

**University of Alberta**

Decomposition Techniques for Non-intrusive Home Appliance Load  
Monitoring

by

Seyed Mostafa Tabatabaei

A thesis submitted to the Faculty of Graduate Studies and Research  
in partial fulfillment of the requirements for the degree of

Master of Science

in

Energy Systems

Electrical and Computer Engineering

©Seyed Mostafa Tabatabaei

Spring 2014

Edmonton, Alberta

Permission is hereby granted to the University of Alberta Libraries to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only. Where the thesis is converted to, or otherwise made available in digital form, the University of Alberta will advise potential users of the thesis of these terms.

The author reserves all other publication and other rights in association with the copyright in the thesis and, except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatsoever without the author's prior written permission.

## **Abstract**

Energy-saving is a key element of Smart Grid. By encouraging consumers to moderate their energy demands, utilities can make more efficient use of their generation assets, and reduce total fuel consumption. For this purpose, we must provide homeowners with appliance energy consumption data, without requiring sensors on each appliance. This means that energy consumption from the house main feeder must be disaggregated into individual appliances.

In this thesis, two novel methodologies for disaggregating household power consumption are evaluated. The first method is multi-label classification, which is used to predict appliance participation in the power signal. The second method is a new signature-based sequence matching algorithm. Two sets of features have been used. In the time domain, a delay embedding of the observed power signal is constructed. The second feature set is a wavelet decomposition of the power signal, using Haar wavelet. We evaluate our techniques and features on two synthetic datasets, and two households from REDD.

# Contents

<b>Chapter 1 Introduction.....</b>	<b>1</b>
<b>Chapter 2 Literature Review .....</b>	<b>4</b>
2.1 A Taxonomy of NILM Research .....	6
2.2 NILM as a Multi-Label Classification Problem.....	13
<b>Chapter 3 Background .....</b>	<b>17</b>
3.1 Delay Coordinate Embedding.....	17
3.1.1 Mutual Information.....	21
3.1.2 False Nearest Neighbor.....	25
3.2 Discrete Wavelet Transform .....	27
3.3 Multi-label Classification.....	32
3.3.1 RAKEL .....	32
3.3.2 ML $k$ NN.....	33
3.4 Clustering.....	34
3.4.1 EM Clustering.....	35
3.5 Similarity Search.....	37
3.6 Dynamic Time Warping .....	38
<b>Chapter 4 Methodology .....</b>	<b>42</b>
4.1 Datasets.....	42
4.1.1 Simulated Datasets.....	42
4.1.2 REDD: The Reference Energy Disaggregation Dataset .....	43
4.2 Delay Embedding.....	43
4.2.1 Delay embedding parameter selection .....	45
4.3 Discrete Wavelet Feature extraction.....	51
4.4 Multi-label classification .....	54
4.4.1 Evaluation of Multi-label Classification .....	55
4.5 Clustering.....	56
4.6 Similarity Search.....	58
4.6.1 Dimensionality Reduction using Haar Wavelets .....	61
4.6.2 Edge Detection.....	61
4.6.3 Coefficients sequence Segmentation .....	64
4.6.4 Segment Modification.....	65
4.6.5 Database .....	68
4.6.6 Similarity measurement .....	69
4.6.7 Decision Making.....	70
4.6.8 Iterative Detection.....	71
4.7 Energy Error.....	73
<b>Chapter 5 Experimental Results.....</b>	<b>75</b>
5.1 Evaluation on Simulated Datasets.....	75
5.1.1 Dataset 1.....	75
5.1.2 Dataset 2.....	76
5.2 Evaluation on REDD .....	78
5.2.1 REDD, House 3 .....	79
5.2.2 REDD, House 1 .....	80
5.3 Clustering based Classification.....	82
5.4 Evaluation of the proposed methods on each appliance .....	88

5.4.1	Simulated Datasets.....	88
5.4.2	REDD, House 3 .....	90
5.4.3	REDD, House 1 .....	93
5.5	Energy Error.....	95
5.6	Comparison with Published Methods on REDD .....	98
<b>Chapter 6 Conclusion .....</b>		<b>102</b>
<b>Chapter 7 References.....</b>		<b>103</b>

## List of Tables

Table 1 Sample appliance waveform .....	16
Table 2 Preserved energy on Haar wavelet decomposition levels in simulated dataset ...	52
Table 3 Preserved energy on Haar wavelet decomposition level in REDD, house 3 .....	53
Table 4 Preserved energy on Haar wavelet decomposition level in REDD, house 1 .....	54
Table 5 Different Scenarios for single state appliances mixture and how to extract features .....	67
Table 6 Mixture of single appliance with pulsive appliances .....	72
Table 7 Evaluation result of multi-label classification in time domain on dataset 1 .....	75
Table 8 Evaluation results of multi-label classification in wavelet domain on dataset 1,	76
Table 9 Evaluation results of Similarity Search on dataset 1.....	76
Table 10 Evaluation of multi-label classification in time domain on dataset 2 .....	77
Table 11 Evaluation results of multi-label classification in wavelet domain on dataset 2, .....	77
Table 12 Evaluation results of Similarity Search on dataset 2,.....	78
Table 13 Evaluation results of Multi-label classification in Time domain on REDD, house 3 .....	79
Table 14 Evaluation results of Multi-label classification in Wavelet domain on REDD, house 3 .....	79
Table 15 Evaluation results of Similarity Search on REDD House 3.....	80
Table 16 Evaluation results of Multi-label classification in Time domain on REDD House 1 .....	81
Table 17 Evaluation results of Multi-label classification in Wavelet on REDD, House	181
Table 18 Evaluation results of Similarity Search on REDD, House 1.....	81
Table 19 Evaluation results of multi-label classification on dataset 1.....	83
Table 20 Evaluation results of multi-label classification on dataset 2.....	84
Table 21 Evaluation results of classification method on House 3, REDD .....	85
Table 22 Evaluation multi-label classification along with clustering on REDD, House	187
Table 23 Performance of identification methods for detecting Refrigerator in Dataset 1	88
Table 24 Performance of identification methods for detecting Microwave in Dataset 1 .	89
Table 25 Performance of identification methods for detecting Refrigerator in Dataset 2	89

Table 26 Performance of identification methods for detecting Microwave in Dataset 2 .	90
Table 27 Multi label classification performance on REDD, House 3 in Time Domain ...	91
Table 28 Multi label classification performance on REDD, House 3 in Wavelet domain	91
Table 29 DTW Similarity search result on REDD, House 3 .....	92
Table 30 Multi label classification performance on REDD House 1 in time domain .....	94
Table 31 Multi label classification performance on REDD House 1 in Wavelet domain	94
Table 32 Similarity Search Performance on House 1 .....	95
Table 33 Energy error on Dataset 1 .....	96
Table 34 Energy error on Dataset 2 .....	96
Table 35 Energy Error on REDD, House 3 .....	97
Table 36 Energy Error on REDD, House 1 .....	97
Table 37 REDD, house 3 comparisons .....	99
Table 38 REDD, house 1 comparison.....	100
Table 39 Comparison of the energy error of proposed method with methods in [64] and [63].....	101

## List of Figures

Figure 1 Normalized P-Q signature space [7].....	4
Figure 2 Different type of appliances [8].....	5
Figure 3 Nonintrusive load monitoring techniques taxonomy.....	8
Figure 4 Sample power waveform demand measured from a real home.....	15
Figure 5 State space trajectory in a 2-D plane. This figure shows a sample case where future point on a trajectory is predicted with delay coordinate embedding for two different points $n$ and $m$ using a two-dimensional delay vector with time delay $\tau$ .....	19
Figure 6 Example of Dynamic Time Warping.....	40
Figure 7 Home Appliance Waveform, Refrigerator features are highlighted.....	44
Figure 8 Mutual Information, Dataset 1, $\tau=35s$ .....	45
Figure 9 Mutual Information, Dataset 2, $\tau=1s$ .....	46
Figure 10 False Nearest Neighbor ratio, Dataset 1, $m=16$ .....	47
Figure 11 False Nearest Neighbor Dataset 2, $m=8$ .....	47
Figure 12 Mutual information, House 3, $\tau=95s$ .....	49
Figure 13 False nearest neighbor ratio, House 3, $m=18$ .....	49
Figure 14 Mutual information, House 1, $\tau=32s$ .....	50
Figure 15 False nearest neighbor ratio, House 1, $m=8$ .....	50
Figure 16 proposed Multi-label classification method along with clustering.....	58
Figure 17 Similarity search method procedure .....	60
Figure 18 Wavelet effect on shape of signal, (a) One day REDD power signal before transformation; (b) level 3 Haar coefficients of (a); .....	61
Figure 19 Sample Approximate coefficients and related detail coefficients which determine events in waveform (a) Approximation coefficients sequence (b) Detail Coefficients which have value at edge locations .....	63
Figure 20 Sequence matching procedure .....	70

## Chapter 1

### INTRODUCTION

Currently, residential houses consume approximately 32 percent of electrical energy in Canada [1]. This means that, if the proposed Smart Grid initiative is to achieve its goal of reducing the growth in fuel consumption and emissions (with respect to the current trend-line), efficiencies will have to be found in the residential market. In a market economy, this means providing a price signal to consumers that encourages them to reduce or time-shift their power consumption during peak load periods (when energy is the most expensive). However, the current power metering and billing infrastructure is simply inadequate for this problem. Currently, the monthly bill shows only the total energy consumption of a home, and provides no insight into the time-of-use cost of operating the individual appliances in a home. Such detailed feedback would allow the consumer to plan their energy usage in order to reduce their monthly bill, while not suffering an unacceptable disruption to their daily lives. The literature indicates that such feedback can lead to a 10-15 percent saving in energy costs [2], or a savings of about \$1.6 billion each year in the Canadian economy.

Power utilities and consumers share an interest in managing electricity loads and costs. From the consumer's side, a fully transparent bill that allows them to examine both historical and real-time costs for using each individual



appliance would be a powerful tool that encourages them to reduce or time-shift their largest electrical loads, saving them money. (The *provision* of effective, easy-to-use tools supporting a behavior is known to promote the adoption of that behavior [3]. From the utilities' side, formulating demand side management and demand response strategies such as changing the time-of use price schedule or load shedding requires detailed information about the mix of appliances in operation [4]. Load component details are used for both short-term and long-term load prediction, and help to determine when conventional and renewable generation asserts must be added to the power grid [5].

While some modern appliances are equipped to communicate with utilities (by receiving and displaying a signal when electricity prices are high, responding to a load-shedding command, etc.) older ones have no such capabilities. Thus, the electricity consumption for an appliance must be inferred by monitoring the appliance's instantaneous energy use, and relating it to a time signal (thus obtaining the load profile and time-of-use cost). Traditionally, this monitoring would be accomplished by placing a sensor in the electrical circuit servicing the appliance. This is an expensive and invasive process, which is unlikely to achieve widespread adoption in North America due to privacy concerns. The alternative is for a utility to monitor the electricity load for a home from the public distribution box, and disaggregate this signal into the individual appliance loads. This well-known problem is referred to as Non-Intrusive Load Monitoring (NILM). Current approaches to NILM may be divided into signature-based and inductive learning techniques. The former require a database of appliance power signals, from which

a signature is extracted for each appliance; the latter do not require such a database, but offer very limited accuracy. What is needed for practical NILM is a technique that requires little to no database assembly, and is highly accurate.

Two novel load decomposition techniques for NILM are proposed in this thesis: two different multi-label classification methods (both with and without an initial clustering step), and a waveform matching method which uses dynamic time warping. These approaches are furthermore evaluated in both the time and wavelet domains. All of these combinations are evaluated on two simulated datasets with different scenarios, and two real homes drawn from the Reference Energy Disaggregation Dataset (REDD) [6].

The remainder of this thesis is organized as follows. In Chapter 2, we review the existing literature on the NILM problem, and offer a new taxonomy of the field. We then review essential background on feature extraction and data mining in time series and signals in Chapter 3. In Chapter 4, we discuss our experimental methodology, and we present our experimental results in Chapter 5. We offer a summary and discussion of future work in Chapter 6.

## Chapter 2

### LITERATURE REVIEW

The technique of non-intrusive load monitoring was first developed by G. Hart in the 1980s [7]. Hart treated the NILM problem as a communication system in which the appliances are transmitters and their signatures are codes; the goal of NILM is then to design a decoder for these messages. In Hart's original work, the signatures were the height of rising and falling edges in the power waveform. He proposed both a supervised classification method, as well as a clustering method in the P-Q plane.

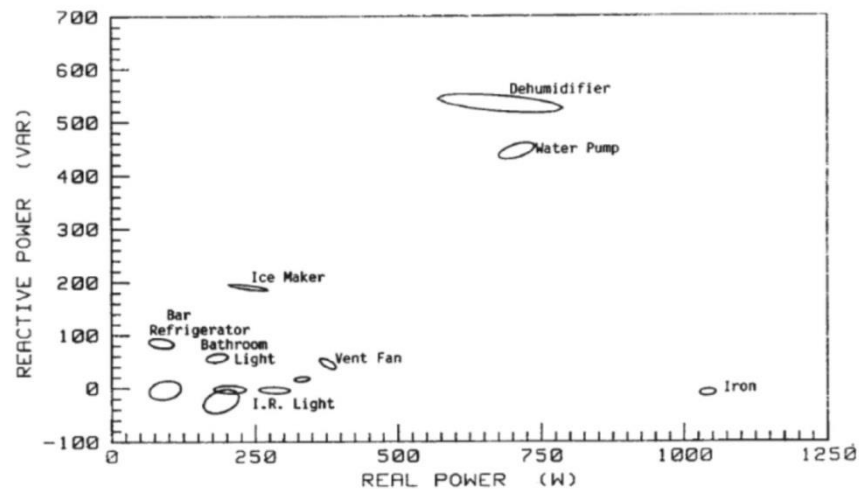


Figure 1 Normalized P-Q signature space [7]

Home appliances are categorized into three groups [7]: single state appliances (ON/OFF), multi state (Finite State Machine), and continuously varying. Figure 2 shows three types of appliances models [8].

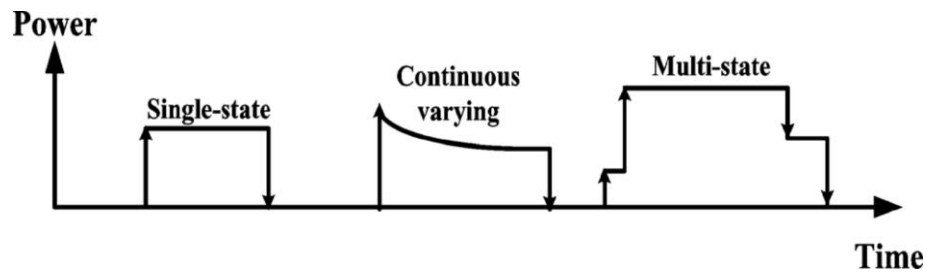


Figure 2 Different type of appliances [8]

Single state appliances have a pair of identical ON/OFF edges and a constant power level. Most appliances, such as a kettle, toaster, and microwave are single state. Multi-state appliances have a set of discrete states and edges. Many heavy appliances including dryer, washing machines and furnace are multi state. Multi-state appliances have a unique sequence of operation, creating a set of changes in the power waveform. For example a clothes washer, in a fixed pattern, follows the following operating modes: water-fill, immerse, rinse, drainage, and spin-dry [8]. A stove is a typical multistate appliance with repetitive pulses (i.e. repeated duty cycles) to avoid overheating [9]. A furnace is another multi-state appliance which has a fixed sequence heating pattern. In a cycle, a pattern such as 0, 230 W, 560 W, 340 W, 0 is observed. Furnace heating cycles may change according to the environment temperature.

