Reynolds, and Patel [59] used electromagnetic interferences in the MHz range to identify appliances with an accuracy of around 94% for 7 to 20 appliances.

The state of the art in NILM comprises numerous approaches, features, experiments and results. A comparison between these approaches is either difficult, due to differences in datasets, data acquisition equipment, appliance models and other factors. Armel, A. Gupta, Shrimali, and Albert [5] review numerous disaggregation algorithms and requirements for smart meters. Their feature related comparison focuses on different sampling frequencies

rather than on the individual feature performances.

The main goal of the energy disaggregation framework NILMTK from Batra, Kelly, Parson, et al. [60] is to allow a consistent comparison of different disaggregation strategies. As of now, NILMTK unfortunately appears to support only low frequency disaggregation. Due to this limitation, NILMTK cannot be used to evaluate approaches for high-frequency sampled energy data.

The work of Froehlich, Larson, S. Gupta, et al. [61] compares some high and low frequency features for disaggregation based on several criteria including installation, costs, sensing technology and ease of calibration. The evaluation criteria focus on the usefulness of the features in terms of environmental situations rather than performance in appliance recognition.

Gao, Kara, Giri, Berg, et al. [62] evaluate several features for PLAID on 5 classifiers. The best results were achieved with a random forests classifier using the VI image feature. The authors state that the combination of features improves the classification performance, which motivates to take a deeper look into feature combinations. Since the authors compare only the most common features, several strong features such as *Wavelet Analysis* are not evaluated.

With the increasing amount of NILM approaches, the necessity for comparability of these works is also growing. The aim of this paper is to give a wide overview of established and novel appliance features, including their stand-alone and combined classification performances for appliance recognition. This work orients on the related work in terms of identification system architecture, typical appliances from common datasets and a wide set of established features. Furthermore it provides a contribution to a better comparison of NILM studies.

3.4 Deep Neural Networks and NILM

Appliance classification can be seen as the second step in the NILM process, see Figure 3.4.1. The NILM community evaluated several classifiers in recent years. Hidden

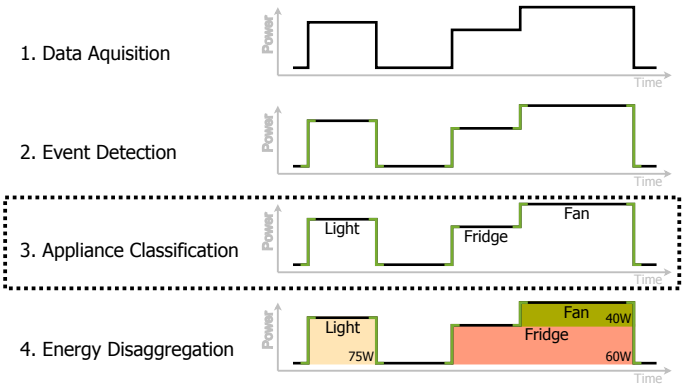


Figure 3.4.1: The focus of this work is on an appliance classification of the general NILM processing-chain. The actual appliance switch-on and switch-off events are retrieved with the help of the provided metadata and low-frequency measurements of the datasets and therefore considered as known in advance.

Markov Models are used in the work of Kolter and Jaakkola [63] and Zhong, Goddard, and Sutton [64] for appliance disaggregation, while Kramer, Klingenberg, Sonnenschein, and Wilken [65] and Du, Y. Yang, He, et al. [66] focus on K-Nearest Neighbor and Support Vector Machines for appliance recognition. NILM studies either belong to the low-frequency or high-frequency domain. Approaches that work on measurements sampled at less than 1 Hz are usually considered low-frequency while measurements with a sampling frequency more than twice as high as the mains frequency are usually characterizing the waveform and considered as high-frequency. Studies that evaluate approaches in the low-frequency domain often aim for solutions of actual energy provider driven smart meters, since their sampling frequency is usually limited due to privacy concerns. High-frequency sampled waveform measurements usually aim for in-house monitoring solutions, driven by the consumer. Armel, A. Gupta, Shrimali, and Albert [5] shows that an increase in sampling frequency also causes an increase in the number of appliances that can be distinguished.

Since waveform-based appliance energy consumption shares similarities to audio signals (signal envelope and appliance events), audio features can be successfully applied for appliance recognition [25] which motivates further studies on deep neural networks for appliance recognition. In computer vision, deep CNNs received a lot more attention due to the paper of Krizhevsky, Sutskever, and G. E. Hinton [67], who reduced the error rate for visual object recognition by almost half. The publication of G. Hinton, Deng, Yu, et al. [29] shows the performance improvements of deep neural networks in speech

recognition from four renowned research groups. CNNs can be successfully applied to visual and audio related classification problems.

J. Lee, Park, Kim, and Nam [68] propose an approach, based on a deep CNN for music tagging on sample-level (raw data). The results of the 10+ layer-sized deep neural networks are comparable to the previous state-of-the-art performances. W. Dai, C. Dai, Qu, Li, and Das [69] use very deep CNNs (up to 34 layers) to classify environmental sounds on raw data. The best architecture comprises 18 hidden layers and reaches the performance of a CNN with the audio spectrogram as input. The complex net architecture of the two approaches shows the potential for promising results on the one side, and that working on raw waveform data is challenging on the other side, especially in finding the right net architecture. Jorde, Kriechbaumer, and H. Jacobsen [70] propose the first approach on an appliance classification that uses deep neural networks on raw measurements. To overcome the issue of a small training-set, data augmentation and a one-against-all classifier composition were implemented to reach state-of-the-art classification performances.

Our approach considers the evaluation of three deep learning architectures (AE, CAE, CNN) in comparison with a comprehensive classical machine learning approach that uses 36 hand-crafted features. The AE and CAE are used for automated feature extraction from raw data, while the CNN is implemented as an end-to-end classification system to gain the full potential of deep learning architectures. The goal is to design an appliance recognition system that keeps the amount of preprocessing and domain-specific knowledge for feature extraction to a minimum, still reaching state-of-the-art classification performances.

3.5 Cross-Dataset Evaluation

Most algorithms in NILM are based on supervised machine learning algorithms to identify appliances and their consumption. Those approaches were already used by Roos, Lane, Botha, and Hancke [52] with neural networks and are still being used in recent works [14, 54, 71]. Several studies [43, 62, 72] in NILM retrieve results in an isolated environment with comparably smaller variances than real-world scenarios.

When focusing on appliance recognition, the idea behind a cross dataset validation lies in applying the trained model to a broader set of variances including measurement equipment, bit resolution and sampling rate of the measurements, and line noise. All these aspects differ along the existing datasets and have an influence on the resulting measurements and therefore also on the resulting features.

Jack Kelly wrote about the different formats of energy datasets: "This is an issue because an important criteria for evaluating any machine learning algorithm is how well it generalises across multiple datasets." [73]

S. Gupta, Reynolds, and Patel [59] implemented a solution for automatic detection and classification of electronic appliances, based on their high-frequency electromagnetic interference signal. One of their evaluations focuses on the stability of signatures across different homes. The characteristics of electromagnetic interference measurements are significantly different to high-frequency voltage and current measurements and are therefore not that easily applicable to a smart meter.

The main motivation for the NILMTK from Batra, Kelly, Parson, et al. [60] is the missing ability to generalize and compare NILM algorithms across different energy datasets. It allows for a consistent comparison of different strategies for power disaggregation across existing energy datasets in the low frequency domain.

The work of [25] consists of a comprehensive feature study that considers the most relevant features for appliance recognition in the high-frequency domain. This set is the foundation for the experiments presented in this paper.

To the best of our knowledge, there is no work that explicitly applies a cross-dataset-validation to retrieve a general appliance classification performance in the high-frequency domain. Another interesting research question is a comparison of features which are able to generalize enough to enable a cross-dataset appliance classification. We aim to answer these two questions with this work.

4

Worldwide Household and Industry Transient Energy Dataset

In this chapter, we introduce a dataset of appliance start-up measurements from several locations. The appliances were recorded with a low-cost custom sound card meter. The recording was mainly done in households and small industry settings in different regions around the world. Thus, it may be possible to extract region-specific grid characteristics from the voltage waveforms in the data. To cover all corresponding transients, we recorded the first 5 seconds of the appliance start-ups for 110 different appliances to date, amounting to 47 different appliance types. The aim of this dataset is to provide a broad spectrum of different appliance types in regions around the world.

4.1 Hardware Components

Our measurement equipment is based on a sound card as inexpensive analog to a digital converter. The idea of a sound card-based measurement system is not new and was already used in [21] and [74]. Sound cards have a very good price vs. performance ratio when using them as an analog to digital converter. Our measurement prototype is based on a modified 3-port extension cord, a current clamp, an AC-AC transformer, a voltage

divider, and an external USB sound card with a *Cmedia CM6206* chipset.

For measuring the current, we use a YHDC current clamp with built-in burden resistor. This current clamp produces a 1 V signal at 30 A primary current. For the voltage measurements, we need to transform the grid voltage from 230 V to 11 V with the AC-AC transformer. To have a corresponding voltage signal that lies in the line-in range of the sound card, we reduce it with a voltage divider to 0.47 V. The voltage divider is located in the black isolation part that merges the current and voltage signal cables into one cable that goes into the sound card. See Figure 4.1.1 for the complete configuration.

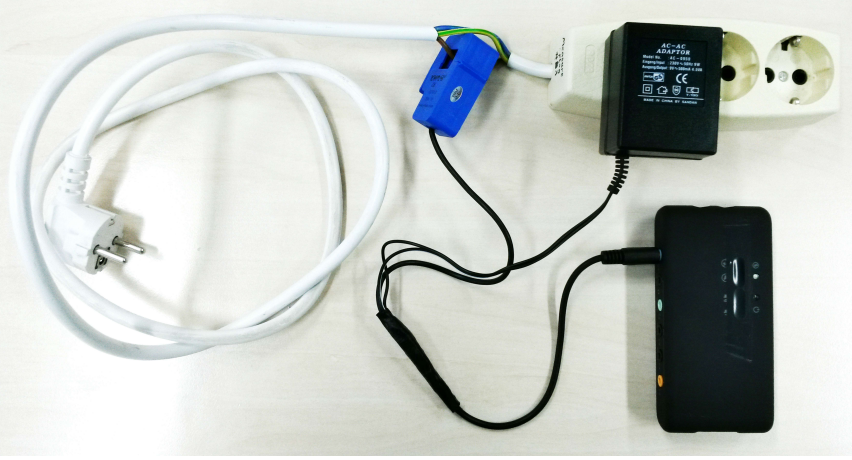


Figure 4.1.1: Measurement equipment prototype

4.1.1 Measurement Methodology

The signals were recorded in 44.1 kHz temporal and 16 *bit* amplitude resolution. To be able to take multiple measurements in different places, it was necessary to build 3 identical measurement kits. Therefore, we also have to deal with three slightly different sets of calibration factors. The calibration itself is done with an *VOLTCRAFT VC-330* multimeter. Since the multimeter provides current measurements with a current clamp, it was possible to measure both signals - voltage and current - and define a voltage and current calibration factor for each measurement kit. Some sample measurements can be seen in Figure 4.1.2 for 4 different appliances.

To cover the start-up transients of the appliances, it is necessary to determine them on

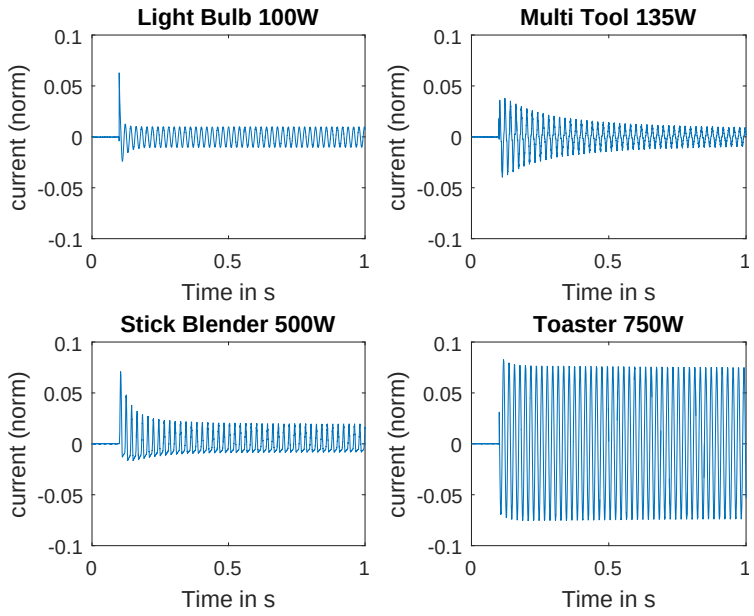


Figure 4.1.2: Start-up of four different appliances. The different in-rush current characteristics are clearly visible.

demand. This is implemented with a *Matlab* routine that uses the internal DSP package to monitor the line-in signal of the sound card. The start-up is defined based on the current signal energy crossing a threshold. If the current signal energy leads to a start-up, the routine starts recording and adds 100 ms of the signal beforehand as pre-start-up window. This window allows difference-based algorithms to work effectively. That means that not the absolute power consumption on the start-up but the difference between the power of the pre-start-up window and the start-up power can be observed. This approach introduces more flexibility for developing algorithms that allow the recognition of concurrently running appliances with different start-up transients.

We decided to measure 10 start-ups for each appliance. These start-ups were triggered manually by the user. Appliances that have no switch (e.g., an iron) were just plugged and unplugged 10 times as it would be the case under real usage. The appliances are measured for 5 seconds which is the duration of each start-up we recorded.

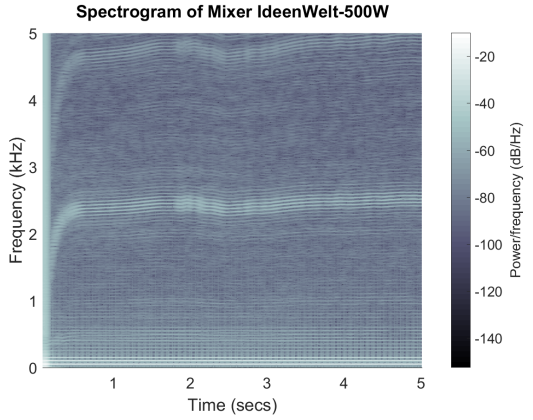


Figure 4.1.3: Mixer

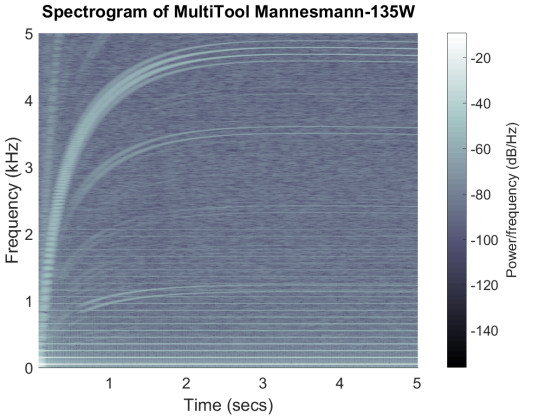


Figure 4.1.4: Multi-Tool

Figure 4.1.5: Comparison of 2 appliances that use motors with relatively high rotations per minute. The different spectral characteristics are clearly visible. The multi-tool has stronger uneven harmonics while the harmonics are more equal in the case of the mixer.

4.1.2 Dataset

To this end, our dataset comprises 1100 different records for 110 different appliances which can be grouped into 47 different types (classes) in 6 different regions. For most appliances, we took a photo of its electrical specification label. These images are located in the sub-folder `images` and `type-labels`. Table 4.2.1 gives an overview of the measured appliances. The signal containing files are saved as `flac` files – a common lossless audio file format. The file names contain meta information and are of following format:

```
[Class]_[Name]_[Region]_[#Kit]_[TimeStamp].flac
GuitarAmp_Marshall8240_R3_MK2_20151115133402.flac
```

The dataset is freely available on the following web page: `https://www.i13.in.tum.de/index.php?id=114`. For demand, load and appliance information retrieval, the most important signal is the current. To give the voltage signal a higher significance, we decided to measure the voltage in several regions that follow the European grid standards. To this end, the dataset contains 4 regions in Germany, 1 in Austria, and 2 in Indonesia.

Since grid characteristics are mainly affected by utilities and the consumption characteristics of the surrounding area, a future research direction is to look for possibilities to determine the region from the voltage signal. This experiment is a similar classification task to the appliance recognition we have already implemented.

4.2 Evaluation

To ensure the quality of the dataset, we applied several signal quality checks and conducted two classification experiments.

4.2.1 Data Quality

Since sound cards do not provide a high level of linearity in frequency response as compared with professional ADCs, we verified that there is no significant impact on the measurements taken.

The sound card manufacturer provides some information regarding the line-in linearity. The strongest damping of around 0.25 dB has its maximum at 3320 Hz. The steepest flank has a bandwidth of around 3300 Hz and lies between 3320 Hz and 6622 Hz which is acceptable for most considered purposes.

To obtain an approximation of the noise level during recording, the energy of a 10 second empty signal is being compared to the energy of a maximum amplitude sine-wave signal. With this calculation, we estimate an effective SNR (signal to noise ratio). We measured an average noise RMS of 4.8 mA where 30 A corresponds to the RMS maximum.

$$SNR = 20 \cdot log_{10} \frac{RMS_{max}}{RMS_{noise}}$$

$$SNR = 20 \cdot log_{10} \frac{30\,\text{A}}{0.0048\,\text{A}} = 75.91\,\text{dB}$$

The effective SNR of this measurement system is 75.91 dB. The maximum measurable peak to peak current I_{p-p} is 30.0 $A_{RMS} \cdot 2\sqrt{2}$ = 84.4 A. Therefore, we calculate an effective current resolution with a step size of 13.5 mA.

$$I_{step} = \frac{I_{p-p}}{Imax_{RMS}} \cdot Inoise_{RMS}$$

$$I_{step} = \frac{84.4\,\text{A}}{30.0\,\text{A}} \cdot 0.0048\,\text{A} = 0.0135\,\text{A}$$

This current step size enables us to calculate the effective power step size P_{step} corresponding to 230 V of grid voltage.

$$P_{step} = 230\,\text{V} \cdot 0.0135\,\text{A} = 3.1\,\text{W}$$

The resolution and noise of the sound card allows a voltage step of 0.313 V, a current step of 0.0135 A which results in a measurable power step of around 3.1 W based on 230 V. To achieve reliable results only appliances with a consumption of at least 20 W are considered in our data set. This covers most household and small industry appliances.

Figure 4.1.3 and 4.1.4 show a spectrogram of a mixer and a multi-tool based on the first 5 seconds after the start-up. Both appliances have a fast spinning motor and look similar in the time domain. However, there are significant differences in the spectral domain that can be transformed into distinguishable features for appliance classification purposes.

4.2.2 Experimental Results

Our appliance recognition experiment is based on a classification task to distinguish appliances on its characteristics in the current signal. The classifier has to distinguish between all 47 appliance types. The classification experiment is implemented in Matlab. All `flac` files are imported and the containing signal is scaled with the corresponding

Table 4.2.1: Appliance types (classes) that were measured

Type	#	Type	#	Type	#
AC	1	Air Pump	1	Bench Grinder	1
CFL	2	Charger	7	Coffee Machine	1
Deep Fryer	1	Desktop PC	1	Desoldering tool	1
Drilling Machine	2	Fan	6	Fan Heater	1
Flat Iron	2	Game Console	4	Guitar Amp	1
Hair Dryer	6	Halogen Fluter	1	Heater	1
HiFi Rack	1	Iron	3	Jigsaw	1
JuiceMaker	1	Kettle	6	Laptop	1
Laserprinter	1	LED Light	9	Light bulb	6
Massage tool	3	Microwave	2	Mixer	4
Monitor	2	Mosquito Repellent	1	Multitool	1
Powersupply	4	Projector	1	Sewing Machine	1
Shoe warmer	2	Shredder	2	Soldering Iron	2
Toaster	4	Treadmill	1	TV	1
Vacuum Cleaner	4	Washing Machine	1	Water Heater	4
Water Pump	1				

calibration factors to determine actual values. After this preprocessing step, a region of interest (ROI) needs to be extracted. Here, we decided to cut the signal right on the start-up until 500 ms after the start-up. These 500 ms samples are given to the feature extraction stage which is an implementation of 13 different characteristics including harmonics, phase shift and total harmonic distortion (THD).

The best results we achieved for the appliance classification were based on a feature set that consisted of a period-based power trend with 25 dimensions, the THD and crest factor of the current spectrum with each 1 dimension in its size. With these three features in 27 dimensions, we achieve an average classification accuracy across all appliances of around 95 % with a 10-fold cross-validation and a support vector machine (SVM) classifier. This confirms the observation that power difference and harmonics contain sufficient information to distinguish among basic electrical appliances [75].

For the region classification experiment, we use the same environment but employ the voltage instead of current for the feature extraction. The labels are not the appliances but the region where the measurements were taken. We apply the voltage, grid frequency and

a few spectral- and waveform-based features. We obtain an almost perfect classification accuracy of 99.13 % with an SVM classifier. Here, we must consider that the feature extraction is based on characteristics that vary over time and are not independently representative for the corresponding region.

5

Appliance Event Detection and Discrimination

NILM is a modern and still expanding technique, helping to understand fundamental energy consumption patterns and appliance characteristics. Appliance event detection is an elementary step in the NILM pipeline. Unfortunately, several types of appliances (e.g., SMPS or multi-state) are known to challenge state-of-the-art event detection systems due to their noisy consumption profiles. Classical rule-based event detection system become infeasible and complex for these appliances. By stepping away from distinct event definitions, we can learn from a consumer-configured event model to differentiate between relevant and irrelevant event transients.

We introduce a boosting oriented adaptive training, that uses false positives from the initial training area to reduce the number of false positives on the test area substantially. The results show a false positive decrease by more than a factor of eight on a dataset that has a strong focus on SMPS-driven appliances. To obtain a stable event detection system, we applied several experiments on different parameters to measure its performance. These experiments include the evaluation of six event features from the spectral and time domain, different types of feature space normalization to eliminate undesired feature weighting, the conventional and adaptive training, and two common classifiers with its optimal parameter settings. The evaluations are performed on two publicly available

energy datasets with high sampling rates: BLUED and BLOND-50.

5.1 Multivariate Event Detection

A reasonable appliance classification and disaggregation performance will be achieved when the NILM system adapts to the deployed environment. The customization may include parameter settings of *base load*, *min/max appliance load* or *max concurrent running appliances*. Besides those parameters, a consumer supervised appliance labeling for system training purposes, over a certain amount of time (e.g., few days/weeks), will result in considerably improved classification and disaggregation performance [25].

Since the temporal appliance event positions are implicitly known from the consumer labeled time range, these event segments can be used to train a supervised event model for the event classification. The a priori known event segments can be used to identify significant event characteristics, which are a major advantage compared to hand-crafted rules. In a supervised classification task, the classifier needs training samples for each individual class. Event detection is related to anomaly detection that faces the problem of not having sufficient training samples for one of the targeting classes. In practice, we explicitly know from examples how an event looks like, but we don't explicitly know how a non-event looks like.

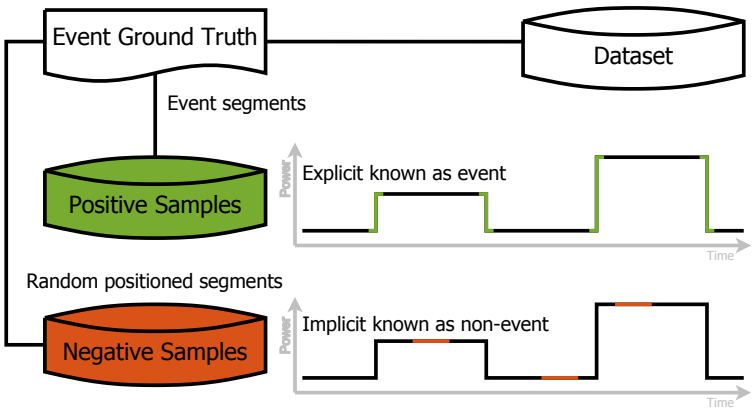


Figure 5.1.1: The explicitly-known events are retrieved from the event ground truth. Therefore, all other regions are implicitly-known as non-events.

To overcome that issue, we make use of the fact, that statistically the majority of the time, no event occurs in the signals time domain. We cut short, randomly positioned regions of the temporal signal from the training area, to use them as non-event samples (see Figure 5.1.1). The probability to hit an event on a randomly selected position in the training area of the temporal signal is low for common residential and office environments. Around 1 250 events occur per phase in one week for the residential environment while it is around 257 for the office environment, based on the utilized datasets BLUED and BLOND-50. Assuming we are interested in the same number of non-events as it is for events, the chance to hit an event via random selection lies at 0.83 % for the residential environment, while it is around 0.17 % for the office environment. To even overcome that small uncertainty, a minimum temporal distance to explicitly known events of minimal 10 s must be fulfilled. The resulting non-events will be named *implicitly-known non-events* throughout this paper. All samples together can be used to train a classifier with a training set that consists of explicitly-known events and implicitly-known non-events.

An observed issue with this approach lies in a high number of event false positives. The randomly selected non-event samples stem mostly from areas of a steady consumption. Therefore the non-event class is a good homogeneous representation of steady non-event areas. A more heterogeneous set of non-event training samples with unsteady event-like transients would be necessary to improve the classification performance of transients from SMPS-driven appliances in favor to non-events.

5.1.1 Adaptive Training

Extracting even more randomly selected samples would be one infeasible way to get a higher variance. The extreme form would be to use every extractable time window in the dataset that is not a ground truth labeled event. Obviously, this would create an infeasible number of training samples for the non-event class. However, the vast amount of training samples would be unnecessary anyway due to a very strong similarity.

Our approach is a so called boosting variant that runs the event detection algorithm on the whole training area to find all ground truth labeled events but also a certain amount of non-labeled transients. These transients are obvious false positives, based on the provided

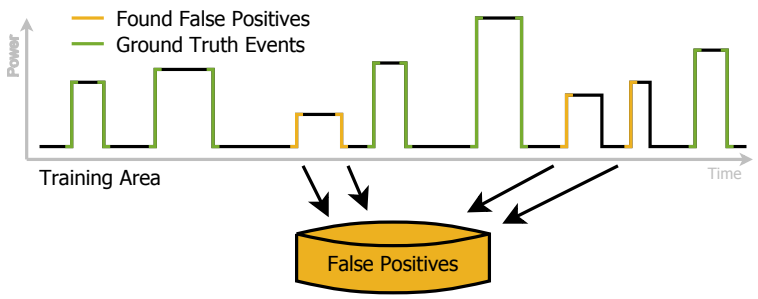


Figure 5.1.2: The event detection runs on the training area and generates false positives that are being stored for the actual event detection.

ground truth (see Figure 5.1.2). They are marginal, uncertain segments of non-events that share similarities with events. These similarities cause the misclassification in favor to the class *event*. Since these false positives are found inside the training set, we can use them freely to improve our classification model.

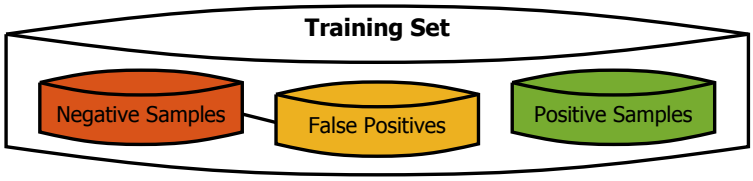


Figure 5.1.3: The collected false positives from the event detection of the training area form together with the negative samples the class non-events. The positive examples are the representatives of the event class.

The idea lies in adding these *edgy* transients to the non-events class of the training set to improve the border between events and non-events (see Figure 5.1.3). The actual training set consists now of ground truth labeled event samples, implicitly-known non-event samples, and false positives that were found in the event classification run on the training area itself. This way, it is possible to overcome the issue of finding proper non-event samples for the event detection algorithm. To even reduce the amount of false positives further, the adaptive training can be applied multiple times.

5.1.2 Event Features

The event ground truth information for BLUED is based on a power consumption change of at least 30 W over a time period of minimal 5 s [20]. Based on this definition, the

appliance events can be identified in a moving time window in the continuous electricity signals. We implemented one spectral and six time domain metrics as appliance event features for the classification between events and non-events. Our design defines that the actual event transient is being aligned in the middle of the extracted time window with 5 s of data before and after the actual event transient. The actual temporal position of the event transient is being extracted from the ground truth information or manual annotation in the case of BLOND-50.

The BLUED provided ground truth information and BLOND-50 annotations from this work, comprises the appliance ON and OFF switch events, including circuit number, temporal position (timestamp) and appliance type. The provided switch-OFF and switch-ON events of these appliances will always cause significant changes in these consumption-related metrics:

Current

The current is the first intuitive metric that contains consumption changes (see Figure 5.1.4-1). The RMS current I_{rms} for each period is calculated as follows, with N as the number of samples per period, calculated as the ratio of the sampling frequency fs and the mains frequency $F0$.

$$I_{rms}(p) = \sqrt{\frac{1}{N}\sum_{k=1}^{N} I_k^2}, \quad N = fs/F0$$

$$\vec{I}_{rms} = [I_{rms}(1), I_{rms}(2), \dots I_{rms}(nPeriods)]$$

Δ(Current)

Since multiple appliances can run at the same time, the actual pre-event current can be a sum of multiple appliances and therefore has a high variance (see Figure 5.1.4). The actual information of interest is the current step change at the event time (see Figure 5.1.4-2). This metric can be retrieved by the numerical difference of the neighboring elements

of the current periods $\vec{I}_{rms}$. The operation is the derivation equivalent for discrete time series.

$$\Delta\vec{I}_{rms} = \vec{I}_{rms_k} - \vec{I}_{rms_{k+1}}$$
$$\vec{I}_{rms_k} = [I_{rms}(1), \dots, I_{rms}(k-1)]$$
$$\vec{I}_{rms_{k+1}} = [I_{rms}(2), \dots, I_{rms}(k)]$$

Admittance

The grids voltage can contain high fluctuations (up to 10 %), which influences the current signal as well. The admittance removes the voltage influence from the current signal and is therefore more precise to the appliance consumption itself (see Figure 5.1.4-3). The admittance ADM, can be calculated by the element wise vector division of the period wise current $\vec{I}_{rms}$ and voltage $\vec{U}_{rms}$.

Spectral Flatness

Our motivation for the only spectral feature we considered is the assumption that all appliances have their individual fingerprints in their harmonic energy distribution. A suitable spectral one-dimensional metric is the spectral flatness. A flat spectral curve f_{bins} would cause a value close to one, while a single strong spike would lead to a value close to zero (see Figure 5.1.4-4). The switch-OFF and switch-ON of an appliance influences the spectral flatness in general way. The spectral flatness $SPF(p)$ for each period is calculated by the ratio of the geometric and the arithmetic mean of the current signal energy spectrum [76].

$$SPF(p) = \frac{\sqrt[N]{\prod_{f \in f_{bins}} x_f}}{\frac{1}{N}\sum_{f \in f_{bins}} x_f}$$

Cumulative Sum

The cumulative sum is a sequence analysis technique that allows to identify small and continuously slow as well as strong and fast changes in a sequential time series (see Figure 5.1.4-5). It is therefore a common technique for change and event detection. The cumulative sum is the sum of the differences to the mean of the signal in between a defined time window.

Δ(Cumulative Sum)

The cumulative sum can have extreme gains in their values and therefore causing undesired weighting of dimensions in the feature space. The derivative of the cumulative sum is a way to prevent this issue and to keep the values in a lower magnitude. The resulting signal is visually comparable with the current itself, but with enlarged transients (see Figure 5.1.4-1 and 5.1.4-6).

$$\Delta\vec{I}_{cms} = \vec{I}_{cms_k} - \vec{I}_{cms_{k+1}}$$

In addition to the mentioned features and training methods, we evaluated the event detection performance through different methods in the feature space normalization and classification step. To avoid undesired weighting across the dimensions of the feature space, a common technique is to apply a feature space normalization. This is often an essential step, of which we evaluate three types. The classification step is being evaluated with two different classifier (KNN and SVM) including their hyper-parameter search.

5.2 Experiments

To compare our event detection performance with state-of-the-art, we applied our algorithm on the BLUED dataset, which is commonly used for event detection evaluation. The experimental setup is oriented on the setup in the work of Baets, Ruyssinck, Deschrijver,

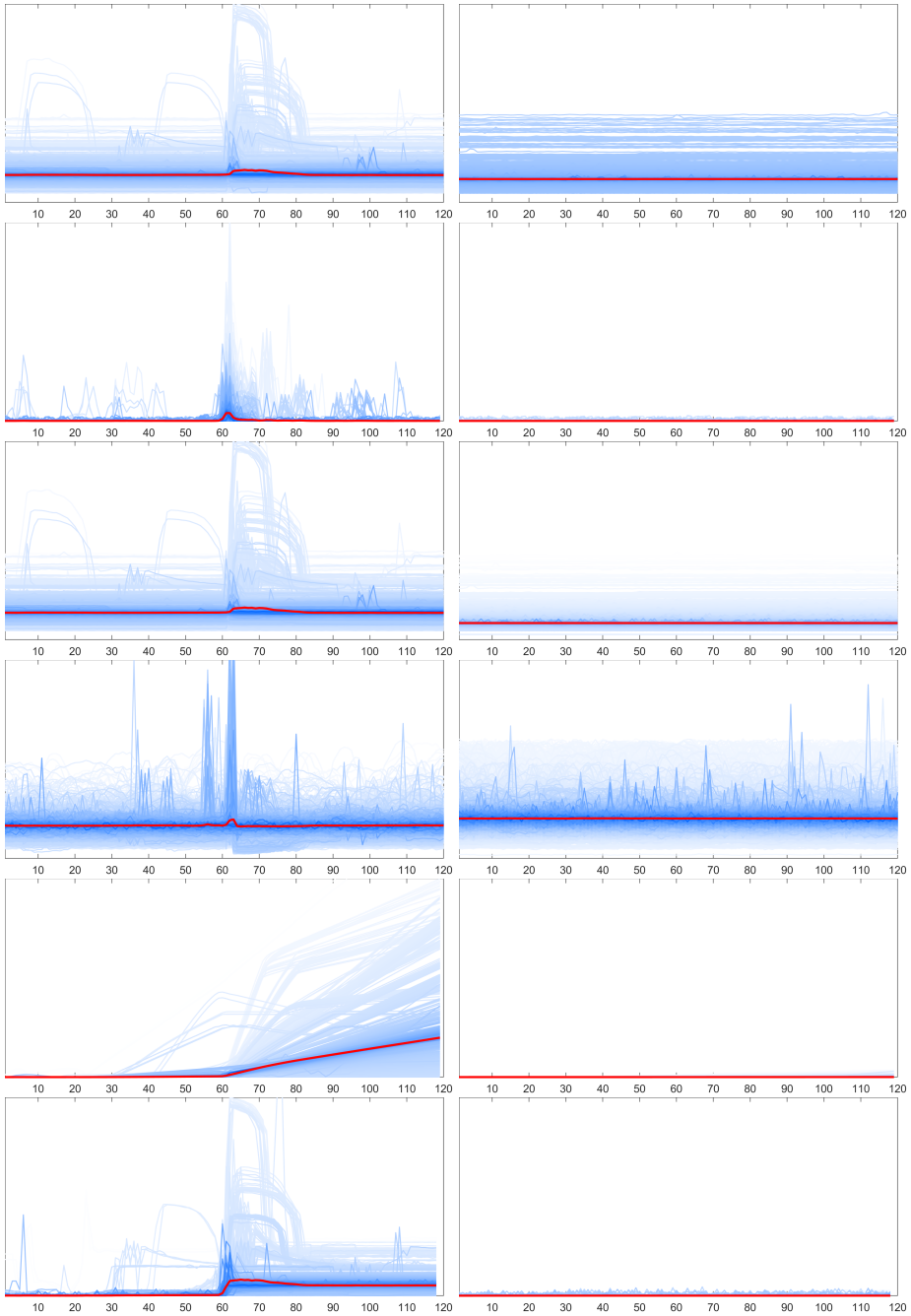


Figure 5.1.4: Events (left) and Non-Events (right) with Periods in the X-Axis and amplitude in the Y-Axis of the 6 event feature metrics: 1. Current, 2. Δ(Current), 3. Admittance, 4. Spectral Flatness, 5. CUSUM and 6. Δ(CUSUM). The color saturation correlates with the average distance to the mean event (red line). The closer the event lies to the mean-event the higher is the saturation.

and Dhaene [18]. While De Baets is using a fixed test area, we are using cross-validation for our performance evaluation. At least K. D. Anderson, Bergés, Ocneanu, Benitez, and Moura [17], Barsim, Streubel, and B. Yang [43] and Wild, Barsim, and B. Yang [44] evaluate their event detection algorithm on the BLUED dataset as well. For BLUED we use the provided ground truth information which stems from hand-crafted annotations.

Unfortunately, neither BLUED nor BLOND-50 provide versatile event information that allows a determination between ON / OFF-switching and user-unrelated transients. In our experiments on BLOND-50, we try to distinguish ON and OFF events from all remaining state transients - identical to the work of Baets, Ruyssinck, Deschrijver, and Dhaene [18]. The appliance ON and OFF events for the BLOND-50 dataset are being collected by visual observation of an instructed person with the help of a self-implemented annotation tool. There are no studies regarding event detection on BOND-50 yet.

Since the benchmark of several parameters using cross-validation takes much computational time, we use a cluster of 60 virtual machines, based on dual Intel Xeon E5-2630v3 with each four cores and 10 GiB RAM to execute the appliance event detection algorithm in parallel. The cumulative CPU time for all experiments, preprocessing and testing lies in a range of 128 000 CPU-core-hours.

5.2.1 Multivariate Event Detection

Instead of monitoring one or few parameters passing thresholds, our multivariate approach enables supervised learning of multiple event characteristics. The explicitly-known event, and implicitly-known non-event sections were used to train the classifier that decides, based on the given feature vector, between event or non-event.

Architecture for BLUED In addition to the 1 577 events, we extracted 6 428 segments of implicitly-known non-events (one for each file) of the same length. The segments are aligned with the ground truth event timestamp in the center. These segments are fed to the feature extraction and normalization after that. The normalization parameters (e.g., means or standard deviation) are saved to apply the corresponding transformation to the

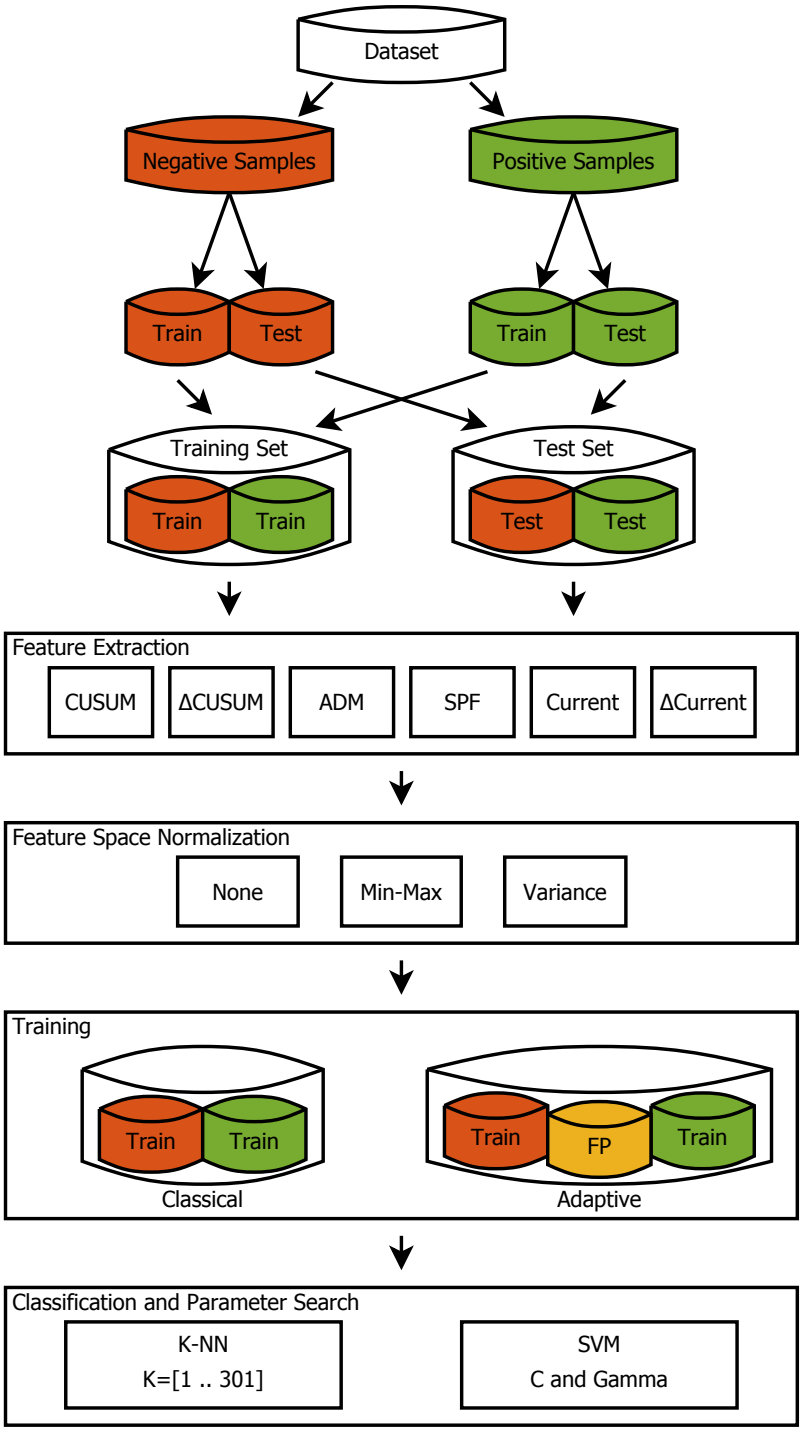


Figure 5.2.1: The architecture of the experimental setup considers the main step of the common machine learning pipeline and includes the evaluation of six event features, three types of feature space normalization, two training approaches and two different classifiers with its optimal parameters. The whole architecture is wrapped by a cross-validation and structured to run on a distributed computation system.

samples of the test area. The following steps include a parameter search for the classifier (e.g., C and Gamma for SVM), classifier training, and classification of the samples of the test area (see Figure 5.2.1). All experiments are implemented within a stratified k-fold cross-validation to ensure reliable results.

Architecture for BLOND-50 The manually annotated temporal time span comprises one month of measured data. We extract all manually annotated events and the implicitly-known non-events in a very similar way as we do for BLUED. This step yields in 3 310 event and 3 264 non-event samples. The events originate from 41 different monitored appliances in the time range of `2016-11-01` to `2016-11-30`.

5.2.2 Adaptive Training

The adaptive training shares the same experimental architecture as the multivariate event detection, with one additional event detection run on the training area itself and its false positives included to the training set. This training run finds events in the training area that can be divided, considering the ground truth information, into true positives, false positives, and false negatives. All false positive segments that originate from the training run are added to the non-event class of the actual training set.

5.2.3 Manual BLOND-50 Event Annotation

Every performance benchmark needs reference information to enable comparisons. For the event detection evaluation, an event ground truth including the exact temporal position of an appliance event is necessary. For the BLUED dataset, the appliance events are provided already, for BLOND-50, the appliance events and the corresponding measurement system, circuit and socket number need to be acquired.

To label the data with a ground truth, we were using a self-designed annotation tool that allows a manual annotation in the per-appliance subset of BLOND-50 (see Figure 5.2.2). The annotating person observes the data of one measurement system instance and all

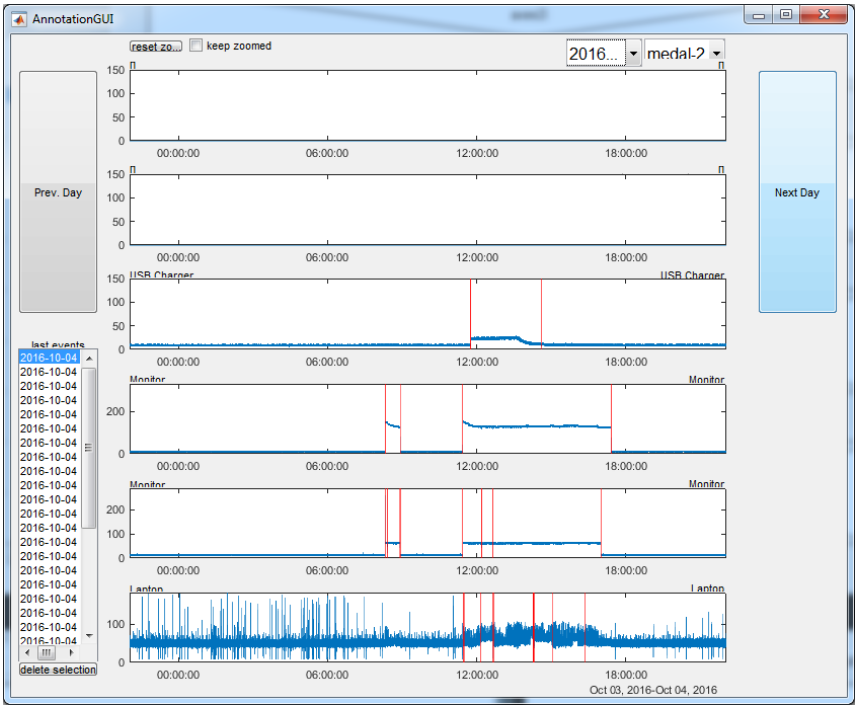


Figure 5.2.2: Annotation tool for BLOND-50 event ground truth annotation. The annotating person specifies the date and measurement system to view the corresponding consumption of the day for each of the six sockets. Zooming into the time series plot allows for a precise event annotation.

6 sockets for one day per screen. The two appliance event constraints (power-rise/fall of 30 W for a minimum time span of 5 s) are communicated to the annotating person to ensure consistent events. In addition, the annotating person is instructed to consider only obvious appliance ON and OFF events. Transients that fulfill event constraints but are not obvious switch ON and OFF events are ignored.

The event ground truth for BLUED and BLOND-50 originate from visual time series observation by humans. Therefore the experimental evaluations in this paper are not performed on the (non-existing) absolute truth but rather subjectively chosen time series segments of the human observation that always contain an individual degree of uncertainty. Since neither an event ground truth nor an appliance event definition has been chosen, the goal is to retrieve an appliance event model from user chosen examples declaratively, a degree of uncertainty from the human observation therefore does not play any role. The manually annotated events, as well as the corresponding annotation tool for MATLAB, can be downloaded at the following link.

5.3 Results

To ensure a consistent evaluation pipeline we decided to use the best parameter or settings from the previous steps. In practice, the evaluation of the normalization method is done with the best-performing feature of the feature evaluation. For all experiments, a search window step-size of 30 periods was used. To the nature of the algorithms, multiple events occurring in between a 5 s window (SCP violation) may be recognized as one event (see Figure 5.3.1). That circumstance causes a small number of false positives. The goal of all experiments is to find all ground truth labeled events (true positives) while keeping the misclassifications (false positives) to a minimum.

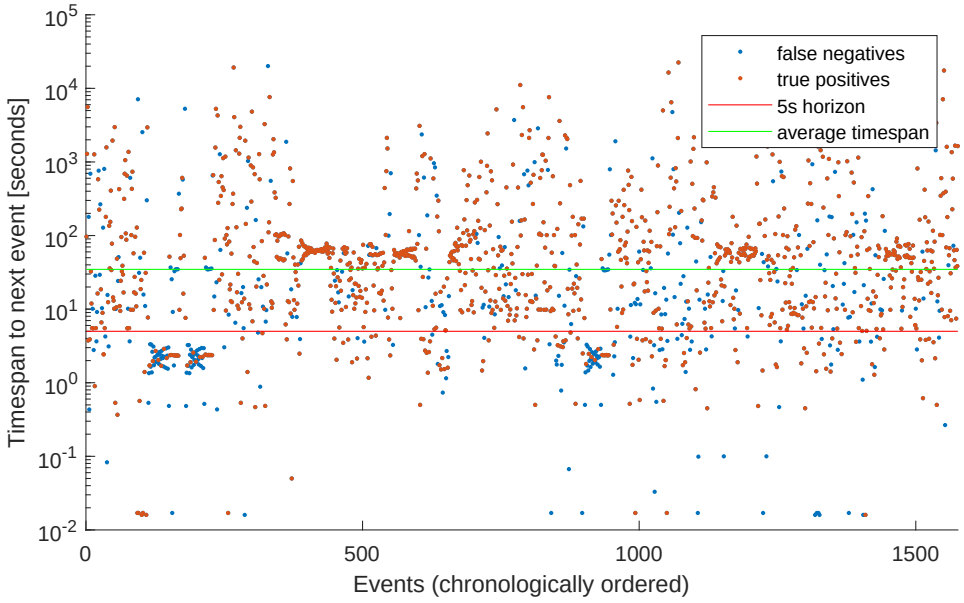


Figure 5.3.1: A BLUED scatter plot of chronologically ordered events with their distance to the next event in the position of the y-axis. The three star-shaped clusters below the 5s horizon are caused by the printer appliance. The detection rate drops below the 5 s horizon, causing more false negatives, due to the fact that 2 events in between 5 s are recognized as one event.