Continuously varying appliances usually have a pair of different ON and OFF edges, and power consumption between the edges gradually varies. Refrigerators and freezers are continuous varying appliances [7]. In continuous varying appliances there is a power demand variation between the rising and falling edges.

After Hart's seminal paper, numerous investigations have attempted to improve upon his results, and NILM is now accepted as an important facet of Smart Grid technology. The main difference between published NILM methods is the machine learning algorithms and features that they have used for appliance identification. Classification methods such as Support vector machine (SVM),  $k$ -nearest neighbor ( $k$ -NN), and clustering methods such as  $k$ -means are commonly applied for NILM. Active power, reactive power, harmonics, current and voltage transients, duty cycles, and/or combinations thereof are commonly used as features.

## **2.1 A Taxonomy of NILM Research**

There is now a significant literature on NILM, using a variety of approaches. In order to organize this review, we have developed a new taxonomy of the field of NILM research. Our review defines the "NILM field" as the set of papers citing Hart's seminal work, and which propose and evaluate a new NILM technique. We classify articles by the machine learning algorithm and feature extraction technique employed. We found that these two categorizations were the most useful in distinguishing between different strands of NILM research.

Machine learning algorithms are either supervised or unsupervised. Supervised learning methods require a known set of input and output data for each class label to build a prediction model for new unknown data. Thus, supervised NILM methods require initial appliance features to train the algorithm.

This initial information is collected by recording appliance features with the help of the customer [8] or installing extra sub-meters inside the home [60].

In contrast, unsupervised learning algorithms discover regularities in measured data, and groups data points based on their common properties (proximity in feature space). The main benefit of unsupervised methods is that they do not need the same initial information for training; however, the accuracy of unsupervised methods is generally lower than supervised methods. In some investigations, clustering has been used to extract appliance features to build a database of appliances for classification purpose [52, 53, 66, 67]. Clustering has also been used to detect appliances from observations. Clustering distinguish between all individual appliances because their features are similar for some appliances; therefore in some applications they have been used to detect groups of similar appliances instead of a single appliance [57]. Statistical methods such as Hidden Markov models are also unsupervised methods [60, 62-64, 68].

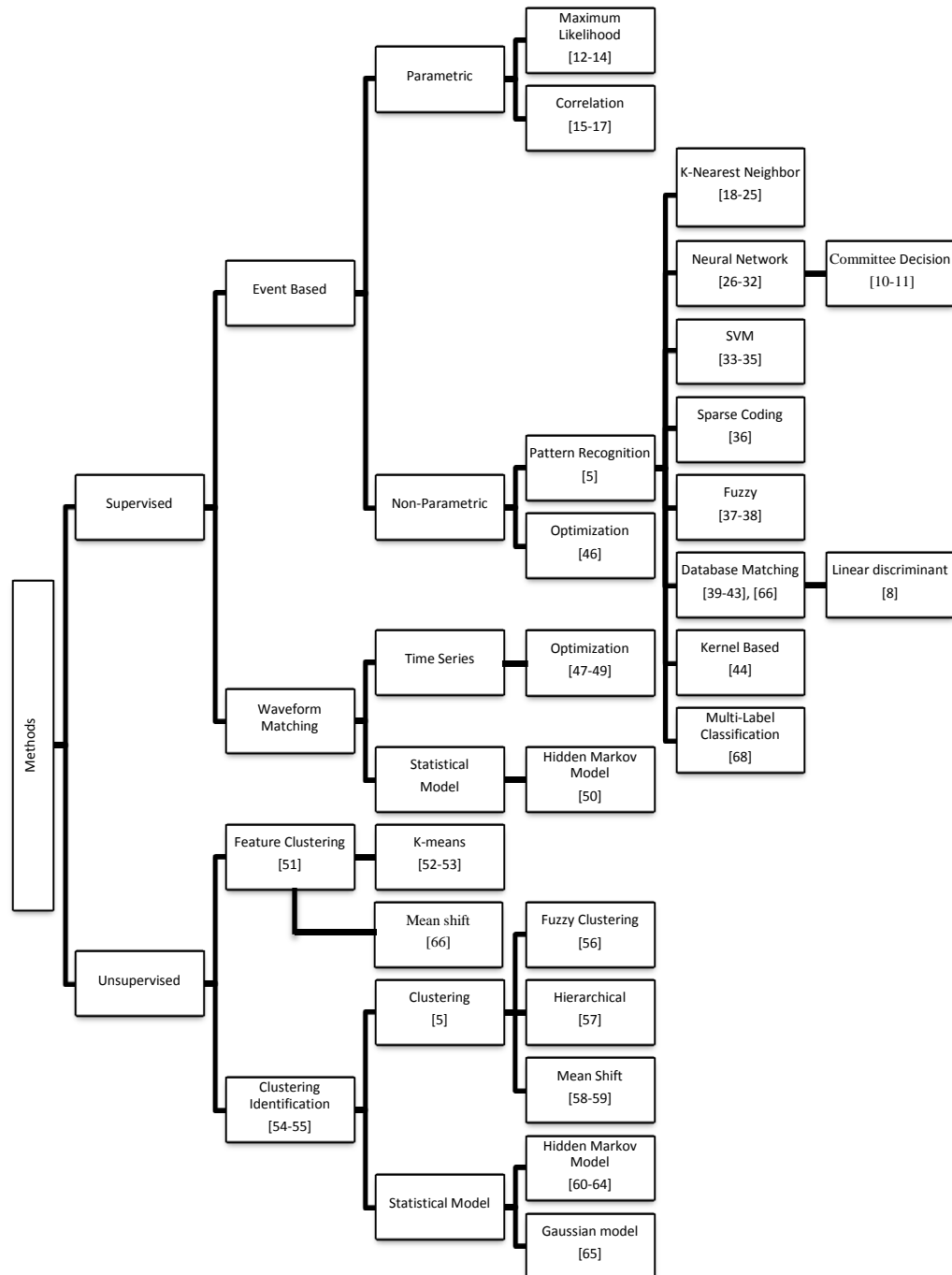


Figure 3 Nonintrusive load monitoring techniques taxonomy

Appliance features are extracted in two ways: Event based or waveform based. If features are just recorded when significant variation is happening in the signal, the collected features are called event based. Changes in active and reactive power magnitude [8] and transients in current and voltage [40] are good examples of event-based features. Most published NILM research employs event based features. The general framework of these investigations is to detect changes in measured signals and then identify the source of the event through a machine learning algorithm [8].

Machine learning algorithms are grouped into parametric and non-parametric models. A parametric model assumes that the data has a type of probability distribution with a finite number of parameters [14-17], and “fitting” the model consists of determining the parameter values for which the model best matches the training data. Non-parametric models, on the other hand, do not assume a probability distribution a priori; the model is instead induced from the data. Most pattern recognition and classification methods (e.g. k-nearest neighbor (k-NN) [18-25], neural networks [26-32], support vector machines (SVM) [33-35] and database matching algorithm [39-43]) are non-parametric.

Waveform-based features essentially treat the power waveform as a time series; the NILM problem is then a particular time-series classification problem. Time series classification algorithms can be categorized into three groups as follows:

- Distance-based classification
- Feature-based classification



- Model-based classification

Distance and feature-based classification both work with the actual time series data [47-49]. Feature based methods such as [8] split the signal into small sequences or windows and then extract features to classify each sequence, while distance based methods use raw sequences directly. Model based methods, on the other hand, convert time-series data into statistical and probabilistic models such as Hidden Markov Models [60, 62-64, 68]. Using Hidden Markov Model (HMM), data sequence is modeled as a Markov chain; HMM learns probability of transition from one state to another state in the Markov chain. Each state represents one or mixture of appliances. Having new test sequence, a sequence of states is predicted using Viterbi algorithm in order to have most similarity between reconstructed signal using predicted states and the test signal. Knowing the corresponding state of each data point, they will get labels which means identifying the appliances.

Furthermore, within all three groups, there are two possible approaches for developing a classification algorithm. The first approach is to design a whole new algorithm that works with a raw time series, usually by creating a new distance measure for sequential data. The other approach is to design a transformation that turns sequential data into a set of feature vectors (e.g. lagged inputs). These feature vectors can then be passed to any standard machine-learning algorithm.

Conventional classification algorithms that work with sequential data usually require a distance metric between two sequences. The selection of a distance (similarity) measure plays a significant role in the quality of the

classification algorithm [69]. Euclidean distance is a widely adopted measurement; it requires the two series in comparison to be of equal length [70, 71]. In addition, it is sensitive to distortions in time. Distortion in the time axis is common in applications such as speech recognition where speech rates are not constant [72]. Similar problems have been noted in applications such as web logs and biomedical data [73]. Some researchers have tried to overcome the time distortion by pre-processing the acquired signal, however such approaches are not practical in most cases [74]. Thus, elastic similarity measures such as Dynamic Time Warping (DTW) have been employed. [71] describes DTW as a non-linear mapping between two unequal sequences where the distance between them is the minimum one among possible distances. Although many researchers [75, 76] agree that DTW solves many of the problems of the Euclidean distance, its computational inefficiency limits its adoption [77]. DTW is calculated using dynamic programming, hence has a quadratic time complexity ( $O(n*m)$ ) where  $m$  and  $n$  are length of two sequences. In a similarity search task (e.g. case based reasoning or signature matching), a new unknown sequence or section of a longer time series called a query sequence is compared with existing archive sequences in the database in order to find the most similar sequence.

An NILM technique based on Dynamic Time Warping (DTW) is proposed in [78]. Power measurements are processed to extract two features, namely total energy consumption and the number of rising edges in overlapped windows. Measured features for each window are compared to the reference set in order to find out the identity of the connected loads. Evaluation results for the proposed

method in [78] are not published.

Like other signals, a power waveform can also be transformed to the frequency domain. Any signal can be expressed as the sum of a (possibly infinite) collection of sinusoidal functions using the Fourier transform. The advantage of frequency-domain analysis is that the coefficients of the Fourier transform (the weights of each frequency in the summation, which are treated as features) often expose important relationships in the data, which lead to a superior classification result. While this information was also present in the time domain, it would have been implicit and difficult for machine learning algorithms to detect. The main disadvantage of the Fourier transform is that it has only frequency resolution and no time resolution. This means that the Fourier transform allows us to determine all the frequencies present in a signal, but the temporal relationship between those frequencies is lost. To overcome this problem several solutions have been developed to represent a signal in the time and frequency domain at the same time. In time-frequency joint representations the idea is to cut the signal of interest into several parts and then analyze each part separately in order to have information about the temporal relationship between different frequency components [79]. Wavelet transforms are one of the best-known examples.

Wavelet transform for NILM Application is used to analyze transient signal which have high frequency resolution, it appears during state changes in some appliances. For example, discrete wavelet transform coefficients [28] or the energy of the coefficients [80] have been extracted as transient features from measured current waveforms. However, the high frequency transient features

extracted from the wavelet transform are not constant across repeated appliance use, so it has been used rarely. The application of wavelet for dimensionality reduction in NILM has not been previously explored; the existing literature in dimensionality reduction for NILM applies different methods such as PCA [81].

## 2.2 NILM as a Multi-Label Classification Problem

In machine learning, “classification” algorithms usually refers to single-label classification, in which a set of instances are each associated with a unique class label drawn from a set of discrete class labels  $L$ . The classification problem is termed binary when  $L$  contains two classes or multi-class when there are more than two classes. We can generalize the idea of classification by allowing an instance to have more than one label, giving us the category of multi-label classification algorithms. As with single-class algorithms, the goal of multi-label classification is learning to predict class labels from a set of instances where each instance could belong to one or more classes.

Multi-label classification was initially developed for automatic text categorization and medical diagnosis. However, a number of other prediction tasks can also be conveniently described as multi-label problems, drawing more research attention to this area [82]. For example, a text document that talks about scientific contributions in medical science can belong to both a science and a health category; genes may have multiple functionalities (i.e. be associated with multiple diseases) [83, 84]; an image that captures a field and fall colored

trees can belong to both field and fall foliage categories [85, 86]; a movie can simultaneously belong to action, crime, thriller, and drama categories [87]; an email message can be tagged as both work and research project [88, 89]. Clearly, traditional binary and multi-class problems both can be posed as specific cases of the multi-label problem [90].

A multi-label classification method has been used to identify some high-power appliances in [91]. Power consumption value at each sample instant, and changes in power consumption in sliding windows were extracted from the power waveform of a house, and the goal was to disaggregate three specific appliances from the power signal. In general, NILM can reasonably be considered a multi-label problem. At each sample instant, the power signal is always associated with a mixture of appliances; thus, if the class labels are the appliances active for a given sample instant, NILM is clearly a multi-label problem. Suppose there are  $n$  appliances inside the home,  $a(t)$  is a  $n$ -component label vector for each moment which describes the state of the  $i$ -th appliance at time  $t$  [7]:

$$a_i(t) = \begin{cases} 1 & \text{if appliance } i \text{ is ON at time } t \\ 0 & \text{if appliance } i \text{ is OFF at time } t \end{cases} \quad (1)$$

The label vector describes the power consumption of individual appliances. The relation between total electric load and its components at each moment is:

$$P(t) = \sum_{i=1}^n a_i P_i(t) \quad (2)$$

where  $P_i(t)$  is power consumption of appliance  $i$  at time  $t$  and  $P(t)$  is home total

power consumption at time  $t$ . For example, Figure 4 depicts a sample power signal with two appliances whose power waveforms are mixed at some points.