Precision, Recall, and F-Score are the most relevant performance metrics for event detection algorithms. These normed metrics allow a general performance conclusion considering the number of correctly detected (true positives), incorrectly detected (false positives) and not detected (false negatives) events. The F-Score is a metric that rises to 1 by an increase of true positives and decrease of false positives. It is combining both relevant performance metrics (true positives and false positives) and is the preferred

performance metric in the following evaluations.

5.3.1 Features

For our first experiment, we implemented the event detection, using adaptive-training and 87 nearest neighbors for the K-NN classifier. These values seemed promising in pre-executed experiments. The highest performance could be achieved with features that are based on the CUSUM (see Table 5.3.1). The CUSUM has already been used for event detection with promising results by Trung, Dekneuvel, Nicolle, et al. [47]. Since the current and ΔCUSUM segments are similar (see Figure 5.1.4), we expected comparable results. A closer look at the segments reveals that the mean event step of the ΔCUSUM segments is broader and more obvious due to the power neutrality of the CUSUM. We assume that this power neutrality leads to a more distinct event model and an improved detection performance. The performance on BLOND-50 supports these assumptions with a similar trend in the results.

Events that have a previous current of near zero are always ON-events, which are easily detectable in the per-appliance measurements (BLOND-50) but not in the case of concurrent running appliances of aggregated measurements (BLUED). The features *ADM*, *SPF*, and *Current* could therefore not be applied to the BLOND-50 dataset due to their strong dependence on the appliance power in combination with the single appliance measurements which would influence the results in an invalid way.

5.3.2 Normalization

To prevent undesired feature weighting, a feature normalization needs to be applied, especially in the case of a strong range variance of the feature dimensions. There are two common ways to normalize the feature space. The first is the min-max scaling that ensures that all dimensions lie in a range of [-1 ... 1] while the second is called

Table 5.3.1: Feature Results for BLUED and BLOND-50

	BLUED			BLOND-50		
Feature	Prec.	Rec.	F-Sc.	Prec.	Rec.	F-Sc.
$\Delta CUSUM$	0.81	0.75	0.78	0,22	0,98	0,36
CUSUM	0.80	0.75	0.78	0.23	0.98	0.38
Current	0.88	0.38	0.53	-	-	-
ADM	0.88	0.38	0.53	-	-	-
SPF	0.87	0.28	0.43	-	-	-
$\Delta Current$	0.20	0.33	0.25	0.18	0.83	0.29

standardization that ensures that the standard deviation of all dimensions lies at exactly 1.

Table 5.3.2: Normalization Results

	BLUED			BLOND-50		
Norm	Prec.	Rec.	F-Sc.	Prec.	Rec.	F-Sc.
None	0.82	0.74	0.78	0.22	0.98	0.36
MinMax	0.82	0.75	0.78	0.23	0.97	0.37
Variance	0.82	0.72	0.77	0.24	0.96	0.38

The min-max normalization performs best in our experiments on BLUED but also shows that the normalization itself does not influence the performance significantly (see Table 5.3.2). For BLOND-50, the best result could be achieved with a variance normalization. However, also here, the performance results remain quite stable. This means that the different value ranges of the feature space dimensions do not add any significant weighting. This is most likely caused by a similar order of magnitude in the value range across the individual feature space dimensions. The fact that the features are based on time series segments, and therefore share the same value range, affirms the low variations in the performance results.

5.3.3 Training Method

The two previously introduced training methods (classical and adaptive training) are being evaluated. The best result for the multivariate event detection (without adaptive training) allows detection of 1 170 out of 1 577 appliance events from BLUED with 490 false positives and a corresponding F-Score of 0.72. This result was obtained with 30 periods of step-size and the K-NN classifier with K=301.

Table 5.3.3: Adaptive Training Improvement on BLOND-50

	K-NN			SVM		
Training	Prec.	Rec.	F-Sc.	Prec.	Rec.	F-Sc.
classical	0.13	0.99	0.24	0.12	0.99	0.21
adaptive	0.22	0.98	0.36	0.28	0.94	0.43
adaptive 3x	0.45	0.87	0.59	0.55	0.85	0.67
adaptive 5x	0.53	0.85	0.65	0.56	0.77	0.65

All experiments for the adaptive training show a significant, absolute improvement of the event detection performance of +0.14 in average for the F-Score regarding the BLUED dataset (see Figure 5.3.3). The individual improvements vary slightly. The primary performance enhancement of the adaptive training is to reduce the number of false positives due to improvements in the non-event class. The best result for BLUED was obtained with 1 175 true positives and an F-Score of 0.78 by using K=137 for the K-NN classifier, a min-max normalization, and one adaptive training round. The number of false positives was reduced to 260. A significant rise of true positives was not expected and did not occur in most experiments with adaptive training.

The main improvement was observed by applying three rounds of the adaptive training to the event detection on the BLOND-50 dataset. Since the event detection on this dataset produces many false positives, due to a high number of SMPS-driven appliances, the adaptive training reduced the number of false positives from 19 463 to 2 297 which is an improvement of more than eight times. An expected side effect of this enormous improvement is a considerable, but still low, decrease in true positives and recall (see Table 5.3.3).

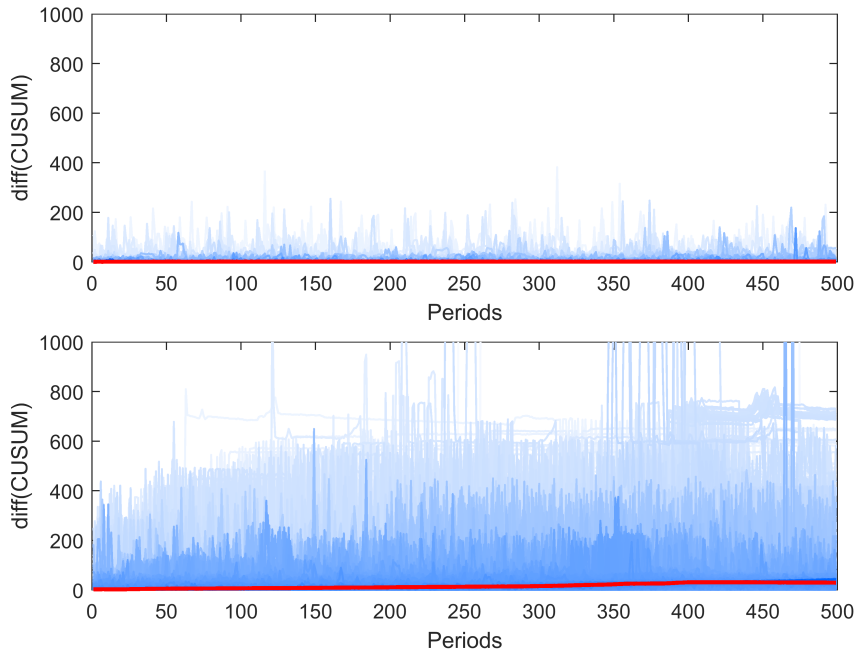


Figure 5.3.2: The first plot shows the non-event class represented only by the implicitly-known non-event segments. The second plot shows the non-event class including the false positives from the adaptive training. The increased diversity due to the false positives is clearly visible. The images are retrieved from non-events of the first 2 weeks in 2016-11 of the BLOND-50 dataset without (first plot) and with (second plot) one adaptive training run.

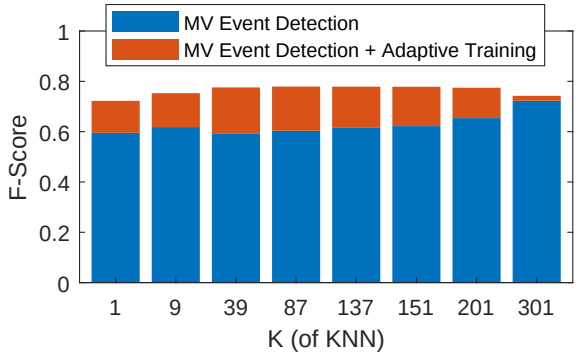


Figure 5.3.3: The individual performance improvements by using the adaptive training. The bars show the achieved event detection F-Score for different K of the K-NN classifier on the BLUED dataset.

Table 5.3.4: Overall best results on BLUED and BLOND-50

	Feature	Norm	Train	Class	Param	F-Sc.
BLUED	ΔCUSUM	MinMax	adap 1x	KNN	K=137	0.78
BLOND-50	CUSUM	Variance	adap 3x	SVM	C/G 128/512	0.67

Using the adaptive training to augment the training set with false positive samples, we were able to reduce the final number of false positives during testing. We conclude that the classifier learns the not explicitly definable heterogeneous model of a non-event by adding the false positives of the training run (see Figure 5.3.2).

5.3.4 Classification

K-NN

Since the event detection performance varies unexpectedly strong, depending on the number of neighbors for the K-NN classifier, we decided to evaluate the performance of eight different K for the classifier. The best general K in our experiments was 301 with classical training, while it was 137 when applying the adaptive training (see Figure 5.3.3). For BLOND-50 the best result with K-NN was achieved by using five rounds of adaptive training (see Figure 5.3.4).

SVM

The best result we could achieve by using the SVM classifier on the BLUED dataset was with an F-Score of 0.72 considerably lower than with 0.78 for the K-NN classifier. The reason is an almost twice the number of false positives - even after adaptive training. The number of true positives with 1 112 lies only slightly below the best result for K-NN. For BLOND-50 the best result by using the SVM lies in a range of 0.67 by using three adaptive training rounds. The optimal SVM hyper-parameter have been retrieved with a grid search algorithm that is provided in the LIBSVM package of C.-C. Chang and C.-J. Lin [77] and could be found at C=128 and Gamma=512 for BLUED and C=1 and Gamma=0.0078 for BLOND-50.

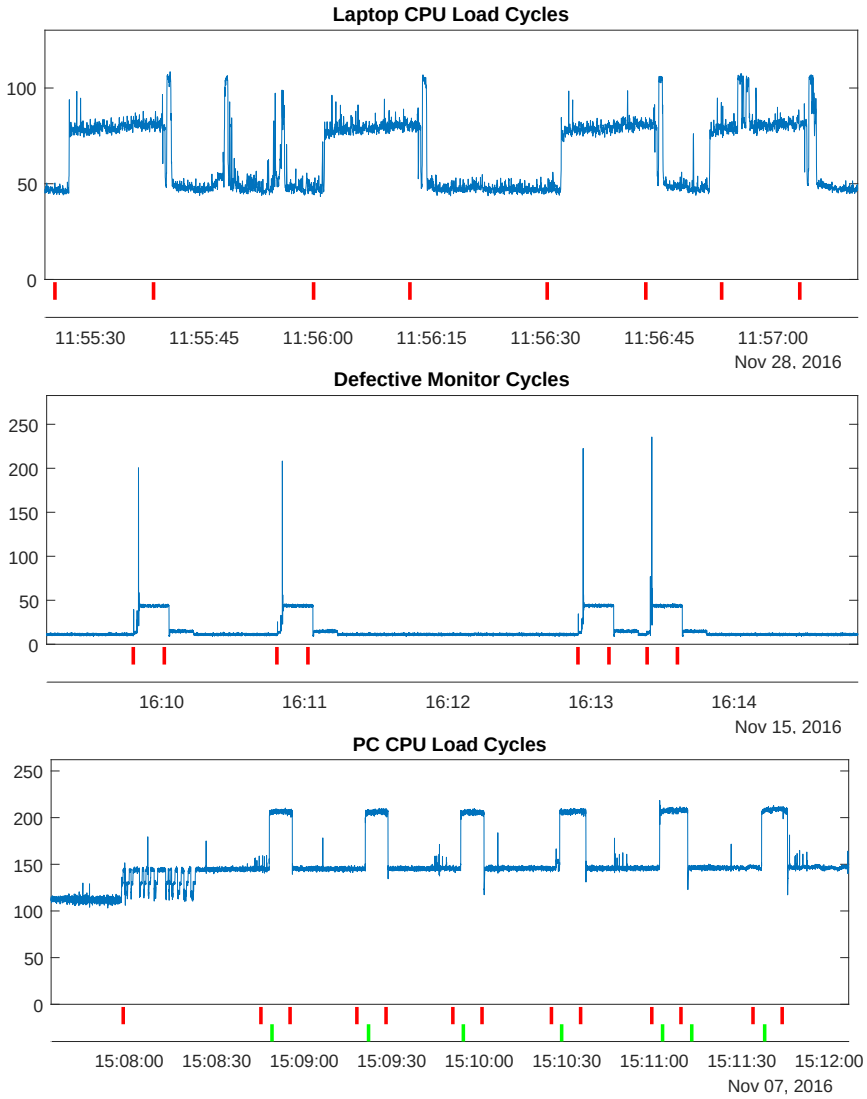


Figure 5.3.4: The three plots show the most prominent reasons for false positive events in BLOND-50: a laptop that produces event-like patterns (first plot), a faulty monitor that immediately goes OFF after switching ON (second plot), and a desktop computer that produces event-like patterns due to CPU load changes. The colored event marker show the false positives that stem from the classical (red) and adaptive (5x) training method (green).

6

Appliance Feature Study

In this chapter, we evaluate a broad set of features for electrical appliance recognition, extracted from high-frequency sampled start-up events. These evaluations were applied on several existing high-frequency sampled energy datasets. To examine clean signatures, we ran all experiments on two datasets that are based on isolated appliance events; more realistic results were retrieved from two real household datasets. Our feature set consists of 36 signatures from related work including novel approaches, and from other research fields. The results of this work include a stand-alone feature ranking, promising feature combinations for appliance recognition in general and per-appliance performances.

6.1 NILM Features

In this work, we considered features that allow us to distinguish appliances based on their start-up events. Many of the discussed features are well documented in the literature and are used frequently by the NILM community. Some features were modified, renamed, or adapted for NILM purposes. All features were extracted from the region of interest I_{ROI}, which is 500ms of the start-up current and voltage.

Only appliance steady-state features satisfy the Feature Addition Criterion *FAC* [57],

necessary for superimposing concurrently running appliances. In this work, all events are handled as isolated events, ignoring the chance of concurrently running appliances in the household datasets which is known as the switch continuity principle [1]. All features extract information from current and voltage signatures. Voltage-only features do not contain any relevant contribution to the current features since appliance-related fluctuations in the voltage signal are dependent on the current.

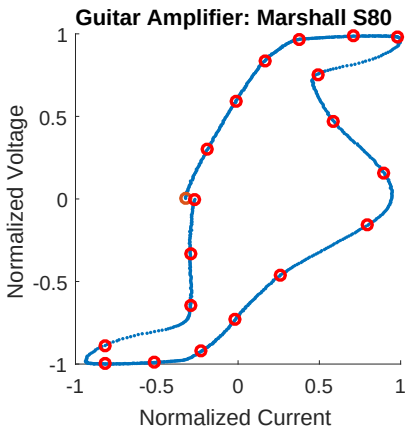


Figure 6.1.1: V-I Trajectory with 20 sampling points from a guitar amplifier.

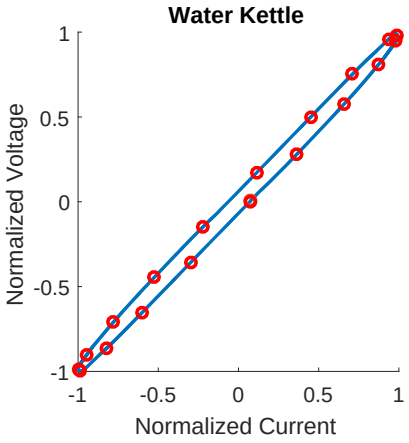


Figure 6.1.2: V-I Trajectory with 20 sampling points from a water kettle.

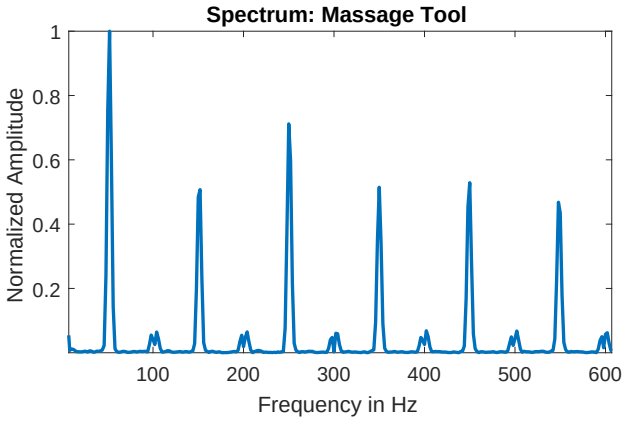


Figure 6.1.3: The spectrum of a motor equipped tool with a strong even odd harmonics imbalance.

6.1.1 Established Features

Early publications on NILM focused mostly on simple-to-compute features. The most used features include electrical power quantities like **Active Power** P, **Reactive Power** Q, and **Apparent Power** $|S|$ [1, 54, 62, 78] which can be considered a standard set for NILM purposes. The **Phase Shift** is the phase angle difference between voltage and current in degrees.

$$P = rms(I) \cdot rms(U) \cdot \cos(\phi) \qquad S = rms(I) \cdot rms(U)$$
$$Q = rms(I) \cdot rms(U) \cdot \sin(\phi) \qquad cos(\phi) = \frac{P}{S}$$

The mutual locus of the normalized instantaneous voltage and current waveforms is called **V-I Trajectory** [26] and is a common feature to visualize signal deformations. Linear (resistive) loads (boiler, heater, kettle, iron, toaster, stove etc.) draw an almost straight line from -1 to 1 due to their waveform equality of current and voltage. The resulting path is sampled by 20 time-equidistant points, creating a vector with 20 coordinates. The belly formed shape of a guitar amplifier and the calculated points are shown in Figure 6.1.1. In contrast, the V-I shape of a linear load, such as that from a kettle can be seen in Figure 6.1.2.

Many works on NILM focus on signatures based on harmonics [3, 36, 74]. Non-sinusoidal currents cause harmonic characteristics, which can be retrieved from a Fast Fourier Transform. Motor equipped appliances such as those depicted in Figure 6.1.3 show usually strong harmonics compared to resistive appliances that show almost none. The relative **Harmonics Energy Distribution** *HED* can be obtained by taking the amplitude of the first 20 harmonics $x_{f_1} \dots x_{f_{20}}$ in a ratio to the mains frequency amplitude x_{f_0}.

$$HED = \frac{1}{x_{f_0}} \cdot \left[x_{f_1}, x_{f_2}, \dots, x_{f_{20}}\right]$$

Wavelet Transformation is a different method for retrieving spectral information. Wavelets can be helpful in handling the uncertainty principle in signal processing. The

output is a frequency dependent size of the time window with an – often wanted – increased time resolution for higher frequency at the expense of a lower frequency resolution. It therefore results in an increased frequency resolution for lower frequencies at the expense of a lower time resolution [79]. The purpose of this feature is to get a non-linear spectral frequency distribution of 50 variable sized frequency bands.

The **Spectral Flatness** *SPF* is a measure for the energy distribution in the frequency spectrum. A theoretical flatness of 1.0 means that all frequencies show the same amplitude, which is defined as white noise. The closer the flatness comes to 0, the stronger are individual frequencies. Linear loads like a toaster show a relative low spectral flatness, whereas show appliances equipped with an SMPS a relative high spectral flatness. The *SPF* is computed by the ratio of the geometric mean to the arithmetic mean of the energy spectrum [76].

$$SPF = \frac{\sqrt[N]{\prod_{f \in f_{bins}} x_f}}{\frac{1}{N} \sum_{f \in f_{bins}} x_f}$$

Many appliances such as motors or SMPS-equipped appliances form strong odd and only modest even harmonics. This imbalance can vary strongly between different appliance types. This makes the **Odd-Even Harmonics Ratio** *OER* [76] feature a useful characteristic for appliance recognition. Figure 6.1.3 shows the harmonics imbalance of a motor-equipped massage appliance.

$$OER = \frac{mean(x_{f_1}, x_{f_3}, \dots, x_{f_{19}})}{mean(x_{f_2}, x_{f_4}, \dots, x_{f_{20}})}$$

A 3-dimensional feature called **Tristimulus** [76] extracts the energy of different harmonic groups. This feature is an audio timbre equivalent to the color attributes in vision [76] and gives a 3-dimensional spectral energy distribution metric. It extracts the intensity for the lower, medium, and higher harmonics, which is different for all appliances.

$$T1 = \frac{x_{f_1}}{\sum_h x_{f_h}} \qquad T2 = \frac{x_{f_2} + x_{f_3} + x_{f_4}}{\sum_h x_{f_h}}$$

$$T3 = \frac{x_{f_5} + x_{f_6} + \cdots + x_{f_{10}}}{\sum_h x_{f_h}}$$

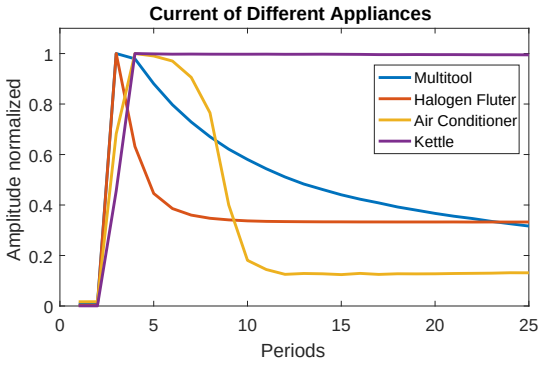


Figure 6.1.4: Different current draws over time on period level. The amplitudes are normalized to 1.

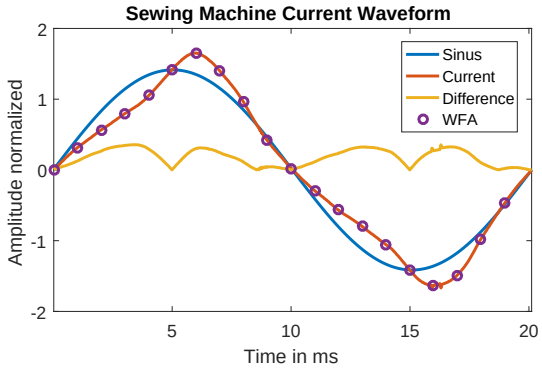


Figure 6.1.5: Current waveform of a motor-equipped sewing machine that deforms the current.

The following scalar features focus on the waveform of a signal: **Form Factor** FF [80], **Crest Factor** CF [81], and logarithmic **Total Harmonic Distortion** THD [82]. These signatures correlate if the waveform is sinusoidal, a compound of strong harmonics or noise. SMPS-equipped appliances produce strong noise in the current waveform which results in a high form factor.

$$FF = \frac{rms(I_{ROI})}{mean(|I_{ROI}|)} \qquad CF = \frac{max(|I_{ROI}|)}{rms(I_{ROI})}$$

$$THD = 10 \cdot log_{10}\left(\frac{1}{x_{f0}} \cdot \sum_{n=1}^{5} x_{fn}\right)$$

Voltage, and therefore also current, can easily drop by a few percent due to rapid changes in the power grid outside the observed circuit. The **Resistance** R is not influenced by voltage fluctuations. Four metrics can be defined using the common Root Mean Square (RMS) and a derived function based on the median (less spike-affected computation). The reciprocal of the resistance is called **Admittance** Y [1] and can be calculated using the quadratic mean and quadratic median accordingly.

$$R_{mean} = \frac{\sqrt{\frac{1}{N} \cdot \sum {U_{ROI}}^2}}{\sqrt{\frac{1}{N} \cdot \sum {I_{ROI}}^2}} \qquad R_{median} = \frac{\sqrt{median\left({U_{ROI}}^2\right)}}{\sqrt{median\left({I_{ROI}}^2\right)}}$$

The Moving Pictures Experts Group (MPEG) published many audio description schemes in the MPEG-7 ISO standard [83]. These descriptors yield to the signal envelope and harmonic characteristics. Both aspects can also be found in energy data. While the signal envelope for musical instruments has an attack, decay, sustain and release state, an electrical appliance has a start-up, decay, steady state and turn off. Musical instrument onsets need to be found in the same way as appliance events. Different instruments draw different harmonic characteristics – similar to electrical appliances. Since high-frequency sampled energy data has similarities to audio data, a subset of these descriptors can be adapted for electricity purposes.

The **LogAttackTime** [76] is an envelope feature describing the logarithmic amount of milliseconds until a sound such as a musical instrument reaches its maximum intensity. For NILM purposes the LogAttackTime depicts the time $ln(t_A)$ until the current reaches its maximum $t_A = max(I_{ROI})$. This feature will often result in a low value, although some appliances, like a power drill, can have an increasing current characteristic due to its speed control.

The **Temporal Centroid** C_t describes the temporal balancing point of the current energy in the region of interest [76]. Appliances with a strong start-up current, such as a vacuum cleaner have a significantly different centroid due to higher consumption in the beginning of the event than appliances with a steady current (toaster).

$$\begin{aligned} I_{W(k)} &= [\text{current sample vector}] \quad ; \; k^{th}\text{period} \\ I_{P(k)} &= rms(I_{W(k)}) \qquad\qquad\qquad\quad ; \; k^{th}\text{period} \\ C_t &= \frac{1}{f_0} \cdot \frac{\sum_{k=1}^{N} I_{P(k)} \cdot k}{\sum_{k=1}^{N} I_{P(k)}} \end{aligned}$$

The **Spectral Centroid** C_f [76] defines the balancing point of the given spectrum. Appliances forming a non-linear current (e.g. SMPS-equipped appliances) have a significant

higher spectral balancing point than linear loads due to the presence of high order frequencies.

$$C_f = \frac{\sum_{f \in f_{bins}} x_f \cdot f}{\sum_{f \in f_{bins}} x_f}$$

The **Harmonic Spectral Centroid** C_h [76] defines the balancing point based on the first 50 harmonics of the mains frequency. This feature covers similar characteristics as the spectral centroid but ignores non-harmonic noise.

$$C_h = \frac{\sum_{k=1}^{50} x_{f_k} \cdot k}{\sum_{k=1}^{50} x_{f_k}}$$

6.1.2 Developed Features

Since some appliances characteristics of the investigated datasets motivate further approaches, we developed features to improve the classification of these appliances. These features are retrieved from individual appliance observations, modifications of known features or ideas from existing literature.

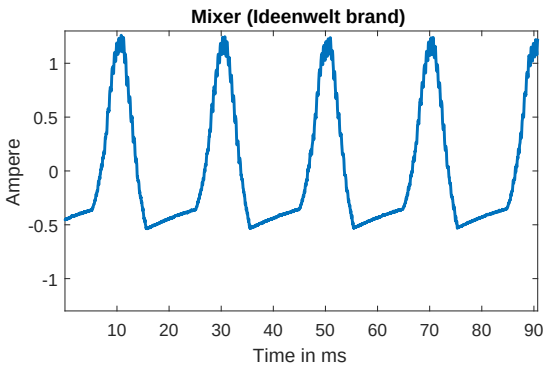


Figure 6.1.6: Current of a mixer with cut negative half waves.

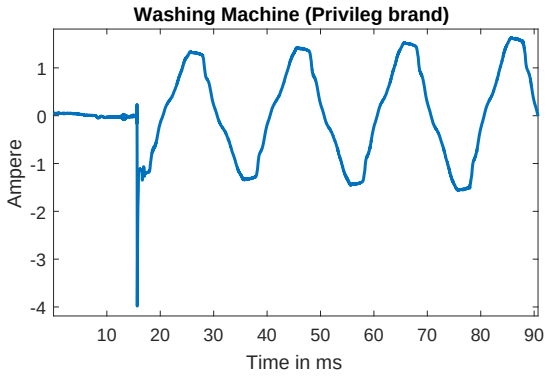


Figure 6.1.7: Current of a washing machine with a strong spike at the start-up.

The feature **Signal to Signal Mean Ratio** *SSMR* extracts information from the spectrum of the current signal and puts the strongest frequency amplitude into a ratio with the

spectral mean. The *SSMR* is a variant of the Signal Noise Ratio *SNR*. Pure resistive appliances such as heaters, water kettles, and ovens are known to have very weak harmonics compared to non-pure resistive loads [78]. Strong harmonics would result in a higher spectral mean. Thus, the *SSMR* can be seen as a scalar representation of the dominance of the strongest frequency in the current spectrum:

$$SSMR = \frac{max(spec)}{mean(spec)}$$

where *spec* is the absolute left-sided frequency spectrum of the whole current sample I_{ROI}.

Appliances that use a SMPS or incandescent light bulbs can have a strong and short current spike in the first period after their start-up (Figure 6.1.7). The amplitude of the first period differs in these cases significantly to the periods of the steady state. The feature **Inrush Current Ratio** *ICR* retrieves the RMS of the first period $I_{W(1)}$, and the last period $I_{W(N)}$ of the current region of interest I_{ROI}:

$$ICR = \frac{I_{P(1)}}{I_{P(N)}}$$

Some household appliances show different characteristics in the positive and negative half cycle of their current, which can be seen in the current waveform of a mixer (Figure 6.1.6). This behavior is usually caused by dimmers or motor speed controllers, which are widely used to reduce the voltage and therefore also the current. Some of these circuits affect only one half of the current cycle. This can be captured by comparing the RMS of 10 averaged positive and negative current half cycles in the **Positive-Negative Half Cycle Ratio** *PNR*. The mixer in Figure 6.1.6 has an imbalance ratio of around 0.6 while most other balanced appliance have a ratio close to 1.0

$$PNR = \begin{cases} \frac{I_{P_{pos}}}{I_{P_{neg}}} & \text{if } I_{P_{neg}} \geq I_{P_{pos}} \\ \frac{I_{P_{neg}}}{I_{P_{pos}}} & \text{if } I_{P_{neg}} < I_{P_{pos}} \end{cases}$$

The **Max-Min Ratio** *MAMI* is an alternative way to cover one-sided waveform characteristics. It puts the maximum and absolute minimum peak current in ratio. In comparison to *PNR*, this feature focuses on the peak values of each half wave that could cover one-sided spikes. The mixer of Figure 6.1.6 shows with around 0.3 a significant lower Max-Min Ratio than a theoretical ideal balanced appliance with a ratio of 1.0.

$$MAMI = \begin{cases} \frac{|min(I_{ROI})|}{|max(I_{ROI})|} & \text{if } |max(I_{ROI})| \geq |min(I_{ROI})| \\ \frac{|max(I_{ROI})|}{|min(I_{ROI})|} & \text{if } |max(I_{ROI})| < |min(I_{ROI})| \end{cases}$$

To determine if an appliance has a pure sine current or spikes from switching-artifacts like polling, the absolute maximum peak can be put into a ratio with the absolute mean current to result in a **Peak-Mean Ratio** *PMR*. Again, linear loads show equal values whereas appliances with strong start-up currents, like an incandescent light bulb, are recognizable with this feature.

$$PMR = \frac{max(|I_{ROI}|)}{mean(|I_{ROI}|)}$$

Appliances equipped with a SMPS or compact fluorescent lights (CFL) show very short (sub-period) peaks and spikes, which won't be covered when focusing only on the power of the whole period. Therefore the RMS current of this period is put into ratio to the maximum peak in the first period. This yields the **Max Inrush Ratio** *MIR* feature, a normalized indication of the peak steepness. A linear load has a theoretical *MIR* of $\frac{1}{\sqrt{2}} \approx 0.707$ while a CFL, for example, can have an *MIR* of around 0.3 due to its non-sinusoidal waveform.

$$MIR = \frac{I_{P(1)}}{max(|I_{W(1)}|)}$$

An indicator of the current steadiness can be retrieved by putting the variance and mean of the absolute current into a ratio. A short and high current spike increases the absolute variance to a higher degree than the absolute mean. Our experiments show that linear

loads have a relatively low **Mean-Variance Ratio** *MVR* and appliances with a short peak (e.g. light bulb) a relatively high *MVR*.

$$MVR = \frac{mean(|I_{ROI}|)}{var(|I_{ROI}|)}$$

The mains voltage follows a relative sinusoidal waveform. Various non-linear appliances distort the current signal by adding individual non-sinusoidal characteristics. Figure 6.1.5 shows the waveform of a sewing machine compared to a generated sine wave. The goal of the **Waveform Distortion** *WFD* feature is to obtain a distortion metric of the current waveform compared to a single period of a sine wave with the same energy Y_{sin}. Therefore, the first 10 post-start-up periods are averaged, normalized (with the RMS of itself), and aligned to the rising zero crossing. The absolute current wave is subtracted from the equivalent absolute sine wave and the differences are summed up. The smaller the value, the more sinusoidal and similar are the half waves of the current waveform.

$$\begin{aligned} I_{WM} &= mean(I_{W(1)}, \dots, I_{W(10)}) \\ WFD &= sum(|Y_{sin}| - |I_{WM}|) \end{aligned}$$

A higher degree of information can be extracted from the waveform itself (e.g. square, saw-tooth, single-pulse). By taking the mean of the first 10 periods point-by-point and down-sampling this vector to 20 points as in Figure 6.1.5, we get the multi-dimensional **Waveform Approximation** *WFA* feature.

$$WFA = downsample(I_{WM}, \left\lfloor \frac{\text{length}(I_{WM})}{20} \right\rfloor)$$

Many appliances, including vacuum cleaners, light bulbs, and motor-based devices, do not have a steady but rather a decreasing power consumption over the initial start-up phase. Due to these variations in the period-by-period current, combining the RMS currents of each period from I_{ROI} into a multi-dimensional vector gives us a **Current Over Time** *COT* distribution. The **Admittance Over Time** *AOT* is less influenced by

voltage fluctuation and is calculated by dividing current $I_{P(k)}$ and voltage $U_{P(k)}$. Figure 6.1.4 shows the current over time of four appliances; the maximum current is normalized to 1.0.

$$COT = \left[I_{P(1)}, I_{P(2)}, \dots, I_{P(25)}\right]$$
$$AOT = \left[\frac{I_{P(1)}}{U_{P(1)}}, \frac{I_{P(2)}}{U_{P(2)}}, \dots, \frac{I_{P(25)}}{U_{P(25)}}\right]$$

The inrush current is higher than the steady state current in many appliances. This inrush current significantly differs in time throughout various appliance types. The index of the first period, after the initial current increase levels off, can be used as the **Periods to Steady State Current** *PSS* feature. The steady state is reached when $I_{P(k)}$ falls below a pre-calculated limit above the median of *COT*.

$$L = \frac{1}{8} \cdot (max(COT) - median(COT)) + median(COT)$$
$$PSS = k \quad ; \text{ first period, where: } I_{P(k)} < L$$

To obtain the advantages of higher sampling frequencies, one can shift the main focus in the spectrum by applying a high-pass filter with a pass frequency of 5 kHz resulting in the following new features: **High Frequency Spectral Centroid** *HFSPC* and **High Frequency Spectral Flatness** *HFSPF*.

Using the high-pass-filtered spectrum one can use the **High Frequency Spectral Mean** *HFSPM* as an indication of the appliance impact on high-frequency regions.

$$HFSPM = mean(x_f \text{ for } f \geq 5\,\text{kHz}, f \in f_{bins})$$

6.2 Experimental Methodology

We set up a machine learning test bed, to evaluate the classification performance of the above defined features. We used start-up events of PLAID [22], WHITED [23], UK-DALE [21], and BLUED [20] to evaluate various features and their combinations with different classifiers.

PLAID and WHITED focus on isolated single appliance measurements. From WHITED, a typical household subset is used (one appliance model per type, low inner-class diversity, 280 events of 27 appliance types). PLAID provides multiple appliance models per type (e.g. 7 heaters, 33 laptops, etc.). All appliances and events are used in our experiments (high inner-class diversity with 1074 events of 11 appliance types).

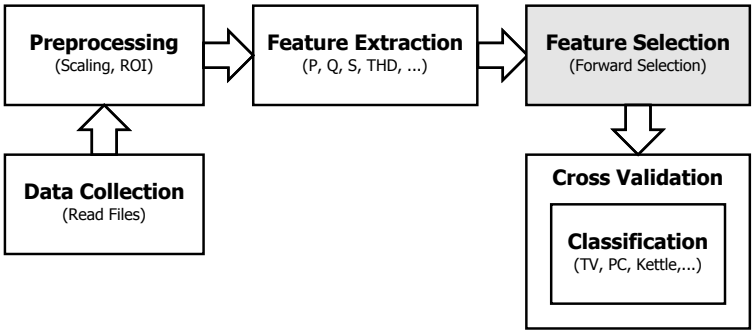


Figure 6.2.1: The simplified steps of our evaluation system based on pattern recognition and cross-validation.

UK-DALE and BLUED provide real household measurements. Start-up events were extracted based on timestamps and event labels from the provided low frequency single appliance measurements. We defined an event as a spontaneous, single and significant rise in the current consumption for a couple of mains periods.

Our machine learning experiments follow a typical pattern recognition approach. Figure 6.2.1 shows the appliance classification pipeline with supervised learning embedded in the evaluation system. We define the classification problem as follows: Classify the appliance type based on a start-up event and represent each appliance type as an individual class.

For reliable results we used a stratified 5-fold cross-validation with four classifiers:

1-Nearest Neighbour (KNN), *Binary Decision Tree* (BDT), *Linear Discriminant Analysis* (LDA), and *Support Vector Machines* (SVM). All computations are performed with a combination of Matlab (with LIBSVM [77] and various toolboxes), and Python (with common scientific computing packages). To remove unwanted feature weighting caused by different ranges of values, we implemented a feature variance normalization, based on the training data in each cross-validation step. Since F1 Score is the most common used metric, all features and feature combinations are ranked based on the F1 Score; the higher the score, the better the classification performance.

Feature	Time Window	Domain	Signal	Source	FAC	Dim	F-Score WHITED	F-Score PLAID	F-Score BLUED	F-Score UK-DALE	ø
Active Power	whole event	time	I, U	electronics	✓	1	0.71	0.49	0.32	0.24	0.44
Admittance	whole event	time	I, U	electronics	✓	1	0.75	0.52	**0.34**	**0.26**	0.47
Admittance (median)	whole event	time	I, U	electronics	✓	1	0.83	0.58	0.28	0.18	0.47
Apparent Power	whole event	time	I, U	electronics	✓	1	0.77	0.52	0.32	0.26	0.47
Crest Factor	whole event	time	I	electronics	✓	1	0.33	0.33	0.23	0.21	0.28
Even-Odd Harmonics Ratio	whole event	spectral	I	audio	✓	1	0.32	0.24	0.17	0.17	0.22
Form Factor	whole event	time	I	electronics	✓	1	0.62	0.45	0.15	0.14	0.34
Harmonic Spectral Centroid	whole event	spectral	I	MPEG 7	✓	1	0.54	0.32	0.16	0.09	0.28
Inrush Current Ratio	first, last period	time	I	developed		1	0.45	0.28	0.27	0.17	0.29
Log Attack Time	first 25 periods	time	I	MPEG 7		1	0.26	0.35	0.16	0.19	0.24
Max Inrush Ratio	first period, max sample	time	I	developed		1	0.27	0.29	0.15	0.07	0.20
Mean-Variance Ratio	whole event	time	I	developed	✓	1	0.28	0.30	0.18	0.14	0.23
Min-Max Ratio	min, max sample	time	I	developed	✓	1	0.17	0.28	0.20	0.19	0.21
Peak-Mean Ratio	whole event	time	I	developed		1	0.32	0.33	0.25	0.20	0.28
Periods to Steady State Current	first 25 periods	time	I	MPEG 7		1	0.18	0.22	0.11	0.15	0.17
Positive-Negative Half Cycle Ratio	mean first 10 periods	time	I	developed	✓	1	0.11	0.15	0.11	0.07	0.11
Phase Shift	whole event	spectral	I, U	electronics	✓	1	0.83	**0.64**	0.27	0.22	**0.49**
Reactive Power	whole event	time	I, U	electronics	✓	1	0.83	0.54	0.26	0.22	0.46
Resistance	whole event	time	I, U	electronics	✓	1	0.77	0.54	**0.34**	**0.26**	0.48
Resistance (median)	whole event	time	I, U	electronics	✓	1	**0.84**	0.57	0.30	0.17	0.47
Signal-Signal Mean Ratio	whole event	spectral	I	developed	✓	1	0.66	0.39	0.16	0.15	0.34
Spectral Centroid	whole event	spectral	I	MPEG 7	✓	1	0.75	0.42	0.15	0.09	0.35
Spectral Centroid HF	whole event	spectral	I	developed	✓	1	0.38	0.23	0.07	0.07	0.19
Spectral Flatness	whole event	spectral	I, U	audio	✓	1	0.68	0.37	0.16	0.08	0.32
Spectral Flatness HF	whole event	spectral	I, U	developed	✓	1	0.27	0.25	0.09	0.07	0.17
Spectral Mean HF	whole event	spectral	I, U	developed	✓	1	0.73	0.41	0.16	0.10	0.35
Temporal Centroid	first 25 periods	time	I	MPEG 7		1	0.50	0.38	0.22	0.17	0.32
Total Harmonic Distortion	whole event	spectral	I	audio	✓	1	0.73	0.50	0.27	0.21	0.43
Waveform Distortion	mean first 10 periods	time	I	developed	✓	1	0.46	0.40	0.13	0.09	0.27
Tristimulus	whole event	spectral	I	audio	✓	3	0.84	0.47	0.28	0.25	0.46
Harmonics	whole event	spectral	I	audio	✓	20	**0.99**	0.90	0.54	0.60	0.76
Waveform Approximation	mean first 10 periods	time	I	developed	✓	20	0.85	0.82	0.49	0.49	0.66
Admittance Over Time	first 25 periods	time	I, U	developed		25	0.98	0.82	**0.61**	**0.76**	0.79
Current Over Time	first 25 periods	time	I	developed		25	0.97	0.87	0.60	0.75	**0.80**
VI-Trajectory	mean first 10 periods	time	I, U	electronics	✓	40	0.93	0.88	0.39	0.45	0.66
Wavelet Analysis	whole event	spectral	I	audio	✓	50	**0.99**	**0.91**	0.56	0.67	0.78

Table 6.2.1: Feature Overview: Features are extracted from different temporal regions of an event, which can be found in column *Time Window*. *Source* describes roughly the usual field of research or standard from which the feature originates. The value *developed* represents features that are newly developed or adapted from existing features in this work. The features with the highest F1 Score – scalar and multi-dimensional – are in bold and the overall winners are highlighted in gray.

6.3 Experimental Results

All experiments for isolated measurements were performed on PLAID and the WHITED subset. The selected appliances can be seen in Table 6.3.6. To cover real-world scenarios, we applied our experiments on two high-frequency sampled household datasets: BLUED and UK-DALE. With these datasets we rank all stand-alone features, find the best 2-dimensional feature combination, and compute the best forward-selected combination. Therefore, we try to find the best-performing features for the following questions:

What is the highest achievable classification performance …

- …for data with a high inner-class diversity? (PLAID)
- …for data with a low inner-class diversity? (WHITED)
- …for real household measurements? (BLUED, UK-DALE)

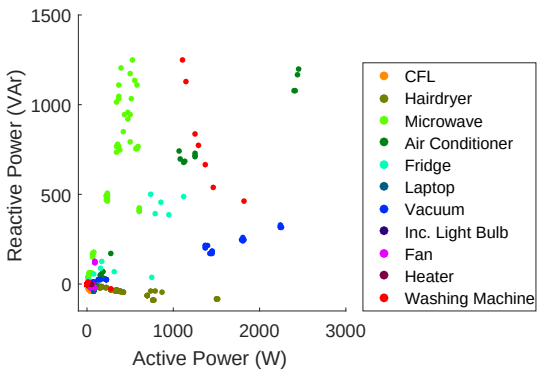


Figure 6.3.1: PLAID in P-Q plane.

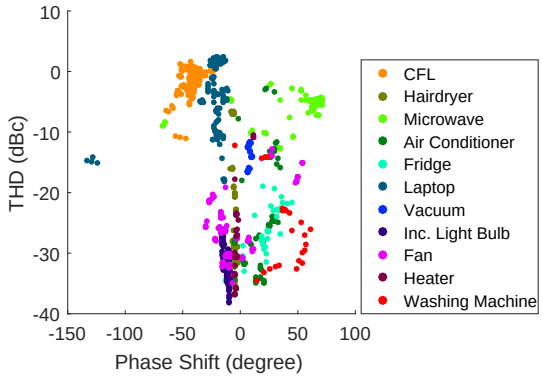


Figure 6.3.2: PLAID in PS-THD plane.

6.3.1 Stand-Alone Feature Ranking

The goal of this ranking is to compute the individual classification performance of each feature and rank them in a table. To be able to compare all features against one another,

it is necessary to compute the stand-alone classification performance of each feature by itself.

The evaluation was performed over all datasets with a *1-Nearest Neighbor* classifier for comparability, traceability, and reproducibility (Table 6.2.1). It is important to differentiate between one- and multi-dimensional features, since multi-dimensional features have a significant higher potential in classification performance than scalar features.

For isolated measurements (PLAID and WHITED), the *Phase Shift* with an average F1 Score of 0.74 has the highest discriminating quality. This means that the most relevant one dimensional metric to recognize appliances lies in the individual voltage and current phase difference of each appliance. The best multidimensional feature is the *Wavelet Analysis* with an average F1 Score of 0.95. This on the other hand, means that, due to individual nonlinearities inside the appliances, complex spectral characteristics offer the highest discriminating quality for isolated events.

For real household measurements (BLUED and UK-DALE), the best scalar feature are the *Admittance* and the *Resistance* with an average F1 Score of 0.30. Note that classification based on a single scalar feature is difficult for any classifier. Since the household data contains more noise than isolated measurements, the recognition quality is significantly lower. The best multidimensional feature is *Admittance Over Time* with an average F1 Score of 0.69. For aggregated measurements, the individual temporal appliance energy consumption shows the highest robustness against unwanted interferences.

When ranking the features of all datasets and environments, the best scalar feature is *Phase Shift*, and the best multi-dimensional feature is *Current Over Time*.

6.3.2 2-Dimensional Feature Combination

Combining multiple features usually improves the classification performance. However, each additional feature increases computational complexity. Therefore, this experiment focuses on evaluating 2-dimensional feature combinations for each dataset and classifier, while keeping the complexity to a minimum. One can use a 2-dimensional feature space

to visualize and examine the class boarders (Figure 6.3.1). A common visualization is the *P-Q* plane [1, 62, 78] where *Active Power* and *Reactive Power* form a 2 dimensional scatter plot.

We evaluated all possible combinations of scalar features, resulting in 406 possible pairs. The *PF-THD* plane shows an interesting sample distribution in terms of appliance clusters (Figure 6.3.2). The advantage of this plane is the high intra-class variance of low-power appliances, compared to the closely grouped clusters in the *P-Q* plane (Figure 6.3.1).

	Features	F1	Pr	Re	Ac
KNN	**Phase Shift** **Admittance**	**0.98**	**0.98**	**0.98**	**0.99**
LDA	Reactive Power Signal-Signal Mean Ratio	0.92	0.94	0.93	0.99
SVM	Total Harmonic Distortion Harmonic Spectral Centroid	0.98	0.98	0.98	0.99
BDT	Spectral Mean (HF) Temporal Centroid	0.97	0.97	0.97	0.99

Table 6.3.1: The best 2-dimensional feature combination for each classifier with WHITED: *Phase Shift* and *Admittance* show promising results. Spectral indicators reach similar results.

	Features	F1	Pr	Re	Ac
KNN	**Active Power** **Reactive Power**	**0.89**	**0.91**	**0.88**	**0.99**
LDA	Phase Shift Temporal Centroid	0.54	0.55	0.58	0.95
SVM	Phase Shift Total Harmonic Distortion	0.86	0.90	0.84	0.98
BDT	Phase Shift Total Harmonic Distortion	0.82	0.84	0.80	0.98

Table 6.3.2: The best 2-dimensional feature combination for each classifier with PLAID: *Active Power* and *Reactive Power*

The best 2-dimensional feature combination for each classifier can be seen in Tables 6.3.1, 6.3.2, 6.3.3, and 6.3.4. For the WHITED subset, the best 2-dimensional feature combination (using KNN) is *Phase Shift* and *Admittance*, compared to *Active Power* and *Reactive Power* on PLAID. For BLUED, the best 2-dimensional feature combination (using KNN)

	Features	F1	Pr	Re	Ac
KNN	**Apparent Power Phase Shift**	**0.59**	0.60	**0.59**	**0.97**
LDA	Active Power Signal-Signal Mean Ratio	0.36	0.36	0.38	0.96
SVM	Resistance (med) Temporal Centroid	0.57	**0.67**	0.57	0.97
BDT	Active Power Temporal Centroid	0.57	0.59	0.57	0.97

Table 6.3.3: The best 2-dimensional feature combination for each classifier with BLUED: The performance metrics show that the *Phase Shift* performs well, even for noisy household data.