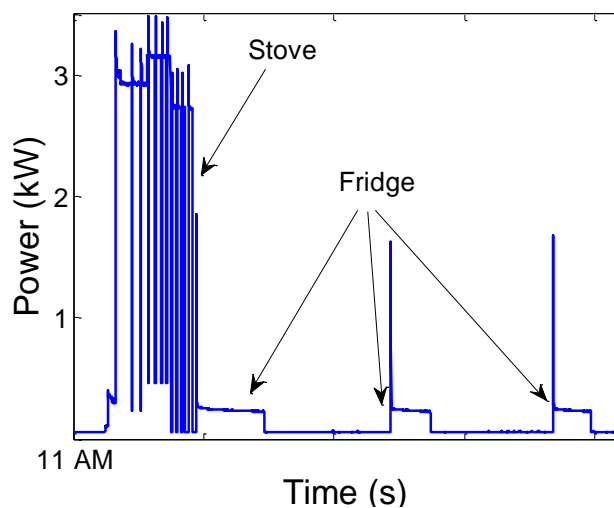
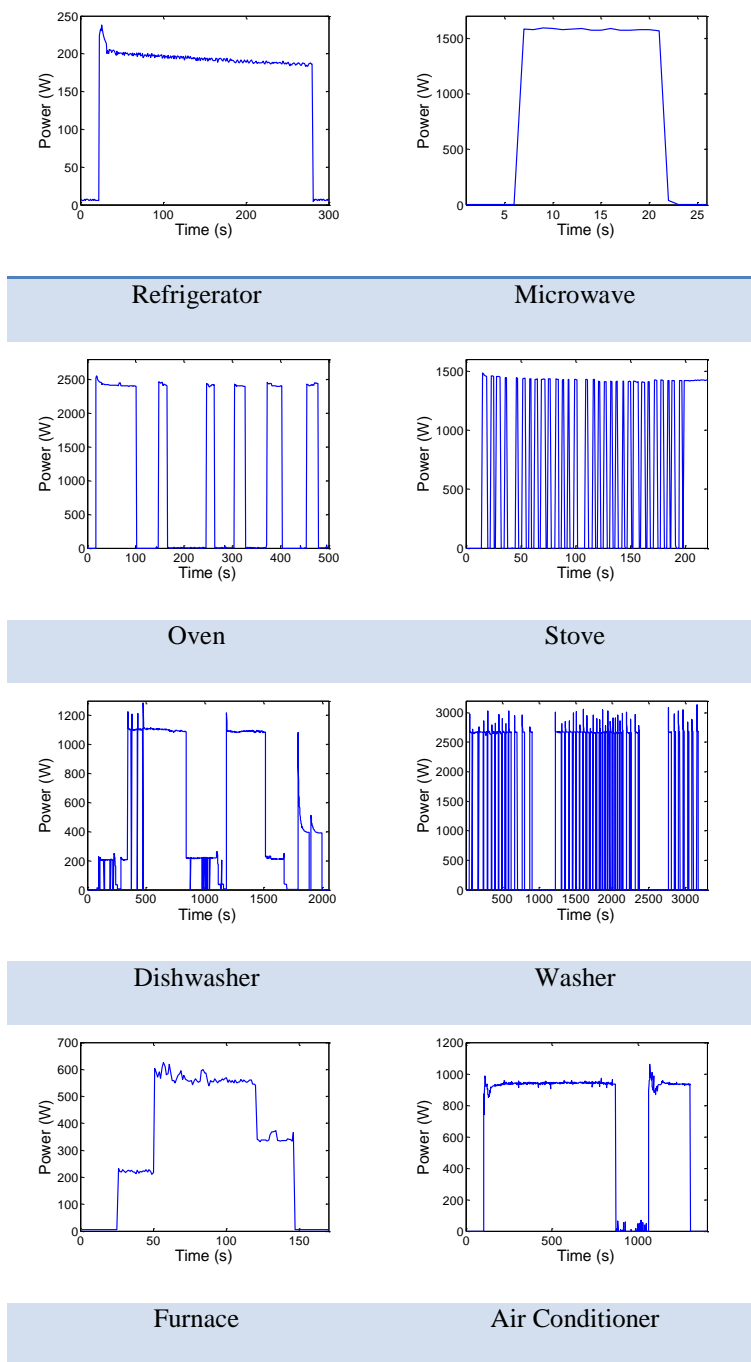


Figure 4 Sample power waveform demand measured from a real home

Each appliance has its own specific features. Levels of active and reactive power, duty cycle, harmonic frequencies and level, and the number of rising or falling edges are just some of the features that can be discriminative for an appliance [92]. However, as with all pattern recognition approaches, we cannot guarantee that appliances will be perfectly separated *a priori*; we must evaluate each dataset empirically. Appliance features are not fixed and they have some fluctuations, but most of the times any deviations are small and the overall shape of the power waveform is preserved. Power waveforms for some of the main appliances in REDD [6] are shown in Table 1.

Table 1 Sample appliance waveform



## Chapter 3

### **BACKGROUND**

In this chapter, we review key concepts in time-series and wavelet analysis, multi-label classification, clustering, and dynamic time warping.

#### **3.1 Delay Coordinate Embedding**

A dynamical system consists of a state space and rules to evolve from one state to another one. The state space describes the system at any given time and contains all the necessary information to predict the future evolution of the system [93]. For a deterministic system, knowing the current state completely determines the future states of the system. Assuming the system is dissipative, over time the system state will converge to a specific subset of states, known as an attractor (alternatively, the system can evolve away from that subset, in which case it is termed a repeller) [94].

In a time series, the state of the underlying system that generated the time series cannot be measured directly; even basic parameters such as the number of dimensions are unknown. A time series is only a sequence of scalar measurements; it is a projection of a  $d$ -dimensional dynamical system into a univariate sequence. In order to forecast the evolution of a time series, we need to reverse this projection (or at least find a mapping that is equivalent to doing so).



This process is called state space reconstruction, and one of the most widely used approaches is the delay coordinate embedding.

The fundamental concept of the delay state space reconstruction is to embed the univariate time series sequence into a multi-dimensional time-lagged state space with appropriate time delay  $\tau$  and embedding dimension  $m$ . In other words, we concatenate the current observation with a number of past observations into a vector (this is known as the lagged input representation). According to Takens' delay embedding theorem, if the number of lags is sufficient (the state vector large enough), then this delay vector is equivalent to the actual state vector of the system.

Takens' delay embedding theorem was published in 1981 [95]. The theorem claims that for a  $d$ -dimension dynamical system, almost every smooth function such as a delay-coordinate map with dimension  $m$  ( $: d \rightarrow m$ ) is one-to-one if  $m > 2d$  and if, for a sampling interval  $\tau$ , the system has no periodic orbits of period  $\tau$  or  $2\tau$ , and at most finitely many periodic orbits of period  $k\tau$  for  $k > 2$ . Therefore if the number of reconstructed dimensions is large enough, such model captures all the relevant dynamics and state space specifications[95]. In the Takens theory,  $d$  is the number of dimension of the phase space containing the attractor which can be much larger than the attractor dimension [93]. Authors in [94] generalised the theorem where reconstruction of state space just requires to satisfy the condition  $m > 2d_F$  where  $d_F$  is the dimension of attractor in phase space.

Figure 5 shows a simple example of a 2-dimensional delay embedding of a univariate time series. Since we assume that the underlying system is

deterministic, if we know the current delay vector, we should be able to predict the next vector on the delay-space trajectory, thereby forecasting the evolution of the time series. One of the important implications of this is that the independent delay vectors each contain sufficient information to predict the next observation(s) in the time series, allowing us to use any standard machine learning algorithm for forecasting them.

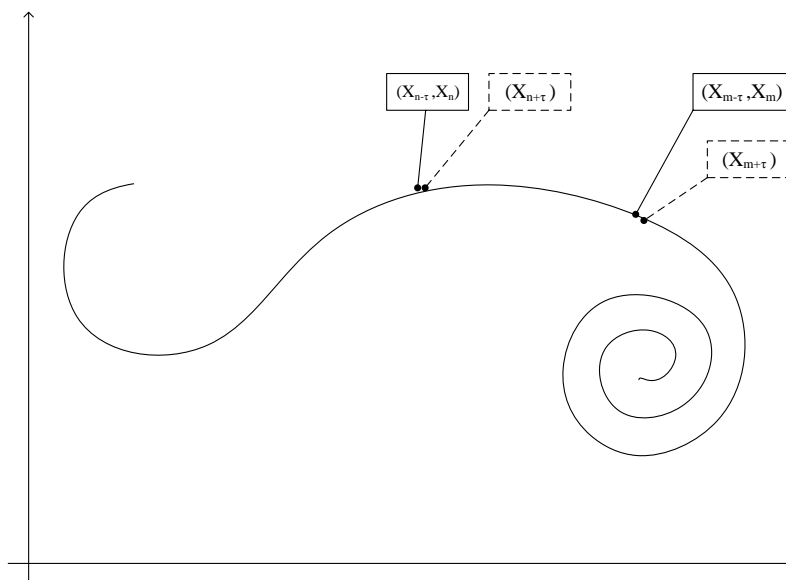


Figure 5 State space trajectory in a 2-D plane. This figure shows a sample case where future point on a trajectory is predicted with delay coordinate embedding for two different points  $n$  and  $m$  using a two-dimensional delay vector with time delay  $\tau$

For time series  $S$  which is a single dimensional vector of length  $M$ :

$$S = \{S_1, S_2, \dots, S_m, \dots, S_M\} \quad (3)$$

the delay reconstruction will construct a family of new vectors, in the form

$$\mathbf{S}'_n = (\mathbf{S}_{n-(m-1)\tau}, \mathbf{S}_{n-(m-2)\tau}, \dots, \mathbf{S}_{n-\tau}, \mathbf{S}_n) \quad 1 \leq n \leq N \quad (4)$$

where  $m$  is the number of dimensions. The time delay  $\tau$  determines whether we select consecutive observations, every second observation, every third, etc. The number of delay vectors  $N$  is given by:

$$N = M - (m-1)\tau \quad (5)$$

For example if  $M= 100$ ,  $m=3$ , and  $\tau=5$  then  $N= (100-2*5) =90$ . After delay reconstruction we have the set of vectors:

$$\begin{aligned} S_{100} &= \{S_{90}, S_{95}, S_{100}\} \\ S_{99} &= \{S_{89}, S_{94}, S_{99}\} \\ S_{98} &= \{S_{88}, S_{93}, S_{98}\} \\ S_{97} &= \{S_{87}, S_{92}, S_{97}\} \\ &\cdot \\ &\cdot \\ &\cdot \\ S_{11} &= \{S_1, S_6, S_{11}\} \end{aligned} \quad (6)$$

The main challenge in delay coordinate embedding is to find the correct number of dimensions and time delay. As mentioned above, the reconstructed space must have dimensionality greater than twice the original state space, but there is no constructive method for determining what that original dimensionality is. As for the time delay, every possible value is mathematically equivalent to every other one for an infinite time series. However, for real-world finite time series, the choice of the time delay has a significant practical effect on forecasting

outcomes. A small time delay makes elements of the delay vector redundant, while on the other hand large time delays make them almost uncorrelated. Several methods have been proposed for determining a time delay and number of dimensions [96]. In this research, the time-delayed mutual information method is used to determine the time delay, and the method of false nearest neighbors is used to choose the number of dimensions.

### 3.1.1 Mutual Information

Autocorrelation is a statistic function which describes the similarity between observations as a function of the time lag [97]. The definition of the autocorrelation,  $R(\tau)$ , of time lag  $\tau$  is:

$$R(\tau) = \frac{E[(X_t - \mu)(X_{t+\tau} - \mu)]}{\sigma^2} \quad (7)$$

where  $\mu$  and  $\sigma$  are mean and variance of data respectively and  $E$  is the expected value operator. In order to approximate the delay value with more information for delay coordinate embedding, in principle the lags equal to  $\tau$  is the best choice where the autocorrelation function is zero. [93] However, autocorrelation only treats the linear dependence of the time series and it does not consider the nonlinearity appropriately, it may find an incorrect value for the delay.[98] Therefore, it is advocated that look for the minimum of the mutual information. [93]

Mutual information is a concept from Shannon's theory of information entropy, which we will briefly review. Consider a system  $X$  with  $N_X$  possible states (i.e. a discrete system). If a measurement is performed on  $X$ , it will yield one of the possible values  $x_1 \dots x_{N_X}$ , each one with corresponding probability  $p(x_i)$ . The amount of information gained from a measurement yielding one particular value  $x_i$  is given by the entropy  $H(X)$  of the system [99]:

$$H(X) = -\sum_{i=1}^{N_X} p(x_i) \log p(x_i) \quad (8)$$

Entropy in essence describes the quantity of surprise one would feel upon taking a measurement of the system [100]. *Suppose that the event  $X=x_k$  occurs with probability  $p_k=1$ , which therefore requires that  $p_i=0$  for all  $i \neq k$ . In such situation, there is no "surprise" and therefore no "information" is conveyed by occurrence of the event  $X=x_k$ , since we know what the message must be. If, on the other hand, the various discrete levels were to occur with different probabilities and, in particular, if the  $p_k$  is low, then there is more "surprise" and therefore "information" when  $X$  takes the value  $x_k$  rather than another value  $x_i$  with higher probability  $p_i$ ,  $i \neq k$ . [101]* For a completely determined system, there is only one outcome which occurs with probability of one and therefore its entropy is zero and measurement providing no new information.

The joint entropy  $H(X, Y)$  of two discrete systems  $X$  and  $Y$  is defined as:

$$H(X, Y) = - \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} p(x_i, y_j) \log p(x_i, y_j) \quad (9)$$

where  $p(x_i, y_j)$  denotes the joint probability that  $X$  is in state  $x_i$  and  $Y$  is in state  $y_j$ .  $N_x$  and  $N_y$  are the number of possible states. In general, the joint entropy can be expressed in terms of the conditional entropy  $H(X|Y)$

$$H(X, Y) = H(X|Y) + H(Y) \quad (10)$$

with  $H(X|Y)$  being defined as

$$H(X|Y) = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} p(x_i, y_j) \log p(x_i | y_j) \quad (11)$$

The mutual information  $I(X, Y)$  between the systems  $X$  and  $Y$  is then defined as

$$\begin{aligned} I(X, Y) &= H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X, Y) \end{aligned} \quad (12)$$

Given a time series  $s(t)$ ,  $t=1, \dots, N$ , let us consider  $X$  as the value of the time series at  $s(t + \tau)$ , and  $Y$  as the time lagged value of the same time series,  $s(t)$ , with  $\tau$  being the time delay parameter from Section 3.1. By definition, the entropy  $H(X)$  is the uncertainty about the time series value  $X$  before observation of the time

series time lagged value  $Y$ , and the conditional differential entropy  $H(X|Y)$  is the uncertainty about the time series value  $X$  after the observation of the time series time lagged value  $Y$ . The difference  $H(X) - H(X|Y)$  is therefore the uncertainty about the time series value  $X$  that is resolved by observing the system time lagged value  $Y$  [101]. In other words, this is the information we already possess about the value of  $s(t + \tau)$  if we know  $s(t)$  [93].

A different approach to the mutual information is given by Kullback [102] who considers two probability density functions as the possible descriptors of the underlying distribution of the input vector  $X$  instead of multidimensional vectors  $X$  and  $Y$ . The Kullback entropy  $K(p|p_0)$  between two probability distributions  $p$  and  $p_0$  is:

$$K(p|p_0) = \sum_i p_i \log \frac{p_i}{p_i^0} \quad (13)$$

The Kullback entropy can be interpreted as the information gain when replacing an initial probability distribution  $p_i^0$  by a final distribution  $p_i$ . Therefore  $K(p|p_0)$  establishes a measure of the distance between the distributions  $p_0$  and  $p$ . However, the Kullback entropy is not symmetric and thus not a distance in the mathematical sense.

$$K(p|p^0) \neq K(p^0|p) \quad (14)$$

The Kullback entropy  $K(p|p_0)$  is always greater than or equal to zero and vanishes if and only if the distributions  $p$  and  $p_0$  are identical [103].