	Features	F1	Pr	Re	Ac
KNN	Inrush-Steady State Ratio Admittance	0.59	0.59	0.59	0.99
LDA	Reactive Power Total Harmonic Distortion	0.31	0.32	0.35	0.98
SVM	Active Power Inrush-Steady State Ratio	0.56	0.61	0.56	0.99
BDT	**Active Power Inrush-Steady State Ratio**	**0.60**	**0.61**	**0.60**	**0.99**

Table 6.3.4: The best 2-dimensional feature combination for each classifier with UK-DALE: In this case, the Binary Decision BDT reaches the highest classification performance with *Active Power* and *Inrush-Steady State Ratio*.

is *Apparent Power* and *Phase Shift*, compared to *Active Power* and *Inrush-Steady State Ratio* on UK-DALE (using BDT). These results suggest that consistent high classification performance can be achieved, by including one of the power indicators (*Active*, *Reactive*, *Apparent Power*, etc.) in the feature combination.

6.3.3 Feature Forward Selection

As shown in Table 6.2.1, many features reach an already high stand-alone F1 Score for the appliance classification. The combination of several features usually improves the classification performance up to a limit. Computation time and classification aggravations anomalies due to large feature spaces like the Hughes phenomenon [84] motivate to keep the amount of features to a minimum. We chose a forward selection algorithm (Figure 6.3.3) to find the best compromise between classification performance and a smaller number of features. An evaluation that considers all possible 2^{36} combinations is not feasible in terms of computational effort and resources. The algorithm starts by selecting the best stand-alone feature, then computes all possible 2-pairs and selects the best one as starting point for the next iteration. This is repeated until the classification performance stops improving.

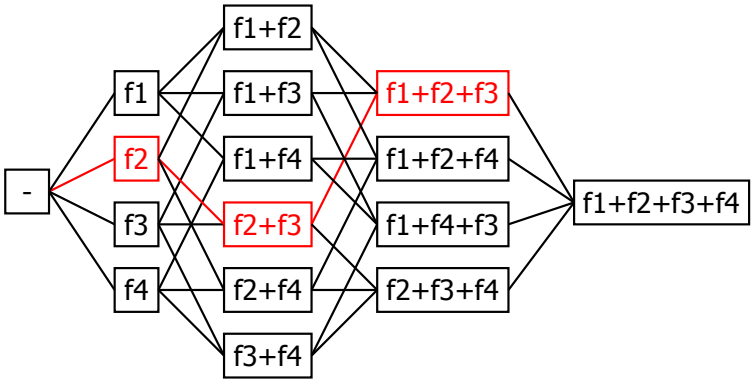


Figure 6.3.3: Feature forward selection: In this example, *f2* is selected as winner of the first iteration, [*f2*, *f3*] in the second, and [*f1*, *f2*, *f3*] as final winner, since the combination of all four features does not further improve the performance.

Table 6.3.5 shows the resulting feature combinations for all datasets computed by the feature forward selection algorithm. Since we used a randomized stratified 5-fold cross validation, the performance results vary slightly for each run. Stratified means in this

context that the amount of training samples are balanced for each class. We achieved an F1 Score of 1.0, due to the intrinsic properties of WHITED (isolated measurements). PLAID showed comparably high classification performance, however, with a different set of features. BLUED and UK-DALE scored slightly below 0.80 due to noise and concurrently running appliances.

Dataset	Features	F1	Pr	Re	Ac
WHITED	Wavelet Analysis, Spectral Mean (HF), VI-Trajectory, Current Over Time	1.00	1.00	1.00	1.00
PLAID	Wavelet Analysis, VI-Trajectory, Admittance Over Time, Form Factor, Phase Shift, Log Attack Time, Current Over Time, Max-Inrush Ratio, Resistance (med)	0.96	0.97	0.95	0.99
BLUED	Admittance Over Time, Resistance (med), Total Harmonic Distortion, Spectral Mean (HF), Temporal Centroid, Phase Shift, Admittance, Admittance (med), Log Attack Time, Spectral Flatness, Sinus Difference Sum, Harmonic Spectral Centroid	0.76	0.78	0.75	0.98
UK-DALE	Admittance Over Time, Resistance, Phase Shift, Temporal Centroid, Admittance, Active Power, Even-Odd Harmonics Ratio, Spectral Flatness	0.79	0.80	0.79	0.99

Table 6.3.5: The results of the forward selection for each dataset. Even with only four features, the optimal classification performance for WHITED is achievable. The results show that not all features are necessary to get the optimal classification performance.

6.3.4 Individual Appliance Performance

Table 6.3.6 shows that the classification performance is not identical for each appliance type. In the case of WHITED, there are no misclassifications since the dataset consists of isolated measurements and only one model per appliance type, which lowers the inner-class variance and therefore improves the classification performance significantly.

Aside from some outliers, most appliance types show a relatively high classification performance. The appliance type *Lights* shows a notably lower classification performance in both datasets. We believe that the main reason for this result is a too broad definition

of lights, which may include a mix of different lighting types (e.g. incandescent, LED, and CFL) and low power lights. A similar problem applies to *Desktop PC's*, which are usually equipped with a SMPS. The total classification fail of the boiler cannot be explained completely. One reason can be that resistive loads have almost no recognizable current signatures that can easily cause a misclassification to another resistive load of similar consumption.

6.3.5 Discussion

What is the best general feature combination for appliance recognition? First of all, there is no one-size-fits-all set of features. The individual composition of appliances in the environment determines the set of features. However, the best feature sets always contain information about a spectral energy distribution (e.g. *Wavelet Analysis*, *Harmonics*, etc.) and time-related power information (e.g. *Current and Admittance Over Time*). The results show that the main information to discriminate between appliances lies in the spectral distribution and the unsteadiness in the power consumption over time.

A poor result of a feature in Table 6.2.1 does not mean that it is worthless. Some features are based on characteristics that are formed only by a rare amount of appliances like the feature *PNR*. The waveform imbalance covered by this feature was only observed in the *Ideenwelt*-mixer of WHITED. The feature may have a poor general discrimination quality but might contribute in a feature combination to distinguish the mixer from a similar appliance such as a multitool.

When focusing on the algorithms, the 1-nearest neighbor classifier leads in performance quality. It seems to be the optimal classifier for this task due to its simplicity and computation performance for small to medium feature spaces. Since the SVM classifier has a strong need for an intense parameter-search, a rough parameter search was used for the 2-dimensional feature combination due to computing time.

Appliance	WHITED KNN	PLAID KNN	BLUED SVM	UK-DALE SVM	Ø
Air Conditioner	1.00	0.93			0.97
Air Compressor			1.00		1.00
Boiler				0.00	0.00
Breadmaker				0.97	0.97
Charger	1.00				1.00
Coffee Machine	1.00			0.88	0.94
CFL	1.00	1.00			1.00
Desktop PC	1.00			[1]0.63	0.82
Dishwasher				0.94	0.94
Drilling Machine	1.00				1.00
Fan	1.00	0.97			0.98
Fridge	1.00	0.75	0.95	0.99	0.92
Game Console	1.00				1.00
Garage Door			0.82		0.82
Hair Dryer	1.00	0.98		0.83	0.94
Heater		0.97			0.97
HiFi	1.00		0.86		0.93
Inc. Light Bulb	1.00	0.97			0.99
Iron	1.00		1.00	0.96	0.99
Juice Maker	1.00				1.00
Kettle	1.00			0.97	0.98
Kitchen Hood	1.00				1.00
Laptop	1.00	0.99			1.00
Lights			[2]0.60	0.04	0.32
Microwave	1.00	0.99		0.95	0.98
Mixer	1.00				1.00
Monitor			0.74		0.74
Printer	1.00		1.00		1.00
Rice Cooker	1.00				1.00
Sandwich Maker	1.00				1.00
Straighteners	1.00			0.86	0.93
Toaster	1.00			0.92	0.96
TV	1.00		0.85	0.83	0.90
Vacuum Cleaner	1.00	1.00		0.97	0.99
Washing Machine	1.00	0.94		0.93	0.96

[1] Mean of HTPC and Office PC [2] Mean of all eight lights

Table 6.3.6: F1 Score of each individual appliance in its corresponding dataset. The values are based on forward selected feature combinations and the best-performing classifier.

7

Deep vs. Machine Learning in NILM

Deep neural networks define the state-of-the-art in several disciplines of learning from large datasets. In this paper, we show how non-intrusive appliance load monitoring (NIALM) can benefit from deep learning. On the basis of an event-based appliance recognition approach, we evaluate seven different classification models: a pattern-recognition approach that is based on a comprehensive hand-crafted feature extraction, three different deep neural network architectures for automated feature extraction on raw waveform data, as well as three simple baseline approaches. The two large-scale, high-frequency sampled energy consumption datasets UK-DALE and BLOND-50 allow us to evaluate the algorithms on more than 50.000 events of 44 appliances. Our study concludes that we are able to reach and surpass performances of state-of-the-art approaches for appliance recognition with an F-Score of 0.75 for UK-DALE and 0.86 for BLOND-50.

7.1 Appliance Recognition Process

We implemented two different appliance recognition systems, a classical machine learning, and a representation learning approach. The typical architectures of these learning systems can be seen in Figure 7.1.1 and 7.1.2. We chose two publicly available energy consumption datasets of a residential and office environment. The datasets are the most

suitable selection of the publicly available datasets for our experiments on the selected deep learning algorithms due to their considerably different set of appliances and usage patterns. Since we use existing datasets, data acquisition does not play any role in this work. Further acquisition details regarding the datasets can be found in the work of Kelly and Knottenbelt [21] (UK-DALE) and Kriechbaumer and H.-A. Jacobsen [24] (BLOND-50).

UK-DALE

The UK Domestic Appliance-Level Electricity (UK-DALE) dataset consists of more than 4 years of energy consumption measurements for a residential building (house-1) with a high number of appliances of many different types. For our experiments, we considered measurements from 2013-04-22 to 2015-01-05. The dataset comprises low-frequency, non-equidistant sampled smart plug measurements (ø 1/6 Hz) for each observed appliance (per-appliance signals) and high-frequency sampled measurements (16 kHz) from a custom sound card meter at the electric cabinet (aggregated signal). The per-appliance measurements allow a coarse determination of appliance events and power consumption to extract the relevant segments from the aggregated signal.

BLOND-50

The Building-Level Office eNvironment Dataset (BLOND) comprises energy consumption measurements from an office building with a high number of appliances of only a few different types. This appliance and appliance type distribution is the main difference between the datasets, covering a wide spectrum of real environments. The BLOND-50 subset comprises 213 days of recording with 50 kHz sampling frequency for the aggregated signal at the electric cabinet and 90 individually observed sockets for the per-appliance measurements with 6.4 kHz sampling frequency.

Our appliance recognition process uses only the first 500 ms of the appliance startup current and voltage as the baseline for the hand-crafted and automated feature extraction of the considered algorithms, categorizing it as a so-called event-based approach.

Figure 7.1.1: The architecture of a classical machine learning approach always requires domain specific expert knowledge for finding types of features with class-discriminative potential in the hand-crafted feature extraction process step three.

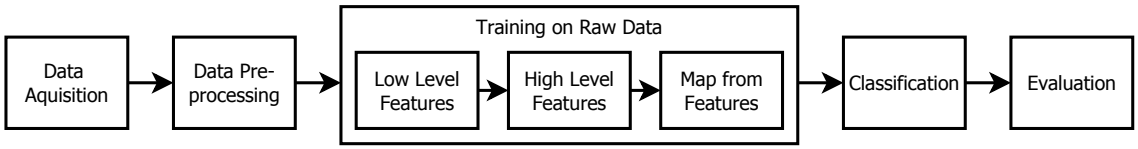


Figure 7.1.2: The architecture of a representation learning approach tries to replace the hand-crafted feature extraction with automated, hierarchical feature extraction without the need of domain specific expert knowledge.

7.1.1 Data Preprocessing and Event Detection

The appliance event time-stamps can be approximated with the help of the additional per-appliance measurements, which are provided in both datasets (1/6 Hz for UK-DALE, 6.4 kHz for BLOND-50). A simple threshold-based event detection algorithm with appliance individual thresholds is used to detect events in these per-appliance measurements (see Table 7.1.1). The resulting appliance event time-stamps are used to extract segments from the high frequency aggregated signal. The power-related switch-on threshold $\Delta_{\uparrow}$ in Watts defines the power state at which an appliance is considered switched on - similarly for the switch-off threshold $\Delta_{\downarrow}$.

The UK-DALE dataset comprises 52 appliances of several types. For this work, we use a subset of 23 appliances by selecting only one appliance per appliance type and ignoring low-power devices such as *ADSL Router*, *Ipad Charger* and *Baby Monitor* that are potentially undetectable in a noisier aggregated signal, due to their low power consumption. The sample numbers of the remaining appliance classes are very heterogeneous and vary between 38 for the coffee machine and 15,766 for the fridge, due to their natural consumer pattern (see Figure 7.1.1).

For the BLOND-50 dataset, a general power threshold of 25 W defines a switch-on and switch-off event. All 21 occurring appliances of the chosen time span are included in the remaining appliance types: *Laptop* (13), *Monitor* (5), *PC* (2) and *Printer* (1), amounting to

Table 7.1.1: UK-DALE Appliance Event Thresholds and Quantity

Appliance	$\Delta_\uparrow$	$\Delta_\downarrow$	# Events	Appliance	$\Delta_\uparrow$	$\Delta_\downarrow$	# Events
Boiler	70	20	1,701	LCD Office	30	4	1,337
Solar Thermal Pump	40	20	5,221	Breadmaker	400	20	649
Laptop	20	2	498	Amp Livingroom	18	10	945
Washing Machine	1,500	1	506	Hoover	400	10	392
Dishwasher	100	20	885	Coffee Machine	1,000	10	38
TV	70	10	907	Hair Dryer	100	20	713
Kitchen Lights	70	20	4,765	Straightener	300	5	264
HTPC	70	20	1,169	Iron	1,000	10	147
Kettle	2,000	10	2,674	Gas Oven	35	10	492
Toaster	1,000	10	1,495	Office Fan	20	2	78
Fridge	70	10	15,766	LED Printer	800	3	159
Microwave	500	10	3,363				

9,321 appliance samples.

Regarding UK-DALE, with the known appliance event time-stamp from the per-appliance measurements, the high-frequency aggregated measurements are observed in a 20 s time-window for the exact event position. For each appliance event that is found in the per-appliance measurements, a segment of the first 500 ms is extracted from the high-frequency aggregated measurements at the corresponding time-stamp. These 500 ms long calibrated startup-transients are the baselines for all following considerations.

It is important to keep in mind that there are two sources of event inaccuracies in this step for UK-DALE. The first lies in the event detection on the low-frequency per-appliance measurements. With 1/6 Hz, the sampling rate is too low to detect short-term consumption patterns. The second lies in finding the exact event position in the 20s time-window in case of multiple occurring events in that time window. The probability to extract the correct appliance segments equals the reciprocal of the number of individual appliance events in the 20s time-window. Reliable error estimation is unfortunately not possible. Since the per-appliance measurements are sampled with 6.4 kHz, the event time-stamps are accurate enough to not cause these issues for BLOND-50.

7.1.2 Hand-Crafted Feature Extraction

For this work, we extracted 36 features that are introduced and explained in our previous work [25]. These features comprise traditional electricity metrics such as *active & reactive power*, *admittance*, *crest factor* and *phase shift*, audio processing features such as *harmonics*, *wavelet analysis* and *total harmonic distortion*, a selection of MPEG7 audio descriptors[1] and novel metrics such as *max-inrush-ratio* and *inrush-current-ratio*. The chosen features are one- and multidimensional with a total number of 212 dimensions. With these features, we reach very high appliance classification performances (F-Score between 0.76 and 1.0) across household-focused subsets of the four publicly available datasets WHITED [23], PLAID [22], UK-DALE [21] and BLUED [20] with the standard classifier (KNN, SVM, LDA, BDT).

7.1.3 Autoencoder

We implemented three different AE architectures that are designed to reduce the raw waveform data to a 212-dimensional feature space. The set comprises a one, two and three-layered encoding and decoding architecture with different dimensionality. The AEs have a mirror-like design, which means that the decoding layers are identical to the encoding layers, but in reverse order.

7.1.4 Convolutional Neural Network Architecture

To ensure an automated hierarchical feature extraction with the 1-dimensional CNN, we applied a sampling and mains frequency (f_s, f_0) dependent layer architecture. The goal is to reduce the layer inputs with max-pooling layers in a way that the output dimension of the last convolutional layer is identical to the number of mains cycles ($n_p = 25$ for 50 Hz mains frequency) of the 500 ms segments. The number of convolutional layers (n_l) and its kernel sizes (k) are calculated as follows:

[1] Since high frequency sampled energy consumption shares similarities to audio and music data, the MPEG7 audio descriptors contain features with significant discriminative potential for electrical appliance

$$\vec{k} = sort_{desc}(prime\_factorization(\frac{f_s}{f_0})), \qquad n_l = \#k$$

Therefore, the pool-sizes of the max-pooling layers for the UK-DALE dataset are [5, 2, 2, 2, 2, 2, 2] (see Figure 7.1.3) while they are [5, 5, 5, 2, 2, 2] for the BLOND-50 dataset.

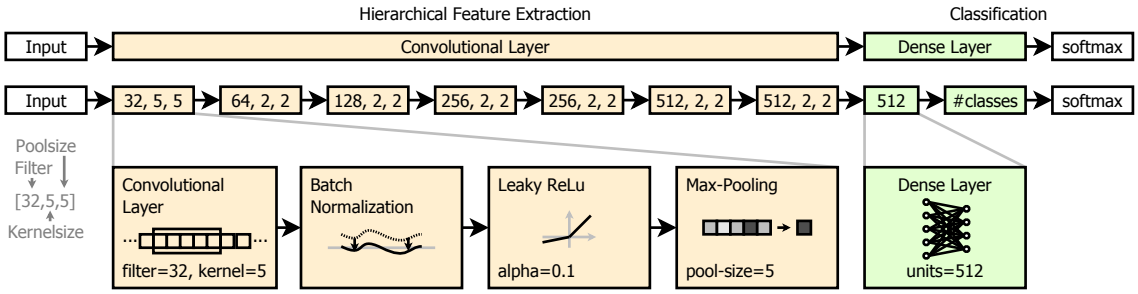


Figure 7.1.3: The architecture of the end-to-end implementation of the CNN for the UK-DALE dataset.

7.1.5 Convolutional Autoencoder

The advantage of CAEs lies in the use of convolutional layers inside an AE network. They benefit from a better locality of natural signals while general AEs handle each input dimension as global. In other words, the order of the input dimensions does not matter for a general, fully connected AE as opposed to using convolutional layers. To keep the model trainable, we implemented three encoding, one coding, and three decoding layers. To ensure comparability with the hand-crafted features extraction approach, we reduced the feature dimensions to 200. This is the closest we could reduce 8,000 and 25,000 to the 212 dimensions of the hand-crafted features approach, using only integer divisors. Therefore, the encoder and decoder pool-sizes (divisors) of the max-pooling layers are [5, 4, 2] and [2, 4, 5] for UK-DALE (see Figure 7.1.4), while they are [5, 5, 5] and [5, 5, 5] for BLOND-50. The number of filter and the kernel sizes of each layer are identical for both datasets, see Figure 7.1.4.

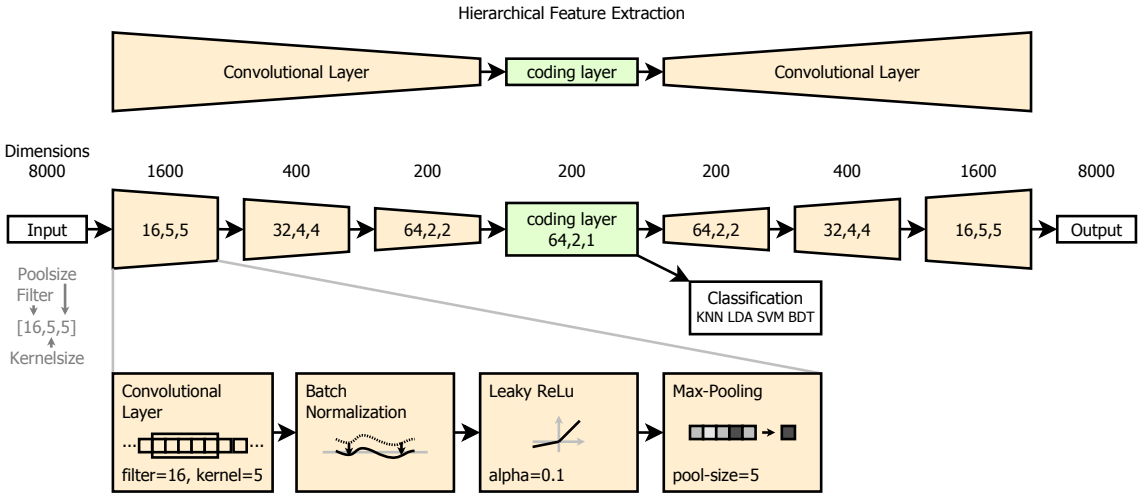


Figure 7.1.4: The architecture of the CAE network for the UK-DALE dataset.

7.1.6 Feature Space Transformation

We implemented two different ways of feature space transformation to avoid undesired feature weighting, caused by different value ranges across the dimensions: the variance-normalization x'_{var} and max-normalization x'_{max}, calculated from x as unprocessed feature vector:

$$x'_{var} = \frac{x - mean(x)}{var(x)}, \qquad x'_{max} = \frac{x}{abs(x)}$$

For the variance normalization, $mean(x)$ and $var(x)$ are calculated from the training-set and test-set independently.

7.2 Experiments

As a preparation step, all samples of both datasets are shuffled and stratified split into 80 % training samples and 20 % test samples to ensure the same class sample number heterogeneity in training and test-set. To better assess the results of the machine and

representation learning approaches, we additionally implemented 3 classification models as baseline-reference models. In total, 7 different models are evaluated. The classification is implemented with four common classifiers: K-Nearest Neighbor (KNN [85]), Linear Discriminant Analysis (LDA [85]), Support Vector Machines (SVM [77]), and Binary Decision Trees (BDT [85]). The coding layer output of the AE and CAE are interpreted as a dimensionally reduced feature space and fed to the classifiers. Only the CNN is implemented as an end-to-end model and therefore already gives a classification as output.

The 36 implemented multidimensional *hand-crafted features* comprise 212 dimensions that are extracted from the raw waveform data to form the feature space. All 212 dimensions are considered in creating the feature space for this model.

The *AE* and *CAE* model are used as a general automated way to reduce the dimensions of the high dimensional raw waveform data. Since a linear AE usually reaches the same performance as a principal component analysis (PCA) [86], it is interesting to see whether the implemented multiple non-linear layers of the AE and CAE would result in any performance improvements compared to the PCA.

The *CNN* model is fed with the raw waveform data to learn from different receptive field sizes. The training-set is again split into 80 % training samples and 20 % validation samples. This validation-set allows for a training performance monitoring after each training-epoch and is used to properly evaluate the current classification performance mid-training and enables training strategies such as early stopping and saving the best model.

For the *random sub-sampling* model, we extracted a random selected subset (without repetition) of the 500 ms raw data samples (see Figure 7.2.1). To facilitate comparability to the hand-crafted feature set, the subset consists of 212 dimensions as well. This approach is equivalent to a non-equidistant sub-sampling and can be regarded as a very simple kind of feature extraction without any expert knowledge or comprehensive model architecture design.

Another model considers the root means square (RMS) energy of each mains cycle. The resulting 25 element long vector *RMS-25* shows the actual absolute current over the

500 ms (see Figure 7.2.2). Since every appliance draws a different consumption during switch-on - the RMS of the 25 mains cycles is simple to calculate, but it is a powerful discriminative model in terms of appliance recognition [25].

The *principal component analysis* is a common method for reducing feature space dimensions. The PCA reduces a high dimensional data space into a lower one by descending ordered variances. These variances form a new cartesian coordinate system. To keep the comparability, the 212 highest variances are considered in this model. Assuming that both sets share the same distribution, the variances are calculated on the training-set and the resulting covariance matrix is used to transform the unseen test-set.

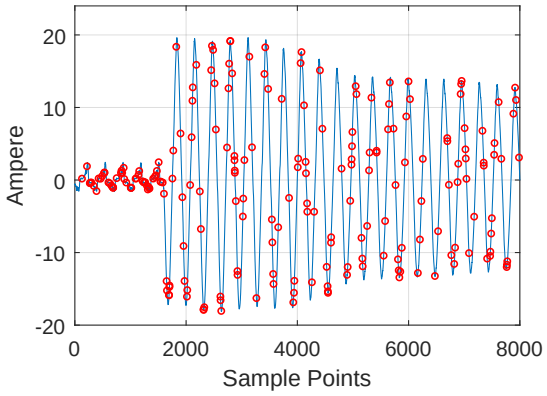


Figure 7.2.1: Random sub-sampling: The red circles show the randomly selected measurement points from the raw waveform current of a dishwasher event.

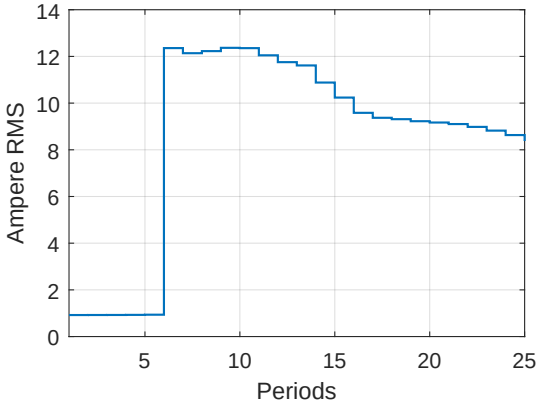


Figure 7.2.2: RMS-25: The blue stepped line shows the current for each mains cycle in the measured 500 ms segment of a dishwasher event.

7.3 Results

The classification performance for all seven experiments is calculated using the predicted output of the corresponding model and its classifier. The results show a rather heterogeneous distribution of performance. The overall best performance including both datasets could be achieved with the CNN model. Regarding the stand-alone classifier, SVM and KNN reach the highest classification performance on average, with a mean F-Score over all six models with 0.60 for KNN on UK-DALE and 0.75 for SVM on BLOND-50 (see Table 7.3.1). The overall best classification performance could be achieved with 0.75 with

Table 7.3.1: Average classifier F-Score

	KNN	LDA	SVM	BDT
UK-DALE	0.60	0.37	0.60	0.53
BLOND-50	0.69	0.54	0.75	0.61

the end-to-end CNN on UK-DALE and 0.87 with the hand-crafted features using the LDA classifier, closely followed by the CNN (see Figure 7.3.1 and 7.3.2).

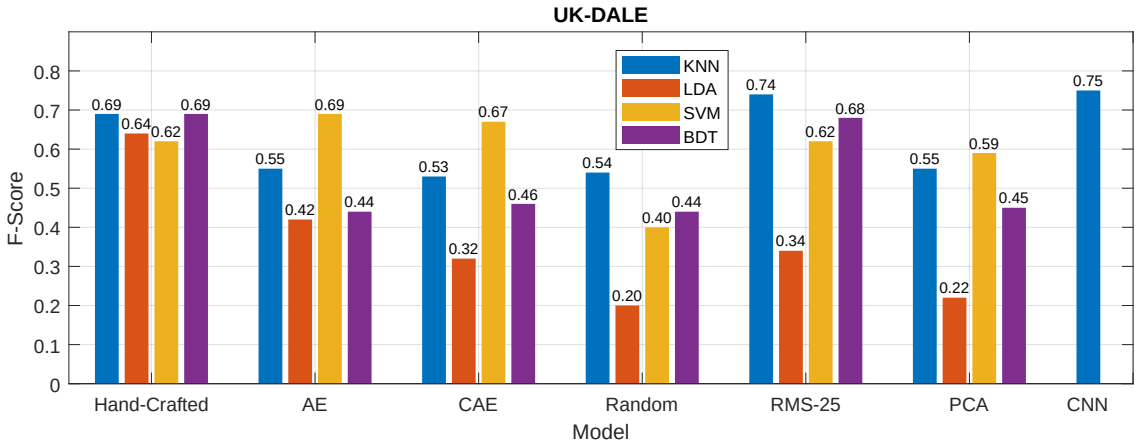


Figure 7.3.1: Appliance classification performance using the seven introduced classification models and five classifiers on the UK-DALE dataset.

Our interpretation of the results substantiates the observation: to replace the expert-driven hand-crafted feature extraction with a representation learning system, a large number of samples is necessary. Prevalent experiments on a lower number of appliance events led to much lower classification performance for the representation learning approaches. Humans are able to identify complex patterns and differences given only a few samples. The process of putting these patterns into metrics and numbers forms very powerful features, which is the main advantage of the expert-driven hand-crafted feature extraction.

7.3.1 Classification Models

Each classification model has been evaluated and performs differently for obvious and non-obvious reasons.

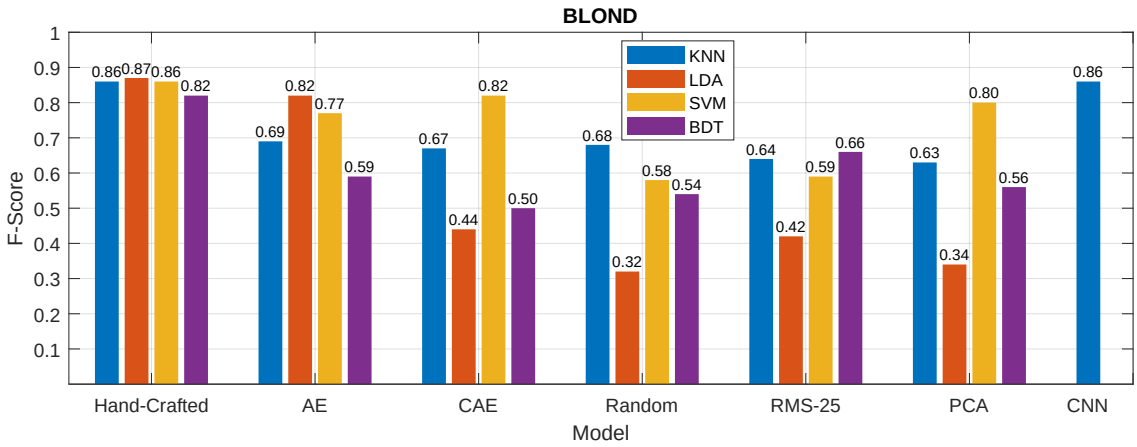


Figure 7.3.2: Appliance classification performance using the seven introduced classification models and five classifiers on the BLOND-50 dataset.

Table 7.3.2: Neural network parameter of the best-performing architectures

Architecture	AE		CAE		CNN	
Dataset	UK-DALE	BLOND-50	UK-DALE	BLOND-50	UK-DALE	BLOND-50
Dim-Scale / Layer	[2,4,5 - 5,4,2]	[10,5,2.5 - 2.5,5,10]	[5,4,2 - 2,4,5]	[5,5,5 - 5,5,5]	[5,2,2,2,2,2,2]	[5,5,5,2,2,2,2]
Batch-Norm	yes	yes	yes	yes	yes	yes
Activation	leaky relu	leaky relu	leaky relu	leaky relu	leaky relu	leaky relu
L2 Regul.	0,00001	0,00001	-	-	-	-
Normalization	variance	variance	variance	variance	variance	variance
Learning Rate	0,0001	0,0001	0,001	0,001	0,001	0,001
Batch Size	30	45	45	45	30	30
Noise	0,005	0,005	-	-	-	-
Optimizer	ADAM [87]	ADAM [87]	ADAM [87]	SGD	SGD	SGD
Loss Function	MSE	MSE	MSE	MSE	cat. cross-entr.	cat. cross-entr.

Hand-Crafted Features

The best results with an F-Score of 0.69 could be achieved by using the max-normalization and the binary decision tree classifier. Figure 7.3.1 and 7.3.2 show a very homogeneous performance across the four classifiers, making the hand-crafted feature extraction a stable and the second best model in this benchmark.

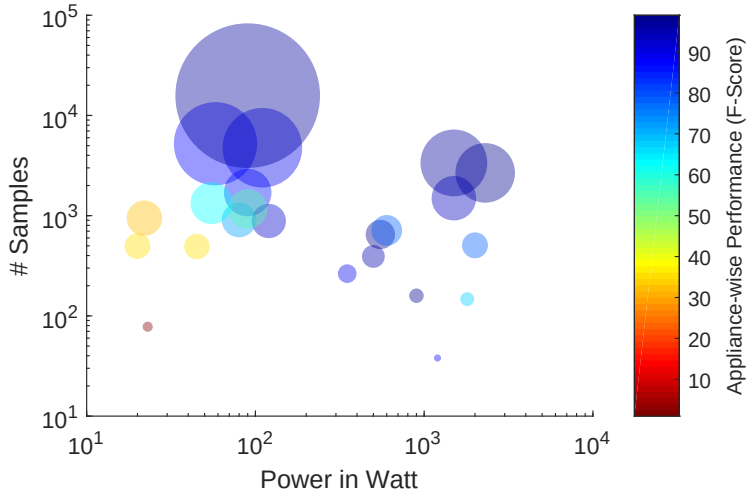


Figure 7.3.3: The figure shows the dependencies of the sample count and appliance power to the actual recognition performance of the individual appliances from UK-DALE. The appliance-marker size correlates with the number of samples true to scale, showing the huge differences in the the frequency with which appliance events occur.

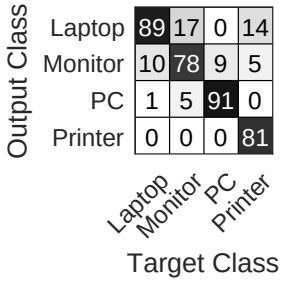


Figure 7.3.4: The confusion matrix of the best performing CNN model (F-Score 0.86) of BLOND-50, normalized to 100.

Autoencoder

The AE architecture with the best performance for UK-DALE comprises three encoding and decoding layers. The three fully connected encoding layers reduce the input by the factors 2, 4 and 5, similarly for the decoding layers. Since the performance with 0.69 for the best classifier is significantly higher compared to the PCA (0.59), some additional patterns in the feature space could be found by the non-linear layers. The best performance for BLOND-50 could be achieved with a three-layered architecture, with different reducing factors of 10, 5 and 2.5, similarly for the decoding layers and a batch size of 45 (see Table 7.3.2 for further details).

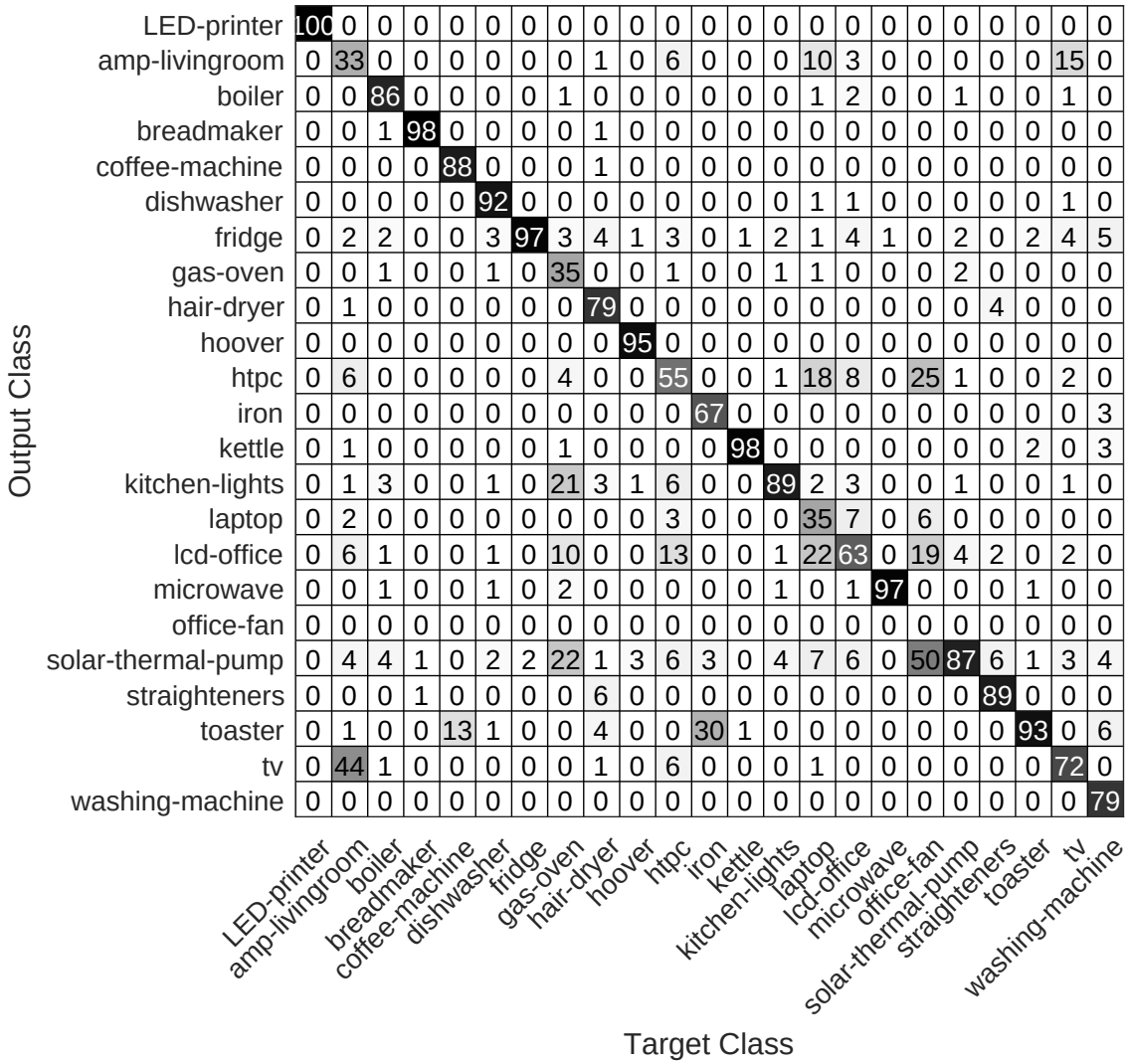


Figure 7.3.5: The confusion matrix of the best performing CNN model (F-Score 0.75) shows the misclassification of each considered class of the UK-DALE dataset normalized to 100. Note that the values are rounded to integers

Convolutional Autoencoder

The expected performance improvement of the CAE due to its convolutional layers could not be reached in our experiments. We assume that the chosen parameter space was too far from the actual optimum. However, the best performing architecture and its parameters for the CAE in these experiments can be seen in Figure 7.1.4 and Table 7.3.2.

Convolutional Neural Network

The best performing end-to-end CNN architecture comprises the architecture of Figure 7.1.3 and the parameter settings of Table 7.3.2. The end-to-end implementation entails that the last layer of the neural network gives a classification as output. The fact that the parameter search gave the identical optimal parameter set for both datasets underlines a good generalization capability of the model.

Random Selected Raw Dimensions

As expected and as the results show, this simple model of dimensional reduction does not allow a reliable and stable classification. With a mean F-Score of 0.39 for UK-DALE and 0.53 for BLOND-50, this model shows the worst performance in both cases. However, a pure random classification for the UK-DALE dataset would result in an F-Score of around 0.04, which is far below the performance of this model.

RMS-25

The energy of the mains cycles forms a powerful feature that allows a very high classification performance in combination with a spectral metric [25]. Surprisingly, the KNN classifier using the mains cycles forms the second best model for the UK-DALE dataset. The appliances of the UK-DALE dataset can be well distinguished, based on their individual startup energy consumption pattern only. Unfortunately, in the case of

the BLOND-50 dataset, the performance is only mid-range due to the different startup pattern of the individual appliances inside one appliance class.

PCA selected Dimensions

PCA is one of the most applied methods for reducing the feature space [30]. Therefore, the performance here is of interest. Since PCA is a linear transformation, not all information can be projected onto the lower feature space. Therefore, PCA usually performs worse than any well configured and trained neural network. Considering the simplicity of the algorithm and the absence of any expert knowledge, this still leaves PCA as an option.

7.3.2 Appliances

Regarding the best representation learning model (CNN) for UK-DALE, the average classification performance (F-Score) across all appliances lies at 0.75 (mean) and 0.86 (median). The four best recognized appliances are the *kettle* (0.97), *fridge* (0.97), *microwave* (0.96) and *breadmaker* (0.95). All of these appliances have in common that they are either represented by a huge number of samples or have a large power consumption.

The four worst recognized appliances are the *gas-oven* (0.42), *laptop* (0.41), *amp-livingroom* (0.39) and *office-fan* (0.0). Further analysis on these appliance events reveals that the *amp-livingroom* shows one very short, small and heterogeneous peak transient while the *gas-oven*, *laptop* and *office-fan* show a very low or even non-visible step in the power consumption. These observations and the fact that these four particular appliances have the lowest energy consumptions (see Figure 7.3.3) of the whole appliance set, leads us to the assumption that their consumption is simply too low to distinguish properly from the background noise of the aggregated signal. The recognition of *laptops* in BLOND-50 is significantly better, supporting the statement that the issue is regarded to these particular appliances. All the remaining appliances in UK-DALE were recognized correctly in most cases. Regarding BLOND-50, CNN could generalize very well over the multiple appliance models inside each class. The remaining misclassification of *monitor* and *laptop* are due to their similar power consumption.

8

Use Case Study

NILM provides several techniques for demand information retrieval to support consumers saving energy usage. Research in NILM often focuses on closed environments, such as single datasets or single households. Disaggregation results are typically not suitable to represent the classification performance under real circumstances due to its data homogeneity of a single dataset. We apply a classification system across four commonly available high-frequency sampled energy datasets. The experiments include classification tasks with four different classifiers on 36 spectral and temporal features to perform a cross-, mixed-, and intra-dataset validation. The outcome of this work is a reliable benchmark for appliance recognition in the high-frequency domain and its efficiency in smart meters for different use cases and appliance features.

8.1 Approach

For our experiments, we chose four publicly available high-frequency sampled datasets that provide a fully labeled appliance ground truth and share a considerable set of appliances. **PLAID** provides a public library of high-resolution appliance measurements in an isolated environment. The appliance events are fully labeled and sampled at 30 kHz with 16-bit resolution. **WHITED** provides multiple measurements of typical domestic

appliance events across different regions of the world. The recording hardware is based on a custom sound card meter that sampled fully labeled isolated events at 44.1 kHz and 16-bit resolution. **BLUED** is a fully-labeled energy dataset of a single-family residence with a 60 Hz mains frequency in the USA. The measurements were sampled with an NI USB-9215A at 12 kHz and 16-bit resolution. The appliance event ground truths are provided as labeled time stamps. **UK-DALE** consists of measurements from multiple domestic houses in the UK that were recorded with a custom sound card meter. The aggregated high-frequency signals are sampled at 16 kHz and 20-bit resolution. The low frequency appliance level measurements serve also as a fully-labeled ground truth. The appliance events occur within a time window of around 15 seconds around the low frequency event time stamp and might be aligned to another event that occurs in the same time window. Only three appliances of the COOLL dataset [34] match with the other datasets appliances which is too low to consider that dataset in this work.

The differences between these datasets can be interpreted as variances of sampling frequency and resolution, line noise, environment (isolated vs. real household), mains frequency, location (different voltage characteristics) and set of appliances. Those variances can be termed as dataset-bias, a common term in image datasets [88], where this issue is already discussed. Since different datasets have their own characteristics, the above mentioned variances can make them distinguishable - which should not be the case! The dataset-bias leads to the problem that the classifier sometimes learns the dataset characteristics instead of the class characteristics. The goal in the following feature evaluation is to identify appliance signatures that generalize best over all datasets and are least influenced by the dataset-bias.

For our experiments, we considered only appliances that occur in at least 3 datasets and have more than 5 start-up events. Therefore it was necessary to leave some appliances out. We define an appliance start-up event as a spontaneous, significant rise in the current consumption that lasts for a couple of mains periods. The rise threshold varies through the datasets and depends on the appliance with the lowest consumption. The individual appliance textual labels given by the authors were consistently renamed and tagged with a textual label of the dataset they belong to (Table 8.1.1). With this dataset label, we are able to divide the appliances in training set and test set for the cross-dataset-validation (Fig. 8.1.1).

Table 8.1.1: Extracted appliance events for each dataset

Appliance	UK-DALE	BLUED	PLAID	WHITED
Fan	78	-	108	60
Fridge	15572	294	34	10
Hair Dryer	324	-	153	60
Iron	114	17	-	30
Laptop	490	-	165	20
Microwave	1295	-	139	30
Monitor	1373	27	-	20
Printer	151	75	-	10
TV	590	27	-	20
Vacuum Cleaner	289	-	38	40
Washing Machine	329	-	25	10

The feature extraction was applied on a 500 ms region of interest from a total of 22017 appliance events, resulting in a 22017x212 sized feature space matrix. Our feature set consists of 36 features with multiple dimensions from related research fields such as music information retrieval, speech recognition and general signal processing. The features are designed to be sampling rate independent. The characteristics of audio and speech signals show similarities to energy data which motivates the usage of features for audio and speech processing. While the signal envelope for musical instruments is build of attack, decay, sustain and release states, electrical appliances always have a start-up, decay, steady state, and turn off. The considered features are introduced and discussed in the work of [25].

With the use of these feature spaces, we applied three classification experiments to evaluate the particular recognition performance and discuss the individual feature ranking for each of the three experiments. The considered classifiers include the k-nearest neighbor (KNN [85]), the linear discriminant analysis (LDA [85]), the support vector machines (SVM [85]), and the binary decision tree (BDT [85]). Initial evaluation of the data suggests that the best results for this setup can be achieved with $K = 4$ for KNN with city block distance metric. For the SVM classification, the external LIBSVM library [77] is used and allows a fast multi-class classification. Due to performance reasons, no `c` and `gamma` parameter selection was considered.

8.1.1 Cross-Dataset-Validation

We applied a leave-one-dataset-out setup to compare the classification performance for each dataset (Fig. 8.1.1). The appliances of the considered datasets are classified. The classification is based on a model trained from all samples of the remaining datasets. In this case, it is ensured that each appliance is trained from a minimum of two different appliances from two different datasets. This experiment represents the use case of a smart meter that is equipped with a factory pre-trained appliance model to recognize appliances in an unknown environment.

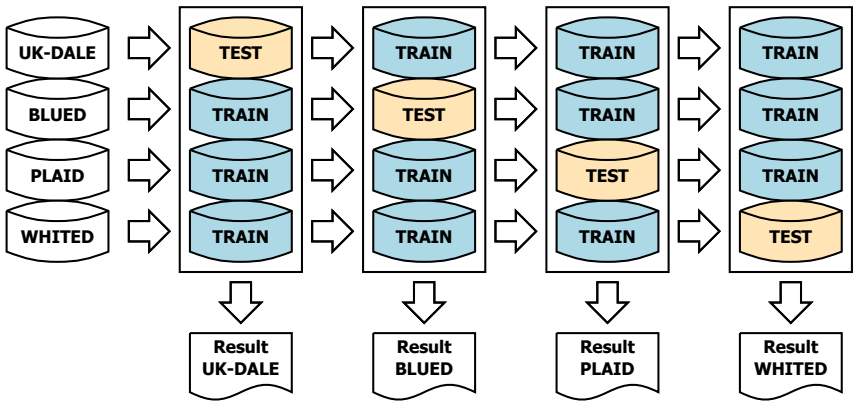


Figure 8.1.1: Schematic overview of the cross-dataset validation experiment

8.1.2 Mixed-Dataset Cross-Validation

In this experiment, the samples of all datasets are mixed together and are only distinguishable by their labels. We applied a randomized stratified 5-fold cross-validation across all samples, ignoring their dataset origin. The classifier trains samples from all datasets and classifies unknown samples of these datasets. The results are expected to be significantly better because the model already knows the environmental aspects of the dataset and the special appliance itself. This experiment represents the use-case of a smart meter that is equipped with a factory pre-trained appliance model including user trained samples.

8.1.3 Intra-Dataset Cross-Validation

In this experiment the datasets are independently classified. This means that all samples of one dataset were given to a randomized stratified 5-fold cross validation to retrieve the classification performance based on samples only for this dataset. Since the samples in one class are very homogeneous, the classifier usually shows promising results. This experiment represents the use case of a smart meter without any factory pre-trained appliance model and is exclusively user trained on measurements in the current environment.