The mutual information between a pair of vectors  $X$  and  $Y$  representation in terms of Kullback-Leiber is:

$$I(X, Y) = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P(x, y) \log \left( \frac{P(x, y)}{P(x)P(y)} \right) \quad (15)$$

In using mutual information to determine a time delay  $\tau$ , we sum the mutual information between  $s(t+\tau)$  and  $s(t)$  for all  $t$  and a fixed  $\tau$ , and repeat for an arbitrary number of values of  $\tau$ . We then plot the sum of mutual information against  $\tau$ , and search for a minimum. [93] recommends that the first minimum of the plot (the extremum with the lowest value of  $\tau$ ) be used, rather than searching for the global minimum. This point retains the greatest correlation between consecutive delay vectors, while also representing a minimum in the redundancy within them [93].

### 3.1.2 False Nearest Neighbor

If there is enough information in the delay vector to predict the future output, then two delay vectors which are nearest neighbors in the delay space should have similar one-step ahead evolutions. On the other hand, if there are not enough terms present in the delay vector to recreate the dynamics of the system, then



there can be some neighbor vectors in the delay space which have very different evolutions. Since they are close in the delay space only because of projection into a space the representation of dynamics of the system is incomplete and topology of the system is not preserved. It is expected that the number of false neighbors will drop to minimum when the dimension of the delay vector is large enough to allow accurate prediction of future outputs. Thus, detecting such false nearest neighbors is a good criterion for determining if a given dimensionality is sufficient for a delay embedding of a given time series. We therefore search for delay vectors which are close in the delay space with vastly different outputs are considered as “false neighbors.”

In order to determine whether neighbors are true or false, false nearest neighbor test has been defined. This test goal is to determine whether the distance between future outputs of time delay vectors is larger than the distance between time delay vectors which are close in the delay space [93]. For each point of the time series, take its closest neighbor in  $m$  dimensions, compute the ratio of the distances between two future points in  $(m + 1)$  dimensions and  $m$  dimensions vectors. If this ratio is larger than a threshold, the neighbor is false.

$$R_i = \frac{|S_{i+1} - S_{j+1}|}{\|S_i - S_j\|} \quad (16)$$

The test calculates this ratio for all objects in time series and then calculates the percentage of points in the data set which have false nearest neighbors. The

algorithm will continue by increasing number of dimensions until the percentage of false nearest neighbor waveform is smooth and drops to acceptable small number. If the percentage of false neighbors is large, then the delay vector must increase to include more delayed terms.

### 3.2 Discrete Wavelet Transform

High dimensionality is a big problem in time series data mining because time series values have a lot of redundancy, which increases the complexity of the resulting models, and may make them less accurate. One solution is reducing the dimensionality of data with feature extraction which maps the original space into a new (lower-dimensional) feature space. The objective of feature extraction is to characterize the object and to map useful data to a new space suitable for the application of pattern recognition techniques [104]. When feature extraction compresses the time series, completeness and effectiveness of data should be preserved. For example the Euclidean distance as a similarity measure in the reduced dimensional space should be less than or equal to the Euclidean distance between the two original time sequences.

A time series in pattern space is showed by a vector with length of  $m$ .

$$X = x_1, x_2, \dots, x_m$$

In feature extraction only features that are necessary for recognition process retained and pattern recognition method only implements on a vastly reduced feature space. [105]. The feature vector can be represented with smaller

dimension by:

$$X' = x'_1, x'_2, \dots, x'_r \quad r < m$$

Fast Fourier transform (FFT) is a method for efficiently computing the discrete Fourier transform (DFT) of a time series. For a sequence of  $N$  numbers  $x_0, x_1, \dots, x_N$ , the FFT coefficients sequence  $X_0, X_1, \dots$  can be calculated using:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i2\pi kn/N} \quad (17)$$

FFT transforms time series into frequency space and provide frequency features but using FFT time localization of frequency components is not possible. It fails to detect the step changes in the signal and gives just an overall view on available frequencies in the signal. FFT is not sufficient for appliance pattern identification because for appliances energy consumption calculations from it is required to find time of use of each appliance in power signal. Therefore Discrete Wavelet Transformation is proposed which represents data in a time-frequency feature domain.

Discrete Wavelet Transformation (DWT) is a tool to map the sequences in time domain to a new feature space. DWT has the multi-resolution and time-frequency localization property. It analyzes the signal at different frequency bands with different resolutions; it decomposes the signal into a coarse approximation and detail information in different level of decomposition. In wavelet analysis, approximation coefficients are the high-scale, low-frequency components of the signal, while detail coefficients are low-scale, high-frequency components.

DWT yields a high dimensional feature vector that generally causes an increase in the complexity of a classifier. Further, the classification performance resulting from using all the original wavelet coefficients is poor when judged either by computation cost or classification accuracy. Dimensionality reduction is thus a required extension of this data transformation before applying a classifier. For this domain, dimensionality reduction is accomplished by selecting certain coefficients to retain, and discarding the others. This potentially makes the classifier algorithm both faster and more accurate, as the discarded coefficients are principally noise. The problem is how to select these coefficients. If we know in what frequencies the desired information is available, we can select the proper level of detail or approximation coefficients with the desired resolution; otherwise (which is usually the case), we must select coefficients based on heuristic criteria. There are two main categories to for wavelet feature selection: supervised and unsupervised method [106].

Unsupervised feature selection methods, independent of time series classes and categories, try to find the optimum set of coefficients with maximum similarity to the original data while minimizing the number of coefficients. Supervised features extraction evaluates the performance of methods based on class separability. Good quality class separation means having classification accuracy as high as possible with maximum separation between classes and minimum data variation inside each class.

Using the first few DWT coefficients and discarding the rest is the simplest method for DWT feature selection. First DWT coefficients tend to have

higher standard deviations, contain more of the signal energy, and carry more information [107]. A better metric for feature selection is to evaluate how similar different coefficient sequences in different decomposition levels are to the original signal by measuring how much energy from the original signal is preserved after the dimension reduction to one of these sequences. Although increasing the decomposition level reduces the data dimensionality more, this also usually decreases the similarity to the original sequence. The best wavelet decomposition level is the highest decomposition level where the retained signal energy of coefficients is over a threshold (e.g. 95%). This potentially reduces the dimensionality of data without severely distorting it [108].

The signal energy content of a signal provides a quantitative measure for signals. The amount of energy contained in a signal  $x(t)$  is expressed as:

$$E_{x(t)} = \int |x(t)|^2 dt \quad (18)$$

The energy content of a signal can be calculated from the signal's wavelet coefficients due to Parseval's theorem [109]. DWT coefficients  $x_j^n$  quantify the amount of energy associated with level of decomposition,  $n$ . The total amount of energy contained in the signal is equal to the sum of the energy in all coefficients. Since the energy content of each sequence of the signal is related to its properties it presents a unique feature for the system [106]. Energy in each level of decomposition,  $n$ , is calculated as:

$$E^n = \sum_{j=1}^{M_n} (x_j^n)^2 \quad (19)$$

where  $M_n$  coefficients are available in the level. The difference between the original time series and its extracted features is equal to the sum of the energy of all removed wavelet coefficients [106].

The Haar wavelet is the simplest wavelet transform, it is a series of averaging and differencing operations on a discrete time function. Haar wavelets are popular because of their simplicity, interpretable output and capability to preserve shape of the power waveform. The wavelet is defined as [110]:

$$\Psi_{Haar}(u) = \begin{cases} -1 & -0.5 < u \leq 0 \\ 1 & 0 < u \leq 0.5 \\ 0 & otherwise \end{cases} \quad (20)$$

It is not a continuous wavelet, so cannot smoothly follow a continuous signal, although this characteristic is beneficial when studying signals with sharp transitions. Moreover, it is a two element wide wavelet, which reduces its resolution. Coefficients of the first level of decomposition are obtained by taking the difference and average of two consecutive values. For example, for a sequence with 4 data points  $\{a_1, a_2, a_3, a_4\}$ , the detail and approximate coefficients are

$$\frac{1}{\sqrt{2}} \{(a_1 - a_2), (a_3 - a_4)\} \text{ and } \frac{1}{\sqrt{2}} \{(a_1 + a_2), (a_3 + a_4)\}, \text{ respectively. In this first}$$

level, the transform does not reflect the relation between all data points, e.g., there is no information showing relation of  $a_2$  and  $a_3$ . Haar wavelet is widely used in areas such as time series analysis, stream data mining and data bases because of

characteristics such as speed, memory efficiency, and ease of computation [111-114].

### 3.3 Multi-label Classification

Multi-label classification associates each instance in a dataset with a set of labels. This obviously means that ordinary classification algorithms cannot be used directly for multi-label classification. The two main approaches to creating multi-label classifiers are problem transformation and algorithm adaptation [115]. We will review method from each category: *RAkEL* is a problem transformation method, and *MLkNN* is an algorithm adaptation method.

#### 3.3.1 *RAkEL*

Problem transformation is a strategy in multi-label classification which divides a multi-label dataset into either one or more single label subsets, constructs a classifier for each data subset with a conventional classification technique, and consequently assembles all classifiers to build a multi-label classifier [115]. The Label Power set (LP) is a problem transformation method which considers each element of the power set of labels in a dataset as a new label, and trains a single-label classifier on the LP labels. While this is conceptually simple, LP classification commonly suffers from having a large number of label subsets, most of which are rarely encountered in the dataset [116].

*RAkEL* (RANdom *k*-labELsets), is an improved version of LP problem

transformation. RAKEL constructs an ensemble of LP classifiers with small random subset of the LP labels. The RAKEL algorithm iteratively and without replacement selects a set of labels with size  $k$  and constructs an LP classifier model. If  $L$  is the number of labels, the maximum number of classifiers  $L^k$  is  $\binom{|L|}{k}$

. For classification of a new instance, each LP model provides binary decisions for each label in the corresponding  $k$ -label set. RAKEL averages the zero one predictions of each model per considered label, and assigns that label to the instance if the average mark is greater than a threshold [117]. The size of the label sets and the decision threshold are user-defined parameters. RAKEL is implemented as a meta-classifier in the Mulan plug-in for the WEKA data-mining environment; any of WEKA's conventional classifiers can be used as the base classifier in the ensemble. We will use C4.5 decision trees, support vector machines, and Naïve Bayes classifiers in our experiments.

### 3.3.2 ML $k$ NN

Another strategy for multi-label classifications is algorithm adaptation. Algorithm adaptation, as its name implies, refers to methods that extend specific learning algorithms such as  $k$ -nearest neighbors to handle multi-label problems. ML $k$ NN is a multi-label learning approach derived from the popular  $k$ -Nearest Neighbor ( $k$ NN) algorithm [90]. ML $k$ NN approach finds the label set for a given test instance using maximum a posteriori (MAP) [118], based on prior and posterior probabilities of each  $k$  nearest neighbor instances. For each instance, ML $k$ NN first



identifies its  $k$  nearest neighbors in the training set and then determines the probability of each label with the following MAP principle:

$$Z_t^\lambda = \arg \max_{b \in \{0,1\}} \mathbf{P}\left(H_b^\lambda \mid E_{Ci(\lambda)}^\lambda\right) \quad (21)$$

$Z_t^\lambda$  is predicted label set for instance  $t$ ,  $H_1^\lambda$  and  $H_0^\lambda$  are the events that  $t$  has label  $\lambda$  and does not have label  $\lambda$ , respectively,  $E_i^\lambda$  is the event that exactly  $j$  instances of  $k$  nearest neighbor of test instance  $t$  has label  $\lambda$ , and  $Ci(\lambda)$  is the membership vector that counts the number of neighbors of  $t$  belonging to class  $\lambda$ . Using Bayes' rule this can be re-written as:

$$Z_t^\lambda = \arg \max_{b \in \{0,1\}} \mathbf{P}\left(H_b^\lambda\right) \left(E_{Ci(\lambda)}^\lambda \mid H_b^\lambda\right) \quad (22)$$

where  $\mathbf{P}\left(H_b^\lambda\right)$  is prior probabilities and  $\left(E_{Ci(\lambda)}^\lambda \mid H_b^\lambda\right)$  is posteriori probability. All these prior and posterior probabilities can be directly estimated from the training data [90].

### 3.4 Clustering

Unsupervised clustering groups instances in a dataset to have maximum similarity inside the clusters and minimal similarity between them, where similarity is

usually defined as proximity in feature space. The goal of clustering is to find hidden categories without using any category labels. Statistically, a hidden category can be a probability distribution over the dataset; mathematically, it can be represented by a probability density function. Expectation Maximization (EM) is a probabilistic clustering algorithm.

### 3.4.1 EM Clustering

Assume that we have a dataset  $D$  containing  $n$  instances, for which we will build  $k$  clusters. Each cluster  $C_j$  ( $1 \leq j \leq k$ ) is associated with a probability  $P(C_j)$  that an arbitrary instance is in that cluster [119]. The probability density function and probability of each cluster are unknowns to be determined. The observed objects in the dataset are the mixture of instances from multiple probabilistic clusters [119]. The mixture model assumes data has been generated independently; the clustering algorithm uses the probability of each cluster and probability density function of that cluster to calculate the probability of belonging to each cluster and makes a final decision based on comparing probability of object in different clusters.

Consider a set of  $k$  probabilistic clusters,  $C_1 \dots C_k$  with probability density function of  $P(x|C_1), \dots, P(x|C_k)$  and probability of  $P(C_1), \dots, P(C_k)$ . The probability that instance  $x$  is generated by the set of clusters is:

$$P(\mathbf{x}) = \sum_{j=1}^k P(C_j)P(x|C_j) \quad (23)$$

For  $n$  independent instances in dataset  $D$  the marginal likelihood function is:

$$P(D) = \prod_{i=1}^n P(x) = \prod_{i=1}^n \sum_{j=1}^k P(C_j)P(x_i|C_j) \quad (24)$$

Our clustering algorithm will maximize the total likelihood of the dataset. Although the best estimate can be achieved by solving the log-likelihood equations, solutions of the likelihood equations cannot be obtained analytically in most circumstances. To make the problem computationally feasible, we assume that the probability density function is a parameterized distribution, and use an iterative suboptimal approach to approximate clusters. Let us assume  $D$  is a random vector which results from a parameterized family. The goal is to find  $\theta$  such that  $P(D|\theta)$  is a maximum. This is known as the Maximum Likelihood (ML) estimate for  $\theta$ . In order to estimate  $\theta$ , it is typical to introduce the likelihood function defined as:

$$P(D|\theta) = \prod_{i=1}^n P(x|\theta) = \prod_{i=1}^n \sum_{j=1}^k P(C_j)P(x|C_j, \theta) = \prod_{i=1}^n \sum_{j=1}^k P(C_j)P(x|\theta) \quad (25)$$

Consequently the task of clustering analysis here is to infer a set of parameters,  $\theta$ , to maximize likelihood function. It is common to show the log-likelihood

function:

$$L(\theta|D) = \log P(D|\theta) = \log \prod_{i=1}^n \sum_{j=1}^k P(C_j) P(x|\theta) \quad (26)$$

Expectation maximization clustering approximates the maximum likelihood of parameters in statistical models. This method starts with initial parameters and iteratively improves the output until no more improvement can be seen. This algorithm has two steps:

The expectation step (E-step):

- Assign objects to clusters according to the current parameters.