8.1.4 Feature evaluation

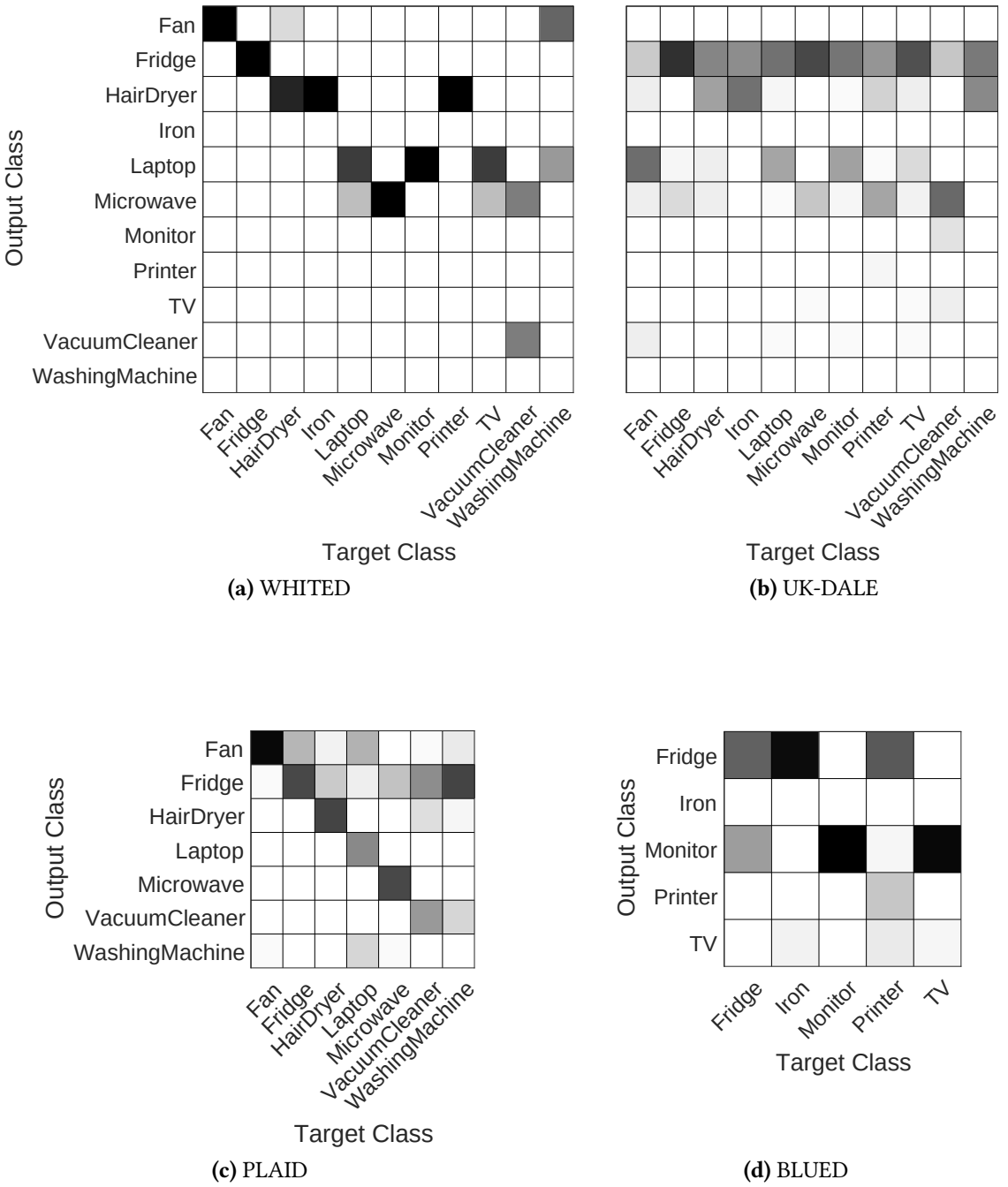
One important outcome of this work is a suggestion of features that are able to generalize through all the dataset biased variances. The workflow of this experiment consists of a single classification performance for each individual feature with the K-NN classifier. The result is a 36 dimensional feature ranking for the cross- mixed-, and intra-dataset validation.

8.2 Results

Our experiments show that it is a difficult challenge to recognize appliances that have not been seen during training of the classification system. The results show that the dataset-bias has a big impact on the classification performance.

8.2.1 Cross-Dataset-Validation

This experiment is divided into four steps, corresponding to each dataset that its samples are being classified with a model that is trained from the remaining datasets.



(a) WHITED **(b)** UK-DALE

(c) PLAID **(d)** BLUED

Figure 8.2.1: Confusion matrices of the cross-dataset validation experiments. The gray-tone intensity correlates to the percentage appliance distribution. The optimal classification would show a black main diagonal.

Table 8.2.1: Cross-dataset classification results

Dataset	Metric	KNN	LDA	SVM	BDT
BLUED	F-Score:	0.25	0.22	0.22	0.18
	Precision:	0.38	0.42	0.37	0.23
	Recall:	0.36	0.37	0.34	0.22
	Accuracy:	0.80	0.72	0.79	0.77
PLAID	F-Score:	0.56	0.16	0.23	0.19
	Precision:	0.66	0.34	0.42	0.40
	Recall:	0.58	0.23	0.25	0.26
	Accuracy:	0.91	0.76	0.81	0.80
UK-DALE	F-Score:	0.13	0.08	0.15	0.07
	Precision:	0.13	0.10	0.14	0.10
	Recall:	0.17	0.12	0.16	0.10
	Accuracy:	0.93	0.83	0.95	0.85
WHITED	F-Score:	0.38	0.15	0.28	0.28
	Precision:	0.36	0.15	0.24	0.28
	Recall:	0.46	0.25	0.40	0.35
	Accuracy:	0.93	0.87	0.90	0.90

The BLUED confusion matrix (Fig. 8.2.1d) shows that the Iron was recognized as a Fridge for almost all of its events. Those consistent misclassifications can be observed several times in all four confusion plots. We believe that those misclassifications are caused by the sum of all variances in the dataset-bias which leads to shifted class centers in the feature space. Since the bias for each dataset is not a priori known, the WHITED Iron will consistently be misclassified as Hairdryer.

We also observe that the isolated samples of WHITED, which have a low inner-class variance, lead to more consistent misclassifications than the UK-DALE dataset. This behavior seems plausible since high, overlapping class variances lead to a broader set of misclassification.

Another observation is the high amount of misclassification in favor of the Fridge without any obvious cause. A test run with using only 10 % of UK-DALE's Fridge events does not change this observation significantly.

A reliable classification is not possible with this setup. Either a much bigger amount of training samples is necessary or the unknown components of the dataset-bias need to be removed. This setup would not allow a sufficient classification performance for a smart meter use case, see Table 8.2.1.

8.2.2 Mixed-Dataset Cross-Validation

Results in this experiment are remarkably better in comparison to the previous experiment. Since the classifier knows samples and appliances of each dataset, the dataset-bias is not playing a role anymore.

Table 8.2.2: Mixed-Dataset Classification Results

Metric	KNN	LDA	SVM	BDT
F-Score:	0.87	0.80	0.88	0.85
Precision:	0.88	0.81	0.92	0.85
Recall:	0.86	0.80	0.86	0.85
Accuracy:	0.99	0.99	0.99	0.99

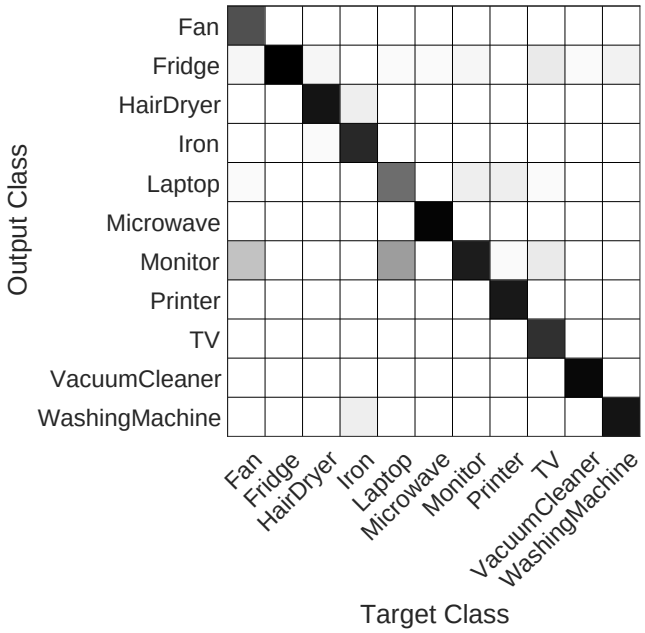


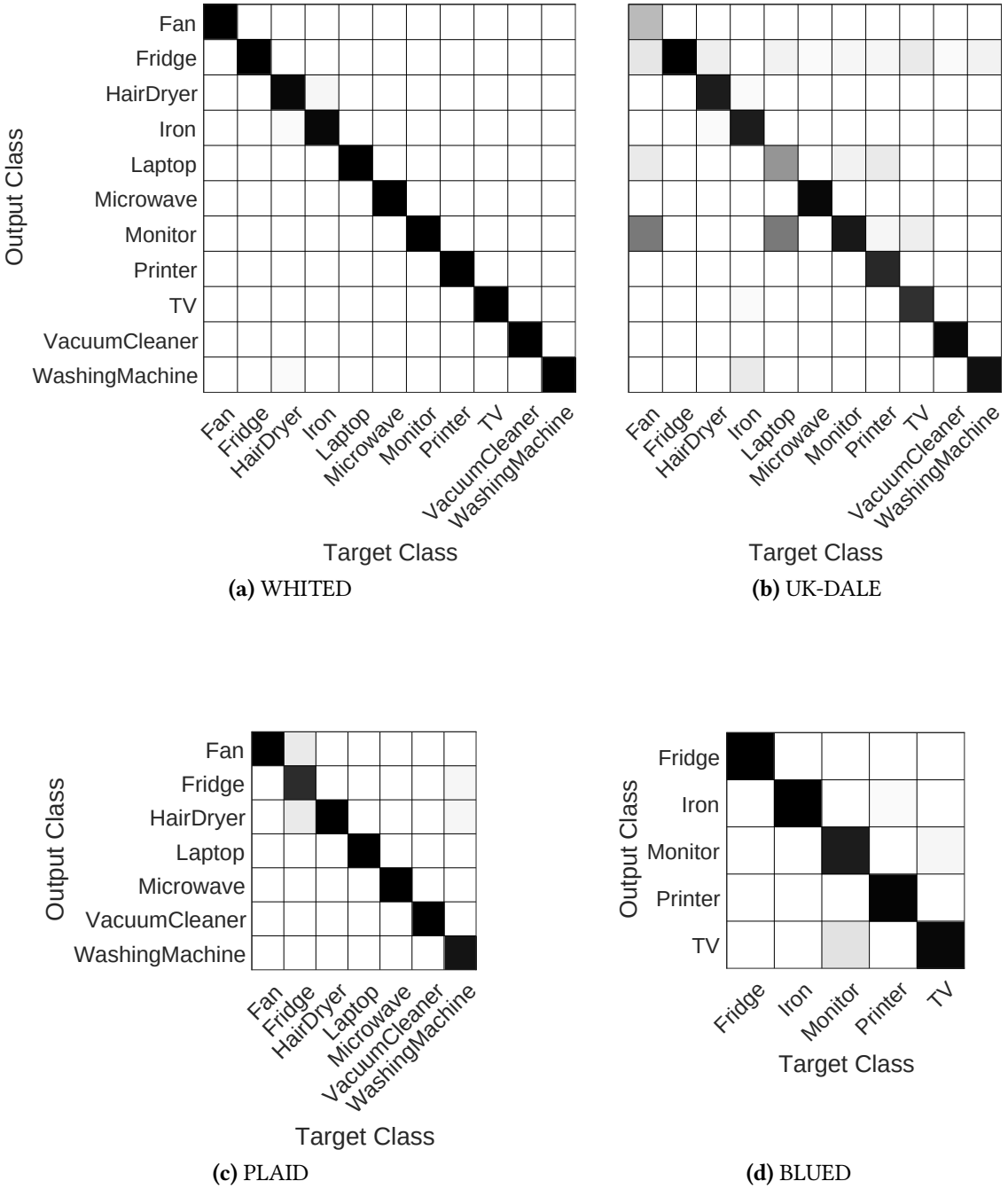
Figure 8.2.2: Confusion matrix - mixed-dataset cross-validation

Only the classes Laptop and Fan are often misclassified as Monitor (Fig. 8.2.2). Laptops and Monitors usually share a similar power consumption and are equipped with a SMPS that draws very unique characteristics. These characteristics help to recognize an SMPS, but makes it harder to recognize the actual appliance behind it. Therefore, Laptops are often recognized as Monitors and some Monitors are recognized as Laptops. Beside some minor misclassification, the recognition performance is on a relative high level and allows a sufficient appliance recognition for a smart meter use case, see Table 8.2.2.

8.2.3 Intra-Dataset Cross-Validation

If the classifier's task is to recognize appliances inside a homogeneous environment such as a single dataset, the results are significantly better than in any of the previous experiments. The classifier usually knows the individual characteristics of the distinct appliances from training samples.

The Fan and Laptop of UK-DALE are mostly misclassified as Monitor due to yet unknown circumstances. This experiment shows that an appliance recognition system is able to distinguish between a limited amount of appliances, even for a real environment that includes challenges like noise and concurrently running appliances, see Table 8.2.3.



(a) WHITED **(b)** UK-DALE

(c) PLAID **(d)** BLUED

Figure 8.2.3: Confusion matrices of the cross-dataset validation

Table 8.2.3: Intra-Dataset Classification Results

Dataset	Metric	KNN	LDA	SVM	BDT
BLUED	F-Score:	0.91	0.96	0.93	0.93
	Precision:	0.93	0.95	0.94	0.93
	Recall:	0.91	0.97	0.93	0.93
	Accuracy:	0.99	0.99	0.99	0.99
PLAID	F-Score:	0.97	0.93	0.94	0.92
	Precision:	0.99	0.94	0.97	0.92
	Recall:	0.96	0.93	0.92	0.91
	Accuracy:	1.00	0.99	0.99	0.99
UK-DALE	F-Score:	0.81	0.81	0.83	0.80
	Precision:	0.82	0.80	0.90	0.80
	Recall:	0.80	0.84	0.80	0.80
	Accuracy:	0.99	0.99	0.99	0.99
WHITED	F-Score:	0.99	0.96	0.94	0.99
	Precision:	0.99	0.96	0.96	0.98
	Recall:	0.99	0.97	0.94	0.99
	Accuracy:	1.00	0.99	0.99	1.00

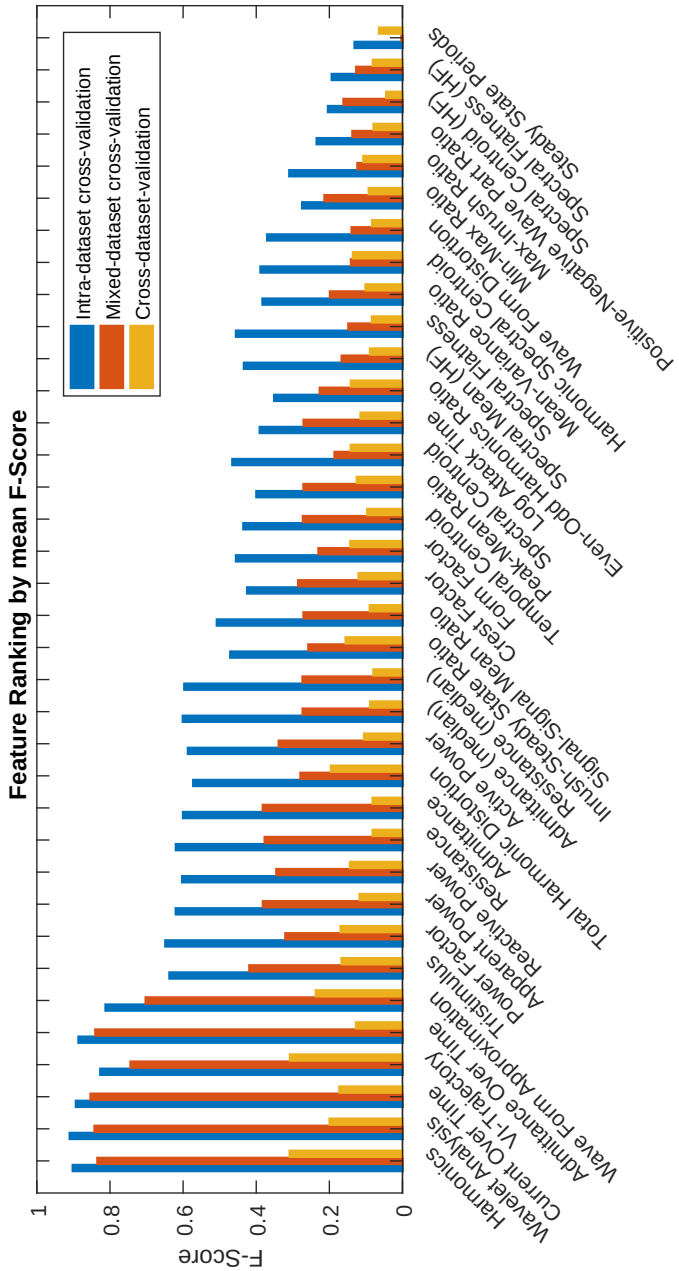


Figure 8.2.4: The feature ranking graph shows the performance for each feature and each experiment.

8.2.4 Feature validation

For each of the previous three experiments, we applied a small feature study. The F-Score of each experiments K-NN classification task was retrieved for every individual feature. In the Section 8.2.1 and 8.2.3, we also get a result for each feature and each dataset. In this case, we calculated the mean performances over the four dataset results. Fig. 8.2.4 shows the individual results for each experiment in another color, while the features are descending sorted, based on their over-all-experiments performance.

The amplitude of the first 20 harmonics in ratio to its mains frequency shows the best overall results. The result is plausible since every appliance has its unique harmonics energy distribution footprint and the harmonics ratio is less influenced by any of the dataset-related variances. Generally we observe that spectral and waveform based features perform better in the cross-dataset-validation, while power- and temporal-based metrics perform also very well for intra-dataset cross-validation. It is fair to say that spectral- and waveform-based features are less influenced by dataset-related variances (dataset-bias) and should therefore be considered for such use cases.

8.3 Discussion

To retrieve a measurable value for the dataset-bias, the set of appliances must be absolutely identical in all datasets, so that the only difference in the measurements is the dataset-bias. The effect of the dataset-bias would have been zero if the results of the cross-dataset-validation lies in the same range as the results for the intra-dataset cross-validation, which is not the case in our experiments. Since each dataset has a different set of appliances and two datasets have multiple appliances for each appliance type (PLAID, WHITED), a measurable influence of the dataset-bias could not be achieved.

9

Conclusions

The potential in reducing residential energy consumption can be considered high, since it depends highly on human behavior and established understanding of comfort. According to Kuckartz, Rädiker, and Rheingans-Heintze [89], around 96% of German citizens agree that the consumer plays a crucial role in saving energy. Supporting consumers with consumption feedback is one of the main use cases of NILM. Targeting the main drawbacks of current NILM systems is the purpose of this work.

We introduced a new dataset of a broad range of household and small industry appliance start-up transients that helps to extract and to evaluate appliance-specific features. We could show that even low-budget hardware allows one to retrieve appliance features with high discriminative potential.

To find appliance start-ups, we proposed a multivariate event detection that learns from a consumer formed event model. The event model stems from event and non-event segments of the training set and allows a user relevant event detection. The challenge to distinguish between relevant and irrelevant events is tackled by multiple runs of the introduced adaptive training process that allows a reduction of false positives by up to factor of eight. The multivariate event detection in combination with the introduced way of adaptive training is an appropriate step towards event detection for the increasing number of SMPS-driven appliances in residential and office environments.

To provide an efficient solution for smart meters that are usually equipped with a limited processing unit, we performed an appliance recognition and feature evaluation. The evaluation comprises four different high-frequency datasets to identify the best appliance features and feature combinations out of 36 implemented signatures from different research areas. We showed that the phase angle difference between voltage and current has the highest scalar performance across all datasets, while the multi-dimensional feature *Current Over Time* shows the most promising results in general. According to our findings, the *Wavelet Analysis* discriminates best for isolated environments while the *Current Over Time* scores best for aggregated environments. Unfortunately, there is no one-size-fits-all combination of features. The composition of appliances, their usage, and the environment determines the combination of features. However, the findings show that the best feature sets always contain information about a spectral energy distribution (e.g., *Wavelet Analysis*, *Harmonics*) and time-related power information (e.g., *Current and Admittance Over Time*). Furthermore, the results of the classification performance for each appliance type shows that almost all appliances are recognized in most cases with the introduced combination of features.

To gain insights into the performance of modern representation learning approaches on appliance classification, we conducted an evaluation of several appliance classification models for two publicly available real-world energy consumption datasets. The classification models include conventional domain expert supported hand-crafted feature extraction, baseline-reference models and promising deep neural network models. The results of our experiments show comparable performances of classical machine learning and representation learning with a slight winning margin for the end-to-end implementation of the convolutional neural network (CNN). Our performance results support the statement that the representation learning approach is a worthy alternative to the classical machine learning processing-chain for appliance recognition systems in NILM. The effort for gaining expert-based features on the one side, neural network architecture and parameter search effort on the other side, as well as training data volume, are most likely the main decision criteria if the recognition system shall be based on a classical machine learning or representation learning framework.

Appliance recognition is a requirement for power disaggregation and a challenging task for advanced smart meters. The recognition performance is significantly influenced by

the type of the trained appliance model. We evaluated a cross-dataset, mixed-dataset, and intra-dataset recognition that represent three possible use cases of a recognition system in a smart meter. While the classification performance - probably due to dataset-bias - is quite weak for the first case, the mixed and intra-dataset recognition provides high performance for almost all appliances in laboratory-like and real environments. The observations additionally show that waveform and spectral features allow the best results for cross-dataset validation while power and temporal changes lead for intra- and mixed-dataset validation. The results allow a recommendation of the mixed or internally trained recognition system for a smart meter that implements a power disaggregation algorithm.

Future work may target the following aspects, which we see as promising:

More appliance Measurements Crowd sourced measurements with comparable measurement equipment, in more regions, as well as more appliances and types, are necessary to increase the general appliance models. The recognition of geographical regions based on characteristics in the voltage signal may be possible, but requires more measurements in more regions than is currently provided in WHITED.

Wide agreement of event definition A systematic evaluation of appliance event detection demands for a uniform understanding of appliance events and their breakdown into an event taxonomy. A broadly accepted event taxonomy might be necessary to distinguish between appliance event types.

Evaluation of NILM process Since most studies focus on specific aspects of the NILM process, it seems worthwhile to evaluate the components that performs best in an end-to-end implementation. Strategies which perform well independently might perform worse in the aggregate context of the process.

Think outside the box It is known that the techniques used in NILM can be applied for other flowing matters including water, gas, and oil. NILM can be used for other research questions beside energy consumption feedback such as building automation, patient care, demand response, and predictive maintenance. Techniques from similar research fields may contribute to NILM. We see chances of interdisciplinary exchange in these aspects for future studies.

Target for Consumer interest The current state of NILM in commercial products can be seen as unsuccessful or niche products. Many people are concerned about the potential of smart meters to reduce energy consumption. The main reasons are privacy, a lack of personal advantage, as well as plain disinterest amongst consumers. Keeping NILM alive as a technique for energy consumption feedback would need to arouse interest in reducing energy consumption with novel strategies such as gamification.

IoT vs. NILM Internet of things is being discussed as an attractive ILM alternative to NILM. We believe that combined solutions may provide higher usefulness than strictly following one paradigm.

Glossary

AC Alternating Current

ADC Analog Digital Converter

AE Autoencoder

BDT Binary Decision Tree

BLOND Building-Level Office eNvironment Dataset

BLUED Building-Level fUlly-labeled dataset for Electricity Disaggregation

CAE Conventional Autoencoder

CFL Compact Fluorescent Light

CLEAR Circuit-Level Energy Appliance Radar

CNN Conventional Neural Network

COOLL Controlled ON / OFF Loads Library

CUSUM Cumulative Sum

DSP Digital Signal Processing

FN False Negatives

FP False Positives

GOF Goodnes Of Fit

HF High Frequency

HFED High Frequency EMI Data Set

ILM Intrusive Load Monitoring

KNN K-Nearest Neighbour

LDA Linear Discriminant Analysis

LILACD Laboratory-measured Industrial Load of Appliance Characteristics

MEDAL Mobile Energy Data Acquisition Laboratory

NILM Non-Intrusive Load Monitoring

NILMTK NILM Tool Kit

PCA Principal Component Analysis

PLAID Plug Load Appliance Identification Dataset

PR Precision

RE Recall

REDD Reference Energy Disaggregation Data Set

RMS Root Means Square

ROI Region Of Interest

SCP Switch Continuity Principal

SMPS Switchin-Mode Power Supply

SNR Signal Noise Ratio

SVM Support Vector Machines

THD Total Harmonic Distortion

TN True Negatives

TP True Positives

UK-DALE UK Domestic Appliance-Level Electricity dataset

WHITED Worldwide Household and Industry Transient Energy Dataset

List of Figures

List of Tables

Bibliography

[1] G. W. Hart. "Nonintrusive Appliance Load Monitoring." In: *Proceedings of the IEEE* 80.12 (1992), pp. 1870–1891.