The maximization step (M-step):

- Find new parameters to maximize expected likelihood in probabilistic model.

### 3.5 Similarity Search

Algorithms that work with time series data commonly compute the distance between pairs of time series as a basis for classification or clustering [120]. Euclidean distance, or its extension or modification is commonly used for this purpose. A similarity search based on Euclidean distance maps two sequences with length  $n$ , into points in an  $n$ -dimensional space and computes the Euclidean distance between those points. For two sequences  $X=(x_1, x_2, \dots, x_n)$ , and  $Y=(y_1, y_2, \dots, y_n)$  the Euclidean distance is defined as [121] :

$$D(X, Y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (27)$$

Dynamic time warping (DTW) is a robust similarity (or dissimilarity) measure for time series which can match similar waveforms even if their signatures are shifted, stretched or compressed on the time axis. DTW does not obey the triangular inequality and thus provides approximate similarity [122].

According to the application either Euclidean distance or DTW can be used to detect similar sequences. Since in NILM the duration of the appliances waveform is not fixed, even a small time variation confounds the Euclidean distance metric. Dynamic time warping is the proper method in this thesis to match the detected segments to the available database for NILM.

### 3.6 Dynamic Time Warping

Dynamic Time Warping (DTW) is a search algorithm which is capable of flexible matching between sequences. DTW is capable of measuring distance between sequences of different lengths, as well as sequences of the same length. Euclidean distance between two sequences is a special case of DTW where the two sequences have same length [123]. DTW is useful for matching sequences where one sequence is a shifted or extended version of another sequence.

For two time series  $Q$  and  $C$  with length  $n$  and  $m$  respectively where:

$$Q = q_1, q_2 \dots q_i \dots q_n$$

$$C = c_1, c_2 \dots c_j \dots c_m$$

To find DTW similarity an  $n$ -by- $m$  cost matrix,  $P$ , is constructed. Each element of the constructed cost matrix,  $(i^{\text{th}}, j^{\text{th}})$  corresponds to the distance cost between two points  $q_i$  and  $c_j$  [124]. The distance measure between time series elements can be Euclidean distance  $P(i, j) = (q_i - c_j)^2$  or other methods such as Manhattan distance

$$P(i, j) = |q_i - c_j|.$$

There are exponentially many warping paths  $W$  which connect starting and ending point of data, in the form of :

$$W = w_1, w_2, \dots, w_i, \dots, w_K \quad \max(m, n) \leq K \leq m+n-1$$

where  $K$  is length of the warping path,  $m$  and  $n$  are length of two compared sequences and  $w_i$  is warping path element in cost matrix. However the final solution of the algorithm is the warping path between  $Q$  and  $C$  with minimum overall cost given the cost matrix  $P$ . The total cost of a warping path  $DTW(Q, C)$  between  $Q$  and  $C$  is defined as:

$$DTW(Q, C) = \min \left\{ \sqrt{\sum_{k=1}^K P(w_k)} \right\} \quad (28)$$

The path with minimum overall cost would run along a “valley” of low costs within the cost matrix [123]. Figure 6 demonstrates an example of DTW matching

between two sequences of refrigerator with different length. This figure shows the warping path in cost matrix which has minimum overall cost.

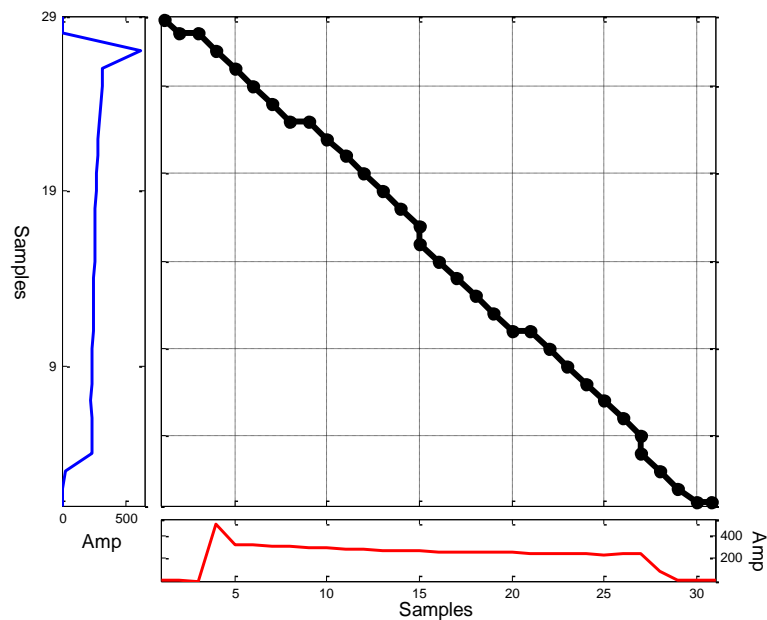


Figure 6 Example of Dynamic Time Warping

The warping path is typically subject to several constraints and shall satisfy the following conditions [124]:

- **Boundary Conditions:**  $w_1 = (1, 1)$  and  $w_K = (m, n)$ . The warping path has to start and finish in diagonally opposite corner cells of the cost matrix.
- **Step size condition:** Given  $w_i$  and  $w_{i-1}$  then  $w_i - w_{i-1} \in \{(1,0), (0,1), (1,1)\}$ .

The warping path in each step is allowed to go only to adjacent cells.

- **Monotonicity Condition:** Given  $w_i=(a,b)$  and  $w_{i-1}=(a',b')$  where  $a-a' \geq 0$  and  $b-b' \geq 0$ . The points in  $W$  must be monotonically ordered with respect to time.

Dynamic programming is used to find minimum warping path between two sequences [125]. It constructs a new matrix  $D$ , which is referred to as the accumulated cost matrix. Its elements are based on the following recurrence relation which defines the cumulative distance,  $\gamma(i, j)$ , for each point as:

$$\gamma(i, j) = d(q_i, c_j) + \min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\} \quad (29)$$

The cumulative distance is the sum of the distance between current elements and the minimum of the cumulative distances of the neighboring points. The dynamic programming algorithm fills the matrix with cumulative distances as the computation proceeds. After constructing the cumulative distance matrix, the optimal warping path can be found by tracing backward in the matrix and choosing the cells with the lowest cumulative distance [125].



## Chapter 4

### METHODOLOGY

#### 4.1 Datasets

To explore the proposed methods, they are evaluated on both simulated and real-world datasets. Simulated datasets are built using available data from measurements of real appliances. Our real-world datasets consist of two houses tracked in the REDD dataset.

##### 4.1.1 Simulated Datasets

Two simulated datasets have been generated in order to test the proposed methods in different scenarios. The power waveforms of several houses are currently being monitored by the PDS lab at the University of Alberta. These measured data have been used to build two simulated datasets with software developed by another student in the PDS group. These datasets contain two appliances: refrigerator and microwave. To provide a practical situation for the test, background power including random appliances and noise has been added to the generated datasets. The main difference between these two datasets is the volume of background power. Dataset 1 consists of two simulated appliances added to a real home power waveform. Dataset 2 has less background power; it contains vampire energy, small disturbances and noise. Simulated power signals are produced for six

independent days.

Dataset 2 is a simple case which simulates the situation where there are a few appliances in a home and appliances are not mixed very much. The objective in dataset 1, the other hand, is to detect two specific appliances from a mixture of appliances while ignoring others by labeling them as unknown or not important. Dataset 1 simulates the situation where labels are not available for all of the appliances, and we only intend to disaggregate a few specific, important appliances.

#### **4.1.2 REDD: The Reference Energy Disaggregation Dataset**

In REDD both whole household and circuit/device power consumption are collected from real houses in the Boston area in the USA. Circuit/device level energy consumption data allow us to label the power waveform with the appliances in use. We are then able to train and test supervised learning algorithms on this data. Labels are assigned to the power signal in any sample instant where the power consumption is greater than 10W. This dataset is available at <http://redd.csail.mit.edu> [6].

## **4.2 Delay Embedding**

We employ the time-delayed mutual information and false-nearest-neighbor techniques of Section 3.3 to determine the time delay and dimensionality, respectively, for a delay embedding for each of our datasets [93]. Both

techniques, and the final construction of the delay vectors, was performed using the TISEAN software package [126]. The output is a matrix with each row being one delay vector. For NILM, we then have to add labels for each row indicating what appliances are active at that sample instant. Elements in the label vector represent each appliance, and are set to 1 if that appliance is ON at that sample instant. An example of a delay matrix and the corresponding label matrix (for two appliances) with corresponding labels is shown below.

$$\begin{array}{cc}
 \textit{Delay matrix} & \textit{Labels} \\
 \left[ \begin{array}{ccc|cc}
 S_n & S_{n-\tau} & S_{n-(m-1)\tau} & 1 & 1 \\
 S_{n-1} & S_{n-1-\tau} & \cdots S_{n-1-(m-1)\tau} & 0 & 1 \\
 S_{n-k} & S_{n-k-\tau} & S_{n-k-(m-1)\tau} & 0 & 0
 \end{array} \right] & 
 \end{array}$$

The power waveforms of a refrigerator within one of the houses monitored by the PDS lab are highlighted in Figure 7. The general waveform of the refrigerator is fixed; however its mixture with other appliances could cause a vertical shift in its location, or produce a combined waveform with a different shape.

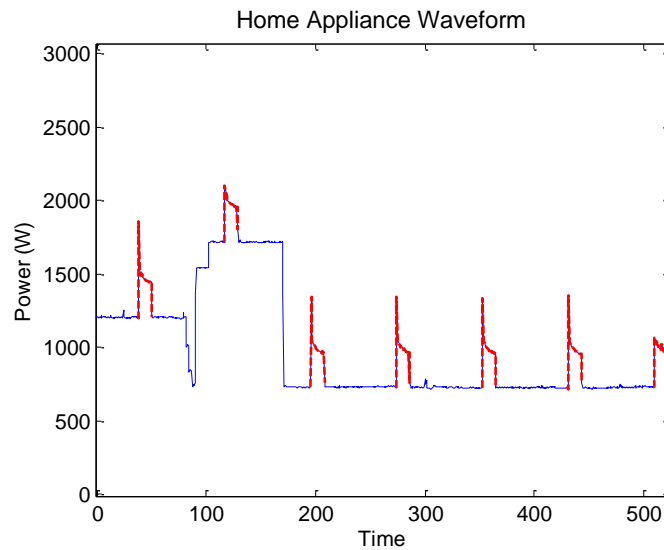


Figure 7 Home Appliance Waveform, Refrigerator features are highlighted

### 4.2.1 Delay embedding parameter selection

Delay embedding parameters should be chosen separately for each dataset. In the following sections its parameters have been selected for available datasets.

#### 4.2.1.1 Delay embedding parameters in simulated datasets

Figure 8 shows the mutual information plot for simulated dataset 1. The first minimum of mutual information curve for dataset 1 is at 35 seconds. Two curves has been plotted which show mutual information for active and reactive power. Sometimes active and reactive power mutual information curves have different shapes and different minimum values. In this case, since all appliances have active power, the delay parameter is selected based on the active power curve.

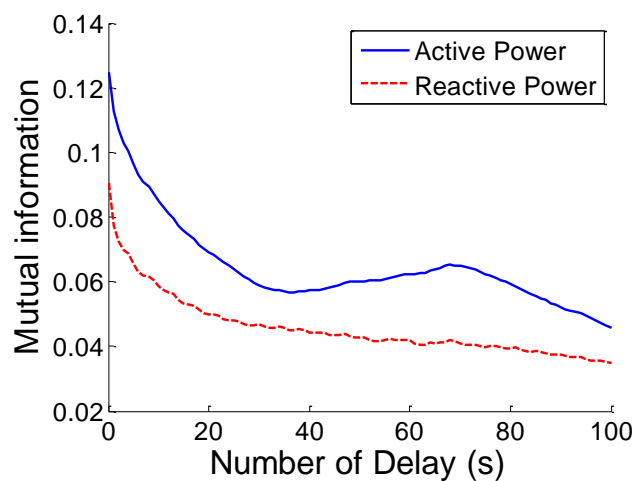


Figure 8 Mutual Information, Dataset 1,  $\tau=35s$

Mutual information for dataset 2 gradually decreases with increasing delay and has no minimum as shown in Figure 9. There are not many power waveform changes in this dataset, and low uncertainty in the data. We therefore choose a time delay of 1 second for dataset 2.

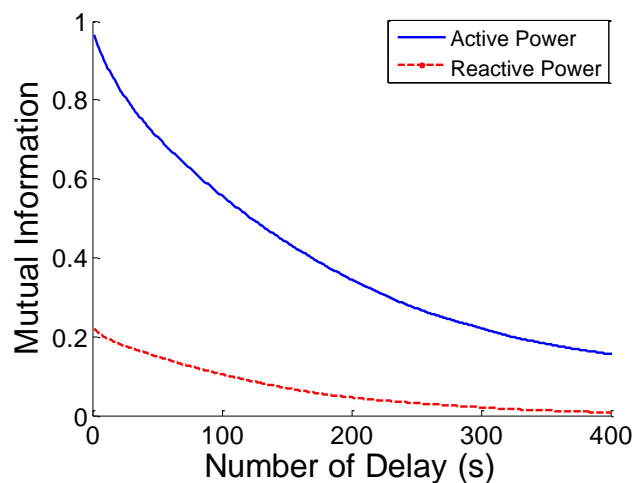


Figure 9 Mutual Information, Dataset 2,  $\tau=1s$

Figure 10 shows the FNN plot for dataset 1. Changes in false nearest neighbor curves are small after  $m=16$ ; therefore we use a 16-dimensional embedding for this dataset.

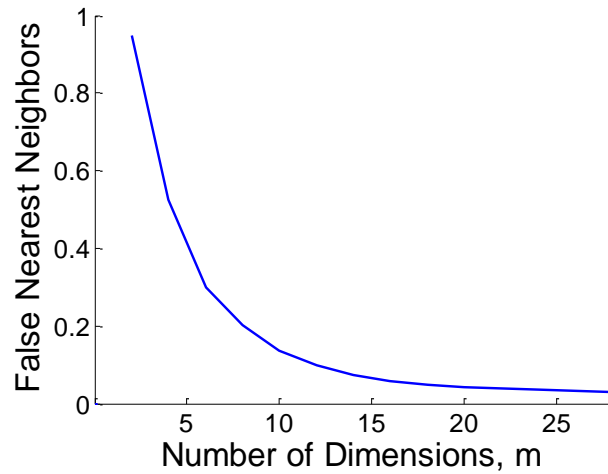


Figure 10 False Nearest Neighbor ratio, Dataset 1,  $m=16$

The dataset 2 FNN curve reaches zero at  $m=8$  in Figure 11, therefore for dataset 2 at least 8 dimensions are required to reconstruct the state space.

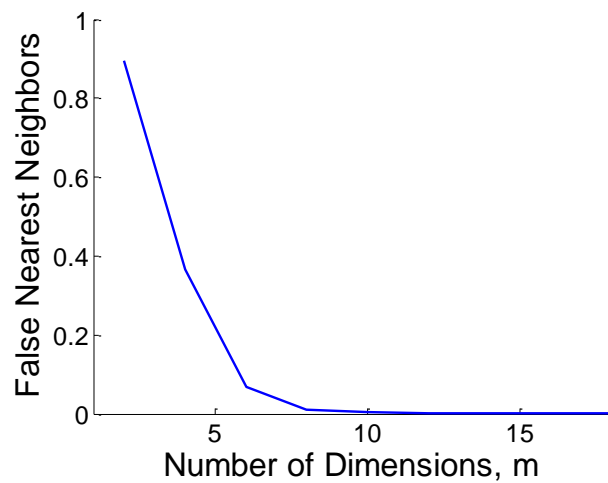


Figure 11 False Nearest Neighbor Dataset 2,  $m=8$

After creating the delay vectors, appliance labels are assigned to the reconstructed delay vectors. The first column of the label vector represents the refrigerator and the second column represents the microwave, and the third represents unknown appliances in the background (this label is always 1 since there is always background power in the signals).

#### **4.2.1.2 Delay embedding parameters in REDD**

In REDD [6] numerous gaps in the data occur, so only continuous sequences of points have been used to reconstruct delay space. We then combine the delay vectors from these sequences, as they represent the same house. Two houses have been used in this research to validate the proposed methods: house 3 and house 1. For REDD house 3, mutual information and false nearest neighbor methods indicate that the best time delay is 95s, and 18 dimensions are required (see Figure 12 and Figure 13).

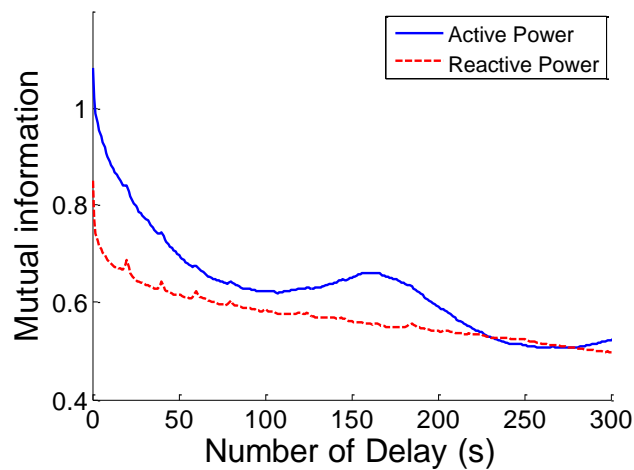


Figure 12 Mutual information, House 3,  $\tau=95s$

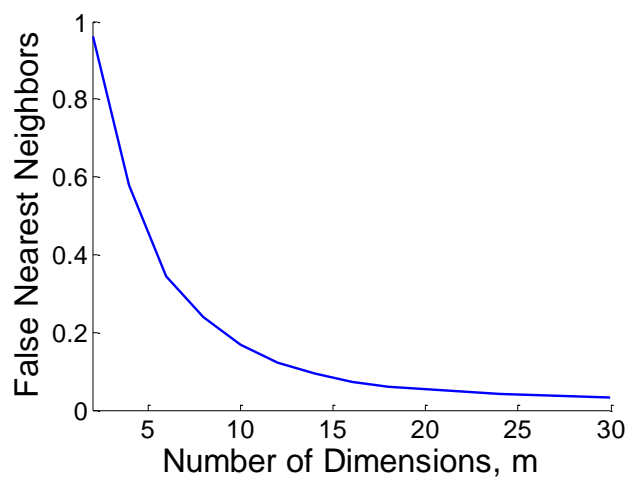


Figure 13 False nearest neighbor ratio, House 3,  $m=18$

For REDD house 1, the data has only active power and reactive power information is not available. For this dataset mutual information and false nearest neighbor curves are depicted in Figure 14 and Figure 15, respectively. From these plots, the minimum number of dimensions is 8 and the time delay is 32.



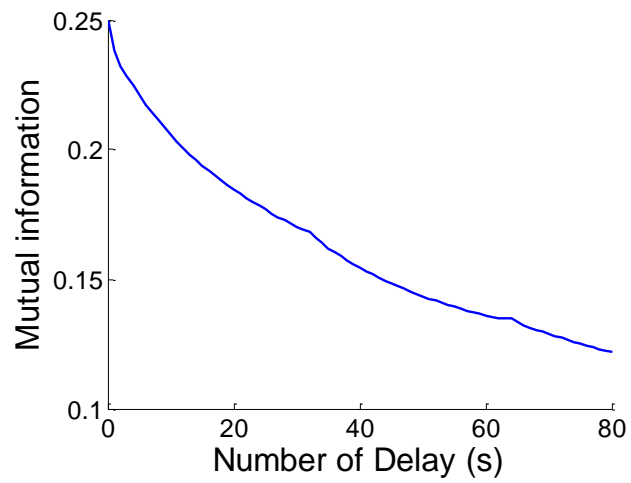


Figure 14 Mutual information, House 1,  $\tau=32s$

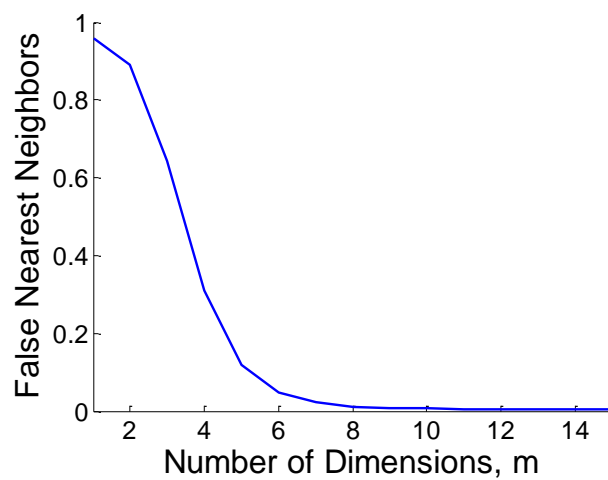


Figure 15 False nearest neighbor ratio, House 1,  $m=8$

### 4.3 Discrete Wavelet Feature extraction

The measured power signal in a home depending on sampling time has different size, for instance there are 86400 data points in one day data measured with sampling time of 1 sec. If the reactive power has also measured then the size of data points will be doubled. This large number of data points shows the importance of dimensionality reduction. In selecting our wavelet coefficients, we wish to use the highest decomposition level which preserves at least 95% of the signal energy.

As simulated datasets, dataset 1 and 2 contain the same appliance waveform, so one test is enough to select the wavelet decomposition level for both. Table 2 show the energy preserved in each level of decomposition. According to the results, the level four coefficients have been selected as our wavelet-domain features. This means the amount of data to be process has been reduced by a factor of 16.

As dataset 1 includes more than just the two registered appliances, high level of decomposition destroy other appliances waveform shapes and make identification complicated, especially when the appliances are mixed together. Therefore, based on experienced on other datasets with more appliances, level three has been chosen for this dataset.

Table 2 Preserved energy on Haar wavelet decomposition levels in simulated dataset

Decomposition Level	Refrigerator	Microwave
1	100.0%	100.0%
2	99.2%	99.4%
3	98.9%	99.1%
4	97.4%	97.3%
5	96.2%	94.6%
6	94.3%	93.2%
7	93.4%	80.7%
8	94.7%	73.0%

Table 3 shows the signal energy preserved for each decomposition level of Haar wavelets for each appliance in House 3 of the REDD dataset. After examining the table, we have selected the third decomposition level as being the best balance of dimensionality reduction and retained energy. Only the Bath GFI outlet (wall plug) shows less than 95% of the signal energy being retained at this level.

Table 3 Preserved energy on Haar wavelet decomposition level in REDD, house 3

Decomposition Level	Electronics	Furnace	Washer	Microwave	Bath GFI	Kitchen Outlet
1	99.6%	100%	99.9%	99.4%	99.8%	99.7%
2	99.0%	100%	98.9%	99.0%	99.7%	99.5%
3	96.5%	99.6%	96.1%	98.5%	94.6%	96.4%
4	95.3%	99.1%	92.8%	93.2%	91.3%	94.2%
5	86.1%	98.8%	88.9%	90.5%	84.7%	89.5%
6	81.2%	98.0%	86.7%	89.1%	81.3%	87.2%
7	60.9%	95.4%	66.9%	54.1%	55.0%	56.2%
8	46.9%	94.4%	53.4%	34.2%	36.8%	43.3%

Table 4 shows the signal energy preserved for each decomposition level of Haar wavelets for each appliance in House 1 of the REDD dataset. Only one appliance drops below 95% retained signal energy at level 3, but that one is the oven – a high-draw appliance with short, frequent duty cycles when in use. It seems unwise to allow this appliance to have less than 95% retained energy, and so we select decomposition level 2 for this dataset.

Table 4 Preserved energy on Haar wavelet decomposition level in REDD, house 1

Decomposition Level	Oven	Refrigerator	Light	Microwave	Bath GFI	Outlet	Washer
1	99.8%	100.0%	100.0%	99.9%	99.9%	99.9%	100.0%
2	97.8%	99.9%	99.9%	99.8%	99.8%	99.9%	99.9%
3	94.5%	99.7%	99.7%	98.5%	99.3%	99.6%	99.5%
4	92.6%	99.3%	99.6%	97.7%	98.1%	99.5%	99.4%
5	90.7%	99.1%	98.7%	94.6%	97.0%	97.3%	98.4%
6	89.7%	98.9%	98.0%	93.1%	94.2%	96.0%	98.0%
7	49.1%	94.7%	93.7%	75.4%	83.4%	86.7%	92.7%
8	28.9%	86.7%	91.7%	69.2%	62.5%	67.8%	81.5%

#### 4.4 Multi-label classification

For our multi-label classification experiments, we divide the data into equal halves, chronologically order. The earlier half is our training dataset (in which we will conduct classifier parameter exploration), and the latter half is our testing dataset (used for test after we have found the best parameterization of our algorithm on the training set). Once that the best parameters of the classifier is found, we will train the classifier on the entire training set, and then evaluate it on the test set.

#### 4.4.1 Evaluation of Multi-label Classification

Multi-label classification requires different evaluation measures than those used in traditional single-label classification [127]. Some of them are based on the average differences of the actual and the predicted sets of labels over all test examples, but others decompose the evaluation process into separate evaluations for each label and subsequently average over all labels.

Evaluation measures can be calculated using one of two averaging operations, called macro-averaging and micro-averaging. Considering a single label evaluation measure for each label, macro-averaging is the average of evaluation measures calculated for each class individually. It shows the ability of a classifier to behave well on all categories even those with a small number of examples [128].

$$M_{macro} = \frac{1}{|L|} \sum_{\lambda=1}^{|L|} M(tp_{\lambda}, fp_{\lambda}, tn_{\lambda}, fn_{\lambda}) \quad (30)$$

where  $L$  is number of labels and  $tp$ ,  $fp$ ,  $tn$  and  $fn$  are true positive, false positive, true negative and false negative respectively.

Micro-averaging combines TP, TN, FP, and FN for examples across all categories into one contingency matrix. It can reflect better classification rates for large classes at the expense of worse results for small ones [128].

$$M_{micro} = M \left( \sum_{\lambda=1}^{|L|} tp_{\lambda}, \sum_{\lambda=1}^{|L|} fp_{\lambda}, \sum_{\lambda=1}^{|L|} tn_{\lambda}, \sum_{\lambda=1}^{|L|} fn_{\lambda} \right) \quad (31)$$

where  $M$  is an evaluation measurement method such as F-measure and accuracy which has been used in this thesis. Precision and recall are proper evaluation parameters for appliance identification. Precision can be seen as a measure of exactness, whereas recall is a measure of completeness. However both of them are not the best at the same time and there is a trade off between them that higher recall is equal to lower precision, it makes comparison and parameter selection hard, therefore F-measure which combines precision and recall results is used.

#### 4.5 Clustering

We evaluate the impact of clustering to improve performance of multi-label classification on NILM problem in this research. The idea is to use clustering method to split data into smaller (and hopefully more homogenous) groups, and then train a multi-label classifier on each cluster. The ideal case is when the clustering method assigns just one appliance to each cluster. The basic assumption in clustering is that two objects belong to the same cluster if they are very similar to each other. If each cluster contains only one of the appliances then the clustering would be the final solution for appliance identification and the method would be unsupervised. However, in multi-label problems the accuracy of clustering drops as the basic assumption becomes less valid due to the high

dimensionality and sparseness of data [129]. Therefore, appliances cannot usually be disaggregated purely by clustering. Instead, we are attempting to create an ensemble of local models to improve our classification accuracy.

To cluster the data, we use the EM clustering algorithm, as implemented in WEKA. After learning EM method with training part of the data, on each cluster one multi-label classifier implemented with the same training data. Labels are ignored in clustering but are necessary to train the classifier. In testing, each new object is assigned to one of the existing clusters and then the local classifier predicts its appliance labels. Figure 16 shows the multi-label classification idea along with clustering to detect appliances features.

The number of clusters in EM clustering is a user defined parameter. To have optimum number of clusters for better classification performance and to study the effect of increasing number of clusters on classification output, number of clusters has manually tweaked. Although classification of data can be done individually without clustering, we expect clustering to improve our classification accuracy.



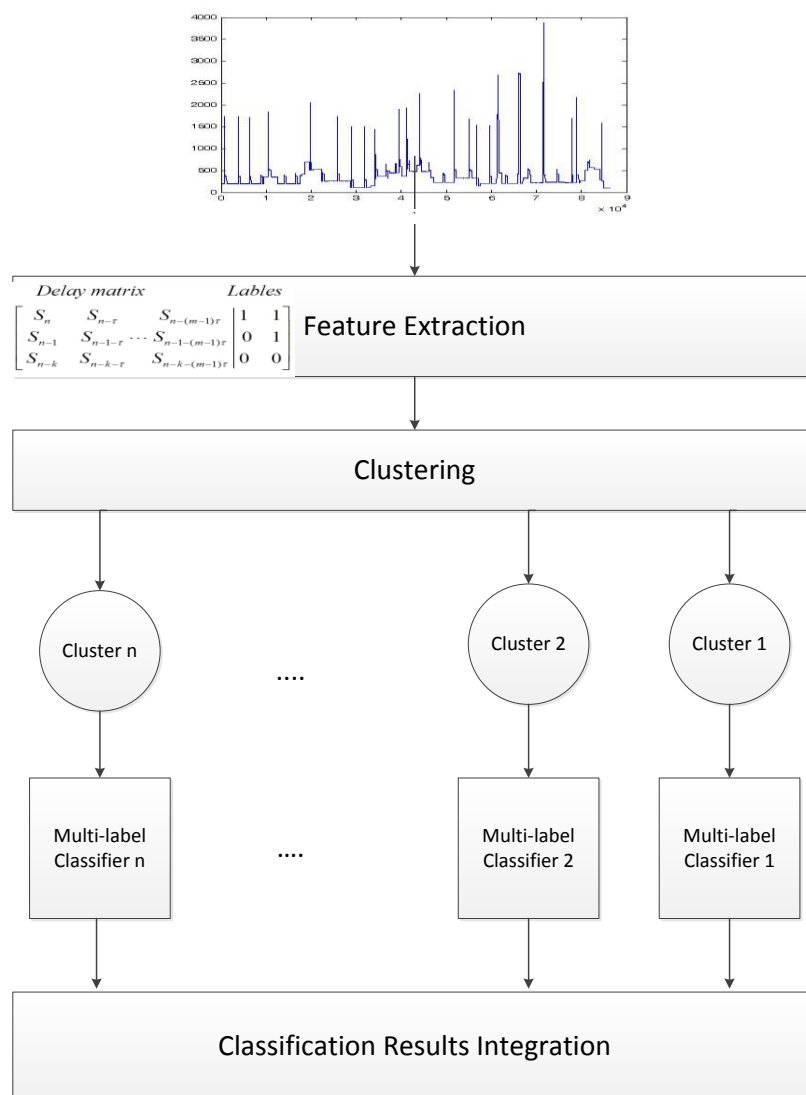


Figure 16 proposed Multi-label classification method along with clustering

## 4.6 Similarity Search

Most NILM methods are based on supervised classification. However, the characteristics of power time series such as high dimensionality, high feature correlation, large amounts of fluctuations, and mixture of features have caused conventional machine learning algorithms to not work well to solve the

identification problem. The multi-label classification method we have proposed needs a considerable amount of training data with labels for all appliances, which in practice is difficult and expensive. Therefore we propose a second method that does not need labels for all data points and has an easier training procedure. The focus of the proposed method is on developing a similarity measure for NILM. This method needs to be robust to time scale and location variations of appliance signatures in the measured power signal.