[2] H. A. D. Azzini, R. Torquato, and L. C. P. d. Silva. "Event detection methods for nonintrusive load monitoring." In: *2014 IEEE PES General Meeting*. 2014, pp. 1–5. DOI: 10.1109/PESGM.2014.6939797.

[3] A. Reinhardt, D. Burkhardt, M. Zaheer, and R. Steinmetz. "Electric Appliance Classification Based on Distributed High Resolution Current Sensing." In: *37th Annual IEEE Conference on Local Computer Networks, Workshop Proceedings*. IEEE, Oct. 2012, pp. 999–1005. DOI: 10.1109/lcnw.2012.6424093.

[4] Y. Lin and M. Tsai. "An Advanced Home Energy Management System Facilitated by Nonintrusive Load Monitoring With Automated Multiobjective Power Scheduling." In: *IEEE Transactions on Smart Grid* 6.4 (July 2015), pp. 1839–1851. ISSN: 1949-3053. DOI: 10.1109/TSG.2015.2388492.

[5] K. C. Armel, A. Gupta, G. Shrimali, and A. Albert. "Is disaggregation the holy grail of energy efficiency? The case of electricity." In: *Energy Policy* 52 (Jan. 2013), pp. 213–234. DOI: 10.1016/j.enpol.2012.08.062.

[6] R. W. Cox, P. L. Bennett, T. D. McKay, J. Paris, and S. B. Leeb. "Using the Non-Intrusive Load Monitor for Shipboard Supervisory Control." In: *2007 IEEE Electric Ship Technologies Symposium*. 2007. DOI: 10.1109/ESTS.2007.372136.

[7] B. Ellert, S. Makonin, and F. Popowich. "Appliance Water Disaggregation via Non-intrusive Load Monitoring (NILM)." In: *Smart City 360°*. Ed. by A. Leon-Garcia, R. Lenort, D. Holman, et al. Cham: Springer International Publishing, 2016, pp. 455–467. ISBN: 978-3-319-33681-7.

[8] A. Veit, C. Goebel, R. Tidke, C. Doblander, and H.-A. Jacobsen. "Household Electricity Demand Forecasting: Benchmarking State-of-the-art Methods." In: *Proceedings of the 5th International Conference on Future Energy Systems*. e-Energy '14. Cambridge, United Kingdom: ACM, 2014, pp. 233–234. ISBN: 978-1-4503-2819-7. DOI: 10.1145/2602044.2602082. URL: http://doi.acm.org/10.1145/2602044.2602082.

[9] A. U. Haq and H.-A. Jacobsen. "Prospects of Appliance-Level Load Monitoring in Off-the-Shelf Energy Monitors: A Technical Review." In: *Energies* 11.1 (2018). ISSN: 1996-1073. DOI: 10.3390/en11010189.

[10] J. Cook, N. Oreskes, P. T. Doran, et al. "Consensus on consensus: a synthesis of consensus estimates on human-caused global warming." In: *Environmental Research Letters* 11.4 (Apr. 2016), p. 048002. DOI: 10.1088/1748-9326/11/4/048002. URL: https://doi.org/10.1088%2F1748-9326%2F11%2F4%2F048002.

[11] International Energy Agency (IEA). *Key World Energy Statistics 2018*. 2018, p. 51. DOI: doi:10.1787/key_energ_stat-2018-en. URL: %5Curl%7Bhttps://www.oecd-ilibrary.org/content/publication/key%5C_energ%5C_stat-2018-en%7D.

[12] J. Hansen, D. Johnson, A. Lacis, et al. "Climate Impact of Increasing Atmospheric Carbon Dioxide." In: *Science* 213.4511 (1981), pp. 957–966. ISSN: 0036-8075. DOI: 10.1126/science.213.4511.957. eprint: http://science.sciencemag.org/content/213/4511/957.full.pdf. URL: http://science.sciencemag.org/content/213/4511/957.

[13] J. Kelly and W. Knottenbelt. "Does disaggregated electricity feedback reduce domestic electricity consumption? A systematic review of the literature." In: *CoRR* (2016).

[14] K. S. Barsim, L. Mauch, and B. Yang. "Neural Network Ensembles to Real-time Identification of Plug-level Appliance Measurements." In: *Signature* 2 (2016), p. 11.

[15] M. Kahl, T. Kriechbaumer, A. U. Haq, and H.-A. Jacobsen. "Appliance Classification Across Multiple High Frequency Energy Datasets." In: *2017 IEEE International Conference on Smart Grid Communications (SmartGridComm)*. 2017. DOI: 10.1109/smartgridcomm.2017.8340664.

[16] L. De Baets, J. Ruyssinck, C. Develder, T. Dhaene, and D. Deschrijver. "Appliance classification using VI trajectories and convolutional neural networks." In: *Energy and Buildings* 158.Supplement C (2018), pp. 32–36. ISSN: 0378-7788. DOI: 10.1016/j.enbuild.2017.09.087. URL: http://www.sciencedirect.com/science/article/pii/S0378778817312690.

[17] K. D. Anderson, M. E. Bergés, A. Ocneanu, D. Benitez, and J. M. F. Moura. "Event detection for Non Intrusive load monitoring." In: *IECON 2012 - 38th Annual Conference on IEEE Industrial Electronics Society*. Oct. 2012, pp. 3312–3317. DOI: 10.1109/IECON.2012.6389367.

[18] L. D. Baets, J. Ruyssinck, D. Deschrijver, and T. Dhaene. "Event detection in NILM using cepstrum smoothing." In: *3rd International Workshop on Non-Intrusive Load Monitoring*. 2016, pp. 1–4.

[19] A. A. Girmay and C. Camarda. "Simple event detection and disaggregation approach for residential energy estimation." In: *Proceedings of the 3rd International Workshop on Non-Intrusive Load Monitoring (NILM)*. 2016.

[20] K. Anderson, A. Ocneanu, D. Benitez, et al. "BLUED: A Fully Labeled Public Dataset for Event-Based Non-Intrusive Load Monitoring Research." In: *Proceedings of the 2nd KDD Workshop on Data Mining Applications in Sustainability (SustKDD)*. Beijing, China: ACM, Aug. 2012.

[21] J. Kelly and W. Knottenbelt. "The UK-DALE Dataset, Domestic Appliance-Level Electricity Demand and Whole-House Demand from Five UK Homes." In: *Scientific Data* 2 (2015), p. 150007. DOI: 10.1038/sdata.2015.7. URL: http://www.doc.ic.ac.uk/~dk3810/data/.

[22] J. Gao, S. Giri, E. C. Kara, and M. Bergés. "PLAID: A Public Dataset of High-Resolution Electrical Appliance Measurements for Load Identification Research." In: *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*. ACM. Association for Computing Machinery (ACM), 2014, pp. 198–199. DOI: 10.1145/2674061.2675032.

[23] M. Kahl, A. U. Haq, T. Kriechbaumer, and H.-A. Jacobsen. "WHITED - A Worldwide Household and Industry Transient Energy Data Set." In: *3rd International Workshop on Non-Intrusive Load Monitoring*. 2016.

[24] T. Kriechbaumer and H.-A. Jacobsen. "BLOND, a building-level office environment dataset of typical electrical appliances." In: *Scientific Data, an open-access NatureResearch journal* 5.180048 (2018). DOI: 10.1038/sdata.2018.48.

[25] M. Kahl, A. U. Haq, T. Kriechbaumer, and H.-A. Jacobsen. "A Comprehensive Feature Study for Appliance Recognition on High Frequency Energy Data." In: *Proceedings of the 2017 ACM 8th International Conference on Future Energy Systems* (May 18, 2017). e-Energy '17. Hong Kong, Hong Kong: ACM, May 18, 2017. ISBN: 978-1-4503-5036-5. DOI: 10.1145/3077839.3077845.

[26] T. Hassan, F. Javed, and N. Arshad. "An Empirical Investigation of V-I Trajectory Based Load Signatures for Non-Intrusive Load Monitoring." In: *IEEE Transactions on Smart Grid* 5.2 (Mar. 2014), pp. 870–878. DOI: 10.1109/tsg.2013.2271282.

[27] W. Wichakool, Z. Remscrim, U. A. Orji, and S. B. Leeb. "Smart Metering of Variable Power Loads." In: *IEEE Transactions on Smart Grid* 6.1 (Jan. 2015), pp. 189–198. ISSN: 1949-3053. DOI: 10.1109/TSG.2014.2352648.

[28] Y. LeCun, Y. Bengio, and G. Hinton. "Deep learning." In: *nature* 521.7553 (2015), p. 436.

[29] G. Hinton, L. Deng, D. Yu, et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." In: *IEEE Signal processing magazine* 29.6 (2012), pp. 82–97.

[30] A. Géron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 1 edition. O'Reilly Media, 2017. ISBN: 978-1491962299.

[31] J. Z. Kolter and M. J. Johnson. "REDD: A Public Data Set for Energy Disaggregation Research." In: *Workshop on Data Mining Applications in Sustainability (SIGKDD)*. Vol. 25. ACM, 2011, pp. 59–62.

[32] S. Makonin. "Investigating the Switch Continuity Principle Assumed in Non-Intrusive Load Monitoring (NILM)." In: *2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*. Institute of Electrical and Electronics Engineers (IEEE), May 2016. DOI: 10.1109/ccece.2016.7726787.

[33] M. Gulati, S. Sundar Ram, and A. Singh. "An in depth study into using EMI signatures for appliance identification." In: *Proceedings of the First ACM International Conference on Embedded Systems For Energy-Efficient Buildings*. ACM. Association for Computing Machinery (ACM), 2014. DOI: 10.1145/2674061.2674070.

[34] T. Picon, M. N. Meziane, P. Ravier, et al. "COOLL: Controlled On/Off Loads Library, a Public Dataset of High-Sampled Electrical Signals for Appliance Identification." In: *CoRR* abs/1611.05803 (2016). URL: http://arxiv.org/abs/1611.05803.

[35] M. Kahl, V. Krause, R. Hackenberg, et al. "Measurement system and dataset for in-depth analysis of appliance energy consumption in industrial environment." In: *tm-Technisches Messen* 86.1 (2019).

[36] J. Liang, S. K. Ng, G. Kendall, and J. W. Cheng. "Load Signature Study–Part I: Basic Concept, Structure, and Methodology." In: *Power Delivery, IEEE Transactions on* 25.2 (2010), pp. 551–560. DOI: 10.1109/tpwrd.2009.2033799.

[37] M. Baranski and J. Voss. "Detecting patterns of appliances from total load data using a dynamic programming approach." In: *Data Mining, 2004. ICDM '04. Fourth IEEE International Conference on*. Nov. 2004, pp. 327–330. DOI: 10.1109/ICDM.2004.10003.

[38] M. Berges, E. Goldman, H. S. Matthews, L. Soibelman, and K. Anderson. "User-Centered Nonintrusive Electricity Load Monitoring for Residential Buildings." In: *Journal of Computing in Civil Engineering* 25.6 (2011), pp. 471–480.

[39] Y. Jin, E. Telebakemi, and M. Berges. "A Time-Frequency Approach for Event Detection in Non-Intrusive Load Monitoring." In: *Proc. of SPIE Vol.* Vol. 8050. 2013, 80501U–1.

[40] S. B. Leeb, S. R. Shaw, and J. L. Kirtley. "Transient Event Detection in Spectral Envelope Estimates for Nonintrusive Load Monitoring." In: *IEEE Transactions on Power Delivery* 10.3 (July 1995), pp. 1200–1210. DOI: 10.1109/61.400897.

[41] S. R. Shaw, S. B. Leeb, L. K. Norford, and R. W. Cox. "Nonintrusive Load Monitoring and Diagnostics in Power Systems." In: *IEEE Transactions on Instrumentation and Measurement* 57.7 (July 2008), pp. 1445–1454. ISSN: 0018-9456. DOI: 10.1109/TIM.2008.917179.

[42] Y. Jin, E. Tebekaemi, M. Berges, and L. Soibelman. "A Time-Frequency Approach for Event Detection in Non-Intrusive Load Monitoring." In: vol. 8050. 2011. DOI: 10.1117/12.884385.

[43] K. S. Barsim, R. Streubel, and B. Yang. "Unsupervised adaptive event detection for building-level energy disaggregation." In: *Proceedings of power and energy student summt (PESS), Stuttgart, Germany* (2014).

[44] B. Wild, K. S. Barsim, and B. Yang. "A new unsupervised event detector for non-intrusive load monitoring." In: *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. Dec. 2015, pp. 73–77. DOI: 10.1109/GlobalSIP.2015.7418159.

[45] S. Houidi, F. Auger, H. B. A. Sethom, et al. "Statistical assessment of abrupt change detectors for non-intrusive load monitoring." In: *2018 IEEE International Conference on Industrial Technology (ICIT)* (2018), pp. 1314–1319.

[46] H. Berriri, M. W. Naouar, and I. Slama-Belkhodja. "Easy and Fast Sensor Fault Detection and Isolation Algorithm for Electrical Drives." In: *IEEE Transactions on Power Electronics* 27.2 (Feb. 2012), pp. 490–499. ISSN: 0885-8993. DOI: 10.1109/TPEL.2011.2140333.

[47] K. N. Trung, E. Dekneuvel, B. Nicolle, et al. "Event detection and disaggregation algorithms for nialm system." In: *the 2nd International Non-Intrusive Load Monitoring (NILM) Workshop*. 2014.

[48] J. Ajmera, I. McCowan, and H. Bourlard. "Robust speaker change detection." In: *IEEE Signal Processing Letters* 11.8 (Aug. 2004), pp. 649–651. ISSN: 1070-9908. DOI: 10.1109/LSP.2004.831666.

[49] S. B. Leeb and J. L. Kirtley. "A multiscale transient event detector for nonintrusive load monitoring." In: *Industrial Electronics, Control, and Instrumentation, 1993. Proceedings of the IECON '93., International Conference on*. 1993, 354–359 vol.1. DOI: 10.1109/IECON.1993.339053.

[50] R. Cox, S. B. Leeb, S. R. Shaw, and L. K. Norford. "Transient event detection for nonintrusive load monitoring and demand side management using voltage distortion." In: *Twenty-First Annual IEEE Applied Power Electronics Conference and Exposition, 2006. APEC '06.* 2006, p. 7. DOI: 10.1109/APEC.2006.1620777.

[51] Y. Jin, E. Tebekaemi, M. Berges, and L. Soibelman. "Robust adaptive event detection in non-intrusive load monitoring for energy aware smart facilities." In: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. May 2011, pp. 4340–4343. DOI: 10.1109/ICASSP.2011.5947314.

[52] J. Roos, I. Lane, E. Botha, and G. P. Hancke. "Using Neural Networks for Non-intrusive Monitoring of Industrial Electrical Loads." In: *Instrumentation and Measurement Technolgy Conference*. IEEE. IEEE, 1994, pp. 1115–1118. DOI: 10.1109/imtc.1994.351862.

[53] S. N. Patel, T. Robertson, J. A. Kientz, M. S. Reynolds, and G. D. Abowd. "At the Flick of a Switch: Detecting and Classifying Unique Electrical Events on the Residential Power Line." In: *Lecture Notes in Computer Science* 4717 (2007), pp. 271–288.

[54] M. Berges, E. Goldman, H. S. Matthews, and L. Soibelman. "Learning Systems for Electric Consumption of Buildings." In: *ASCI international workshop on computing in civil engineering*. Vol. 38. American Society of Civil Engineers (ASCE), June 2009. DOI: 10.1061/41052(346)1.

[55] K. Ting, M. Lucente, G. S. Fung, W. Lee, and S. Hui. "A Taxonomy of Load Signatures for Single-Phase Electric Appliances." In: *IEEE PESC (Power Electronics Specialist Conference)*. IEEE, 2005, pp. 12–18.

[56] H. Y. Lam, G. Fung, and W. Lee. "A Novel Method to Construct Taxonomy Electrical Appliances Based on Load Signatures." In: *Consumer Electronics, IEEE Transactions on* 53.2 (May 2007), pp. 653–660. DOI: 10.1109/tce.2007.381742.

[57] H.-T. Yang, H.-H. Chang, and C.-L. Lin. "Design a Neural Network for Features Selection in Non-intrusive Monitoring of Industrial Electrical Loads." In: *11th International Conference on Computer Supported Cooperative Work in Design*. IEEE, 2007, pp. 1022–1027.

[58] Y. H. Lin, M. S. Tsai, and C. S. Chen. "Applications of Fuzzy Classification with Fuzzy CMeans Clustering and Optimization Strategies for Load Identification in NILM Systems." In: *2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011)*. Institute of Electrical and Electronics Engineers (IEEE), June 2011, pp. 859–866. DOI: 10.1109/FUZZY.2011.6007393.

[59] S. Gupta, M. S. Reynolds, and S. N. Patel. "ElectriSense: Single-Point Sensing Using EMI for Electrical Event Detection and Classification in the Home." In: *Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM. ACM Press, 2010, pp. 139–148. DOI: 10.1145/1864349.1864375.

[60] N. Batra, J. Kelly, O. Parson, et al. "NILMTK: An Open Source Toolkit for Non-intrusive Load Monitoring." In: *Proceedings of the 5th International Conference on Future Energy Systems*. New York, NY, USA: ACM, 2014, pp. 265–276. DOI: 10.1145/2602044.2602051.

[61] J. Froehlich, E. Larson, S. Gupta, et al. "Disaggregated End-Use Energy Sensing for the Smart Grid." In: *IEEE Pervasive Computing* 10.1 (Jan. 2011), pp. 28–39. DOI: 10.1109/mprv.2010.74.

[62] J. Gao, E. C. Kara, S. Giri, M. Berg, et al. "A feasibility study of automated plug-load identification from high-frequency measurements." In: *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE. Institute of Electrical and Electronics Engineers (IEEE), Dec. 2015, pp. 220–224. DOI: 10.1109/globalsip.2015.7418189.

[63] J. Z. Kolter and T. Jaakkola. "Approximate Inference in Additive Factorial HMMs with Application to Energy Disaggregation." In: *Proceedings of the 15. International Conference on Artificial Intelligence and Statistics*. Vol. 22. Proceedings of Machine Learning Research. PMLR, Apr. 2012, pp. 1472–1482.

[64] M. Zhong, N. Goddard, and C. Sutton. "Signal Aggregate Constraints in Additive Factorial HMMs, with Application to Energy Disaggregation." In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger. Curran Associates, Inc., 2014, pp. 3590–3598.

[65] O. Kramer, T. Klingenberg, M. Sonnenschein, and O. Wilken. "Non-intrusive appliance load monitoring with bagging classifiers." In: *Logic Journal of the IGPL* 23.3 (2015), pp. 359–368. DOI: 10.1093/jigpal/jzv016. eprint: /oup/backfile/content_public/journal/jigpal/23/3/10.1093/jigpal/jzv016/2/jzv016.pdf.

[66] L. Du, Y. Yang, D. He, et al. "Support vector machine based methods for non-intrusive identification of miscellaneous electric loads." In: *IECON 2012 - 38th Annual Conference on IEEE Industrial Electronics Society*. Oct. 2012, pp. 4866–4871. DOI: 10.1109/IECON.2012.6389580.

[67] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Imagenet classification with deep convolutional neural networks." In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[68] J. Lee, J. Park, K. L. Kim, and J. Nam. "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms." In: *arXiv preprint arXiv:1703.01789* (2017).

[69] W. Dai, C. Dai, S. Qu, J. Li, and S. Das. "Very deep convolutional neural networks for raw waveforms." In: *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE. 2017, pp. 421–425.

[70] D. Jorde, T. Kriechbaumer, and H. Jacobsen. "Electrical Appliance Classification using Deep Convolutional Neural Networks on High Frequency Current Measurements." In: *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*. 2018. DOI: 10.1109/SmartGridComm.2018.8587452.

[71] J. Kelly and W. Knottenbelt. "Neural NILM: Deep Neural Networks Applied to Energy Disaggregation." In: *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*. ACM. Association for Computing Machinery (ACM), 2015, pp. 55–64. DOI: 10.1145/2821650.2821672.

[72] N. Iksan, J. Sembiring, N. Haryanto, and S. H. Supangkat. "Appliances identification method of non-intrusive load monitoring based on load signature of VI trajectory." In: *International Conference on Information Technology Systems and Innovation (ICITSI)*. IEEE. 2015, pp. 1–6.

[73] J. Kelly and W. Knottenbelt. "Metadata for Energy Disaggregation." In: *Computer Software and Applications Conference Workshops, 2014 IEEE 38th International*. IEEE. 2014, pp. 578–583.

[74] F. Englert, T. Schmitt, S. Kößler, A. Reinhardt, and R. Steinmetz. "How to Auto-Configure Your Smart Home? High-Resolution Power Measurements to the Rescue." In: *Proceedings of the fourth international conference on Future energy systems*. ACM. Association for Computing Machinery (ACM), 2013, pp. 215–224. DOI: 10.1145/2487166.2487191.

[75] K. N. Trung, O. Zammit, E. Dekneuvel, et al. "An Innovative Non-Intrusive Load Monitoring System for Commercial and Industrial Application." In: *International Conference onAdvanced Technologies for Communications (ATC)*. IEEE. Institute of Electrical and Electronics Engineers (IEEE), Oct. 2012, pp. 23–27. DOI: 10.1109/atc.2012.6404221.

[76] G. Peeters. *A large set of audio features for sound description (similarity and classification) in the CUIDADO project*. Tech. Rep. IRCAM, 2004.

[77] C.-C. Chang and C.-J. Lin. "LIBSVM: A Library for Support Vector Machines." In: *ACM Transactions on Intelligent Systems and Technology* 2.3 (3 Apr. 2011). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, pp. 271–2727. DOI: 10.1145/1961189.1961199.

[78] F. Sultanem. "USING APPLIANCE SIGNATURES FOR MONITORING RESIDENTIAL LOADS AT METER PANEL LEVEL." In: *Power Delivery, IEEE Transactions on* 6.4 (1991), pp. 1380–1385. DOI: 10.1109/61.97667.

[79] I. Daubechies et al. *Ten Lectures on Wavelets*. Vol. 61. SIAM, Jan. 1992. DOI: 10.1137/1.9781611970104.

[80] R. Patzelt and H. Schweinzer. *Elektrische Meßtechnik*. Springer-Verlag, 2013. DOI: 10.1007/978-3-7091-6557-7.

[81] T. J. Rouphael. *RF and digital signal processing for software-defined radio: a multi-standard multi-mode approach*. Newnes, 2009.

[82] D. Shmilovitz. "On the Definition of Total Harmonic Distortion and Its Effect on Measurement Interpretation." In: *IEEE Transactions on Power Delivery* 20.1 (Jan. 2005), pp. 526–528. DOI: 10.1109/tpwrd.2004.839744.

[83] MPEG. *MPEG-7 Audio Descriptions*. 2005. URL: http://mpeg.chiariglione.org/standards/mpeg-7/audio (visited on 09/07/2016).

[84] G. Hughes. "On the Mean Accuracy of Statistical Pattern Recognizers." In: *IEEE Transactions on Information Theory* 14.1 (Jan. 1968), pp. 55–63. ISSN: 0018-9448. DOI: 10.1109/TIT.1968.1054102.

[85] T. A. Runkler. *Data Analytics*. Springer, 2012. ISBN: 978-3-658-14074-8.

[86] A. C. Ian Goodfellow Yoshua Bengio. "Deep Learning." Book in preparation for MIT Press. 2016. URL: http://www.deeplearningbook.org.

[87] D. P. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization." In: *CoRR* abs/1412.6980 (2014). arXiv: 1412.6980. URL: http://arxiv.org/abs/1412.6980.

[88] A. Torralba and A. A. Efros. "Unbiased Look at Dataset Bias." In: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE. 2011, pp. 1521–1528.

[89] U. Kuckartz, S. Rädiker, and A. Rheingans-Heintze. "Umweltbewusstsein in Deutschland 2006 - Ergebnisse einer repräsentativen Bevölkerungsumfrage, im Auftrag des Bundesministeriums für Umwelt." In: *Naturschutz und Reaktorsicherheit (BMU)* (2006).