The objective of the similarity search method is finding the most similar match of a measured candidate sequence to a set of known sequence of appliances in the database. This is achieved by sequentially scanning and comparing each and every candidate to the database query to find the most-similar sequence. To make this solution more practical, the dimensionality of the power signal has been reduced with the discrete wavelet transform (using Haar wavelets).

The proposed similarity search method proceeds as follows:

- First, execute the discrete wavelet transform over the query sequence.
- Detect and mark events in the power sequence. Changes and events in the power signal show change in behaviors of appliances e.g. turning an appliance ON.
- Divide the power time series into smaller power segments. Detected events in the signal are the basis for this division. Each segment is a candidate for appliances inside the home.
- Find the best match of each segment with registered appliances in the

database using similarity search.

- Assign the label of that best match to that segment, if the similarity is greater than a predefined threshold.

Figure 17 summarizes the proposed method. In the following sections the detail of each part of the proposed method are introduced.

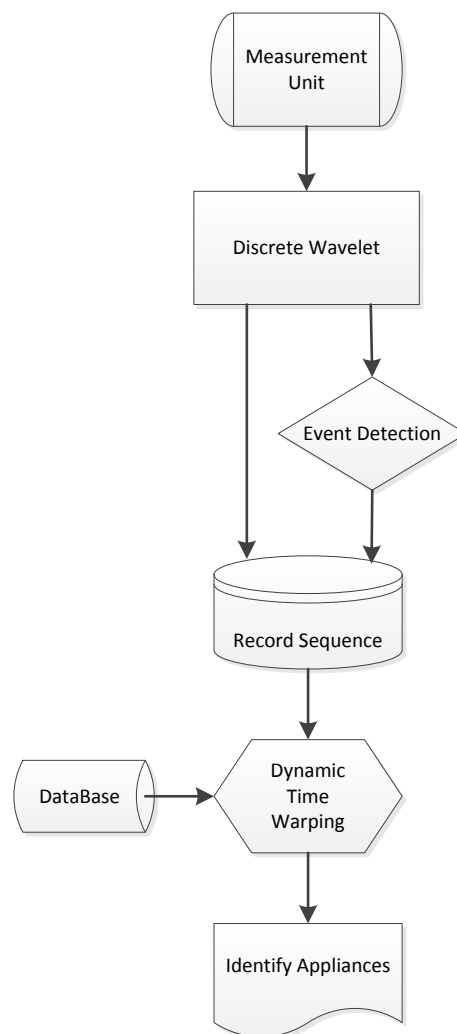


Figure 17 Similarity search method procedure

#### 4.6.1 Dimensionality Reduction using Haar Wavelets

Figure 18 depicts the power signal before and after the wavelet decomposition and dimensionality reduction. The shape of the waveforms is almost completely preserved whereas the number of data points has decreased. Normalization factor of coefficients in this level is  $(\sqrt{2})^3$  times that of the original signal. This constant multiplier is for all available data including power signal and database objects, and thus makes no problem in appliances identification process.

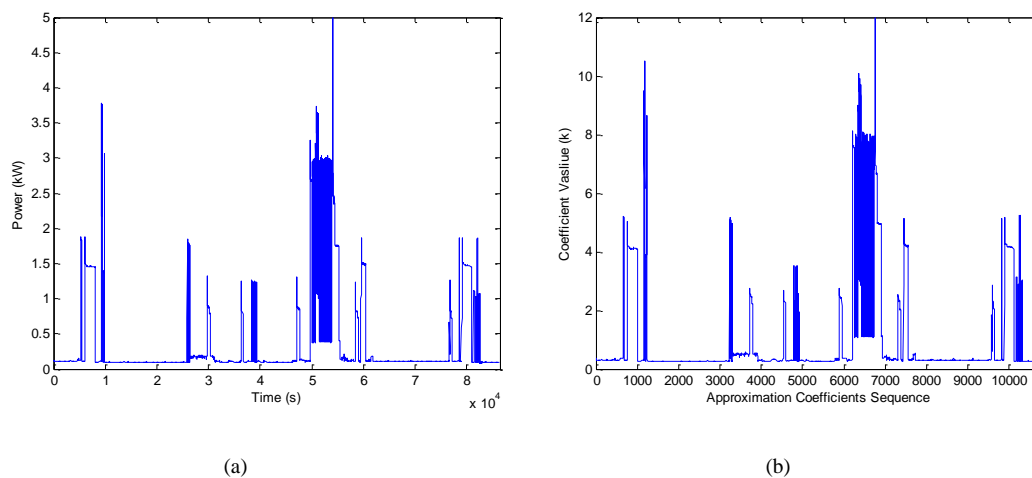


Figure 18 Wavelet effect on shape of signal, (a) One day REDD power signal before transformation; (b) level 3 Haar coefficients of (a);

#### 4.6.2 Edge Detection

After transforming the power time series into the wavelet domain, the new dataset

is a sequence of wavelet coefficients with smaller length than original power signal. The objective is to optimize the search space for appliance disaggregation from the whole time series to small sequences; each sequence is a candidate to be a specific appliance. Dividing data into smaller segments is based on the events inside the data.

An event in the power sequence means a change in behavior of an appliance. When turning ON an appliance, the power level in the waveform increases and a rising edge appears in the signal, in contrast when turning it OFF the power level decreases and a falling edge appears in the signal. Edges in the power signal can also be created by multi-state appliances, when there is a change in the operating state of the appliance while it is working.

There are different edge detection methods for signals [130]. Most of these methods are based on detecting a variation in a signal when comparing each point with its neighbors. In this project wavelet high frequency coefficients are used to detect edges in power sequence. Wavelet coefficients are available due to wavelet dimension reduction and new calculations are not required. Haar detail coefficients have the same locality properties as approximation coefficients so the time of change in detail coefficients corresponds to the time of change in approximation coefficients.

Detail coefficients show non-zero rapid changes just when there is an event in the signal. Therefore detail coefficients determine the time of event, type of event and edge size. Figure 19 shows a sample case where the location and

type of event (rising/falling) can be determined with Haar wavelet detail coefficients. Edge detection with wavelet transform application in NILM is novel.

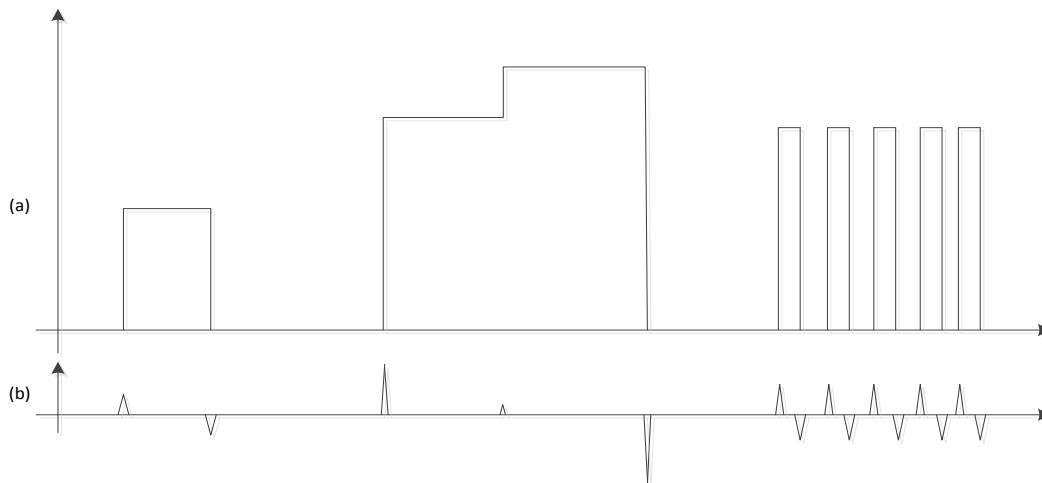


Figure 19 Sample Approximate coefficients and related detail coefficients which determine events in waveform (a) Approximation coefficients sequence (b) Detail Coefficients which have value at edge locations

To detect a change with discrete wavelet transform, the wavelet has to be sufficiently regular without discontinuity. Haar wavelet is not sufficiently regular for this analysis and some of the changes in signals would be missed. To solve the regularity problem of Haar wavelet, the length of the detail filter (high pass filter) is modified. A Haar wavelet with optimal length of the high pass filter would reflect all the edges in high frequency coefficients via local maxima. The high frequency filter is up-sampled by adding zeroes between each filter coefficient to increase the filter length from two to four terms and double the frequency

response [131]. Approximation coefficients remain unchanged and all changes would appear in the detail coefficients.

If the nonzero value of detail coefficient is greater than a threshold corresponding to the minimum edge height of appliances in database, the location is flagged as an event and then the sequence is segmented.

### **4.6.3 Coefficients sequence Segmentation**

In order to make event based segments when a rising edge is detected in the power signal, the waveform is recorded until detecting a falling edge. Single state and continuously varying appliances have a pair of ON/OFF edges in each time of use; one segment with a pair of rising/falling events is enough to contain all features of these appliances. However, any number of unrelated edges may be present in the segment due to the instantaneous appliance mixture. To be confident that whole features of an appliance are collected in the segment, segments with more than one falling edge (e.g. two falling edge) have also been considered for appliance matching. In this situation the assumption is that the starting and ending edges of the segment are related and edges between them belong to other appliances and appeared because of simultaneous usage.

Another challenge is multi-state appliances, which have more than two edges. Selecting more than one falling edge is required for detecting this type of appliances. To make the problem more accessible for multi-state appliances, the method records more edges in an event segment. Therefore the method considers more than one candidate segment for each rising edge. The number of candidates

segment for each rising edge depends on number of edges in multi-label appliances. The distance between recognized segments and database will decide which candidate segment is really an appliance or it is just a part of waveform that is not related to any specific appliance.

#### **4.6.4 Segment Modification**

Segment modification simplifies appliance matching for single state and continues varying appliances when they are mixed with other appliances. Appliance mixtures occur when there are edge(s) between the related rising and falling edge. (Multi-state appliances are exempt from this step because they naturally have more than a single pair of edges.) For the segments that have the possibility to be multi-state appliance (i.e. their initial rising edge is similar to that of a multi-state appliance in the database), we keep a copy of the segment unchanged, and then proceed with segment modification.

If assume that only first and last power level changes in the candidate segment are related to each other, any unrelated rising or falling edges and their related values can be removed from the segment. Table 5 shows different scenarios when two appliances with a pair of edges can mix together and how their waveforms are modified. For example, when an appliance is working and another appliance start working the first scenario will happen. If the first falling edge becomes candidate to match the rising edge, there would be one extra unrelated rising edge and unrelated features start with the rising edge between events. The power demand can be reduced from the time of the extra rising edge



until we reach the falling edge which is close in value to that rising edge.

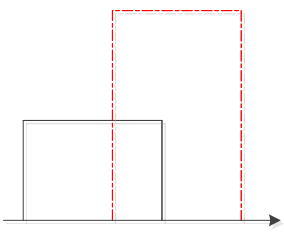
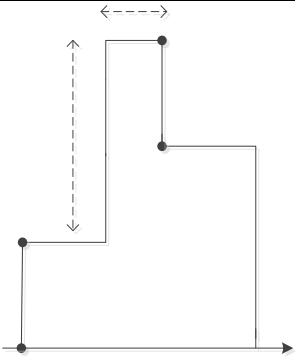

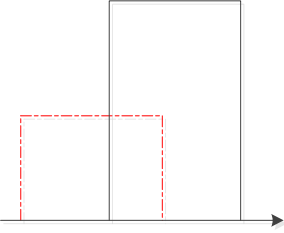
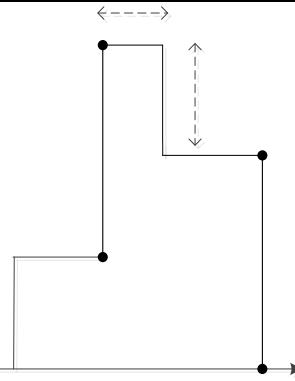
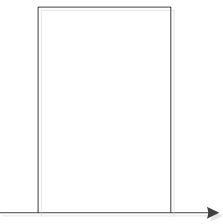
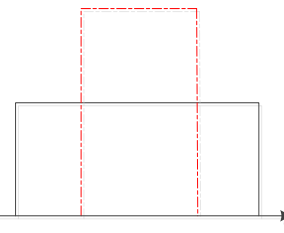
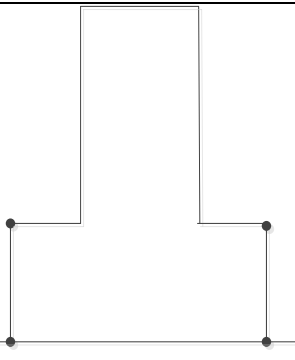
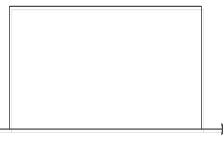
More complicated scenarios may happen especially when a repetitive appliance mixes with another appliance, or more than two large appliances work simultaneously; in such cases only the large appliance with higher power consumption is detectable. When talking about simplifying a mixture of appliances, the small appliances are not included in these procedures. Small appliances with small rising/falling edges are ignored in edge detection because their contribution to the overall classification error is small.

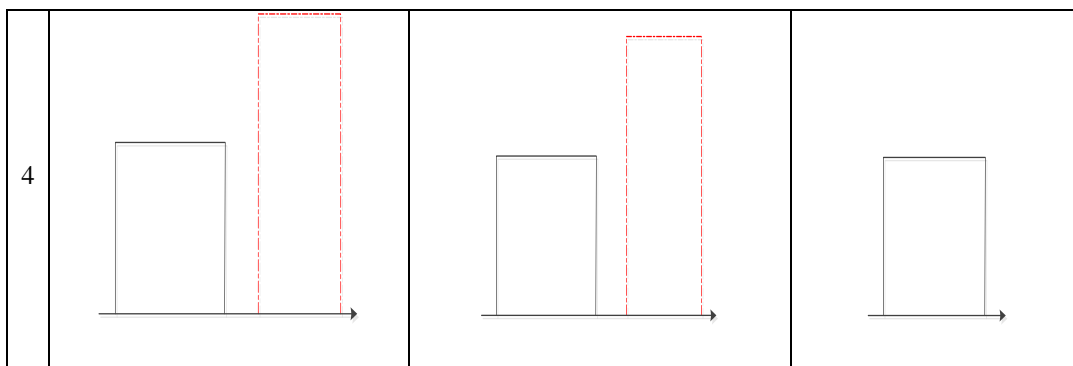
Vampire energy (appliances in standby mode) is another problem which causes offsets and vertical shifts in appliance waveforms, leading to errors in similarity measurement. In order to remove background energy effects, data in each segment undergoes a uniform downward vertical shift that reduces the minimum observed power demand in the segment to zero. This should remove the appliance signature distortion caused by vampire energy.

$$S_i^{new} = S^i - \min\{S^i\} \quad (32)$$

Modified segments with different lengths and numbers of events provide an opportunity to assign more than one label on each data point. Therefore the problem can be solved like a multi-label problem, it means for example in scenarios 1 and 2 in Table 5 the common region between two appliances will have labels for both appliances.

Table 5 Different Scenarios for single state appliances mixture and how to extract features

#	Two Appliances	Appliance Mixture waveform	Extracted Waveform
1			
2			
3			



#### 4.6.5 Database

One of the main issues in similarity search is to have a strong and generic database. If each appliance waveform in the database is general enough, database objects will have close similarity to all similar appliances in the dataset. The priority in building a database is to minimize redundancy as much as possible to reduce search time; however, all modes of operation of each appliance must also be stored in the database. It is not possible to have an ideal database in practice with exact similarity to all candidates, so we should relax the identification threshold to include all the cases; however it should be little to do not add much false identification. To prepare a signature database, appliance waveforms are collected ahead of time before implementing the method.

Collecting data for supervised methods is one of the challenges in nonintrusive load monitoring problem. Depends on the identification method, information about appliance waveform should be collected. Power consumption data for individual appliances can be collected by installing sub-meters inside the

home on each device or on circuits. Even temporarily installing sub-meters inside a house makes the method intrusive and defeats the whole purpose of NILM methods. Installing sub-meters is also quite costly, and thus cannot be done on a large scale.

Building a reference database is easier, but still requires customer help to determine the time of use of each appliance to extract appliance features from recorded power demand. However, without customer help registration is impossible. If the registration process of an appliance failed due to some reasons such as there being mixture of appliances during the registration process, it will be difficult to convince the customer to register the appliance again. For appliances such as clothes washer, we must also wait until customer uses the appliance, which may create a long delay. Another problem is registering automatic appliances such as refrigerators which do not have an ON/OFF switch. These automatic appliances can be registered and added to the database by signal experts.

The database for the synthetic datasets in this project has been built manually from the available appliance information. The database for REDD has been built using sub-meter measurements for each appliance/circuit, which are available in REDD. The number of repetitions of each appliance in the database is once the same as MIT research in [60].

#### **4.6.6 Similarity measurement**

The main step in similarity search method is finding the distance of the candidate

segments from waveforms in the reference database. Each query segment is compared against stored sequences in the database, and the sequence with the minimum distance to the query segment is identified. The distance between waveforms is measured with dynamic time warping.

#### 4.6.7 Decision Making

After identifying the stored segment with minimum distance to the query segment, the label of that matched segment is applied to the query segment – but only if the distance is smaller than a predetermined threshold. This threshold is used to cut down on the false positives, e.g. when a query segment only represents a part of the power waveform of an appliance. Figure 20 shows the database search and decision making procedure.

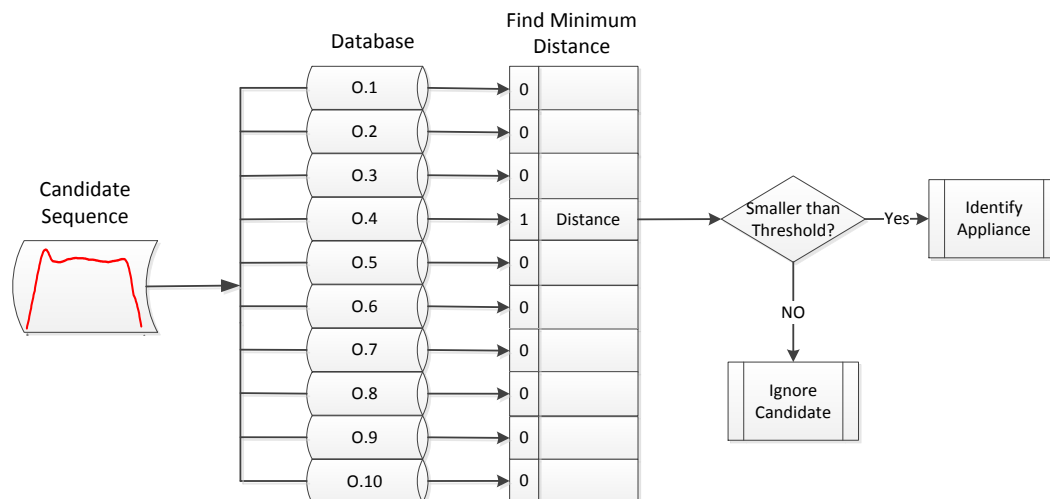


Figure 20 Sequence matching procedure

DTW is an efficient tool to identify appliances for matching with database, however when using DTW for similarity measure, the length of the sequences

cannot be used as a feature in identification procedure. The time duration of appliance usage would be ignored because the warping path length varies between  $\max(m, n)$  and  $m+n-1$  and does not reveal the length of the waveforms.

Time duration of an appliance is, however, an important distinguishing feature, especially for those that have a routine operation and changes in their duration are small. To add the effect of duration back into our similarity metric, the difference in lengths between matched segments has been added as a penalty factor to the calculated distance. The penalty factor is high for change in duration of fixed length appliances. The penalty factor function and related constants, as well as the decision threshold, are calculated heuristically from the available training data.

To identify features of appliances two measured signals are available: active power, and reactive power. These signals have been used separately for similarity comparison if available and then their results (distance) have been added. Decision about assigning label has been made based on the sum of both distances. Reactive power is especially important to distinguish between appliances which absorb reactive power and those that do not; for example resistive appliances only have active power and motor driven appliances have significant reactive power.

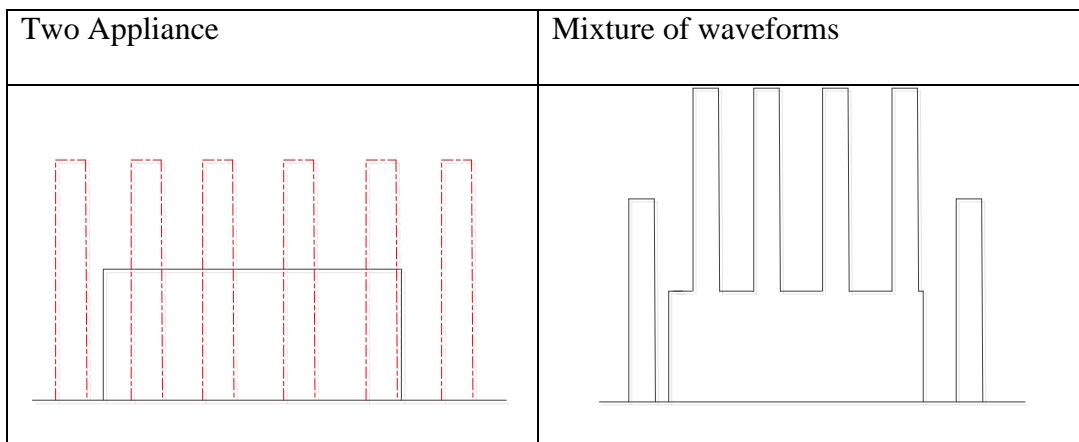
#### **4.6.8 Iterative Detection**

The objective of NILM is to detect all appliances in the measured power signal. However when a big appliance is mixed with other appliances, the signatures of

the small appliances can be overwhelmed by the bigger appliance and are not recognizable. It is one of the challenges in NILM: detecting big appliances is easily achievable but small appliances are often missed. The mixture of a single state appliance with an appliance with repetitive narrow pulses is another example of appliance mixture; it has been shown in Table 6. Repetitive waveforms could hide the single state appliance from detection.

If two or more appliances are mixed together and their features are not distinguishable, then the identification methods could not identify any of them. Preprocessing the segments before similarity search is one of the suggestion in this research which helps when two single state appliance are mixed together.

Table 6 Mixture of single appliance with pulsive appliances



The proposed idea is to remove appliance features when they are recognized in identification process. Removing properly detected appliance features provides the chance to detect underlying appliance features. The idea is to repeat the

identification process after removing detected appliances. In each identification step the dataset would be sparser than the previous step and fewer appliances exist. It would improve the overall performance of identification methods. The success of the iteration depends on performance of similarity search method, which should be able to assign right label on separated segments with minimum false positives and false negatives. There is no specific way to determine the number of iterations needed. For our simulated datasets which have just two appliances and the background power is unknown iteration is not applied, but for REDD the identification is repeated once after removing the detected appliances.

#### **4.7 Energy Error**

In addition to the standard measures of classification accuracy (true & false-positive rates, F-measure, etc.), the NILM field also employs an additional measure called the energy error. The idea is to measure how closely the energy consumption of an appliance matches the energy consumption assigned to that appliance by a NILM technique. Clearly, a smaller energy error indicates a more accurate identification method.

The energy error calculation is different for each proposed method. In multi-label classification, labels are assigned to each data point and each point could have several labels. The problem is how to determine the share of each label on energy consumption in each point. As the detailed energy consumption of each appliance is available, the average energy of each data point has been



calculated. Multiplying the number of labels in average energy consumption each appliance give us approximate energy consumption of each appliance in test data.

$$P_i = P_{ave,i} \times N_i \quad (33)$$

where  $P_i$  is total energy consumption of appliance  $i$ ,  $P_{ave,i}$  is its average power consumption, and  $N_i$  is number of detected appliance label  $i$ .

In the similarity search method, labels are assigned to each section of the signal. Adding up energy consumption of sequences with same label is the method used to calculate total energy consumption of each appliance. Energy consumption of each segment is equal to the area under its associated curve. Each sequence is just related to one specific appliance.

In the literature, the relative energy error is most commonly used to evaluate the performance of NILM methods on the REDD dataset. The available sub-meter data has been used as a ground-truth for error calculation. Relative energy error is given by:

$$error = \frac{|E_{predicted} - E_{actual}|}{E_{actual}} \quad (34)$$

## Chapter 5

### EXPERIMENTAL RESULTS

In this chapter we present our experimental results and evaluate NILM methods we have proposed. We contrast our various multi-label classification techniques against each other, as well as the similarity search method introduced in Chapter 4, and (in the case of the REDD datasets) the existing literature.

#### 5.1 Evaluation on Simulated Datasets

##### 5.1.1 Dataset 1

Table 7 shows the evaluation results of implementing multi-label classification on dataset 1 in the time domain. From Table 7, it can be understood that the best results in time domain are achieved when  $MLkNN$  is used.

Table 7 Evaluation result of multi-label classification in time domain on dataset 1

	Accuracy	micro F-measure	Macro F-measure
RAkEL	0.742	0.501	0.520
$MLkNN$	0.788	0.541	0.619

The results of Multi-label classification when the used features are wavelet coefficients are shown in Table 8. Comparing the results with Table 7 shows that

there is no improvement in classification performance when the data are transformed to the wavelet domain. Therefore between these two feature sets for this dataset, delay vectors are the best feature for  $MLkNN$ .

Table 8 Evaluation results of multi-label classification in wavelet domain on dataset 1,

	Accuracy	micro F-measure	Macro F-measure
RAkEL	0.796	0.133	0.412
$MLkNN$	0.747	0.470	0.587

The results of the similarity search on dataset 1 are presented in Table 9. The method appears to essentially match the performance of the time-domain  $MLkNN$  classifier.

Table 9 Evaluation results of Similarity Search on dataset 1

Dataset 1	micro F-measure	Macro F-measure
Similarity Search	0.611	0.602

### 5.1.2 Dataset 2

Dataset 2 has less background power compared to dataset 1, and thus we expect the results on this dataset to be superior for all techniques. Results from the time-

domain multi-label classifiers on dataset 2 are presented in Table 10; *RAkEL* appears to have performed better on this dataset.

Table 10 Evaluation of multi-label classification in time domain on dataset 2

	Accuracy	micro F-measure	Macro F-measure
<i>RAkEL</i>	0.957	0.890	0.873
<i>MLkNN</i>	0.813	0.649	0.540

The results for the wavelet-domain multi-label classifiers are presented in Table 11. In this dataset, the wavelet-domain classifiers were superior, and the wavelet-domain *RAkEL* was the best of the four.

Table 11 Evaluation results of multi-label classification in wavelet domain on dataset 2,

	Accuracy	micro F-measure	Macro F-measure
<i>RAkEL</i>	0.995	0.996	0.931
<i>MLkNN</i>	0.936	0.857	0.717

The similarity search method results are presented in Table 12. Similarity search performed better than time-domain multi-label classification, however it is not better than multi-label classification in the wavelet domain.

Table 12 Evaluation results of Similarity Search on dataset 2,

Dataset 2	micro F-measure	Macro F-measure
Similarity Search	0.921	0.925

Overall, the performance of all methods on dataset 2 (which is a mixture of two appliances) is good. For dataset 1, which is mixture of many appliances with only two of them labelled, the methods make many more errors. Both wavelet-domain and time-domain classifiers have been superior to the other on one of the datasets, and the similarity search has always been better than at least one of them. At this point, we cannot rule out any of our methods, and so we will proceed with an experimental evaluation of all three on the REDD dataset.

## 5.2 Evaluation on REDD

House 1 and house 3 have been selected for the tests because data are collected for a longer period of time for these houses. While the collection periods and appliances in the houses are different, we feel the most important difference is that both active and reactive power can be determined for house 3, while only active power is available for house 1. (The house 3 data includes current and voltage measurements that are used to calculate reactive power.)

### 5.2.1 REDD, House 3

Evaluations of the time- and wavelet-domain multi-label classification methods on REDD house 3 are shown in Table 13 and Table 14, respectively. The evaluations show that there is not much difference in values when using either *RAkEL* or *MLkNN*. The performance of the classification methods is by some measures superior in the wavelet domain and by others superior in the time domain. It is impossible to select one as the best method just from these overall results.

Table 13 Evaluation results of Multi-label classification in Time domain on  
REDD, house 3

	Accuracy	micro F-measure	Macro F-measure
<i>RAkEL</i>	0.922	0.923	0.492
<i>MLkNN</i>	0.915	0.921	0.471

Table 14 Evaluation results of Multi-label classification in Wavelet domain on  
REDD, house 3

	Accuracy	micro F-measure	Macro F-measure
<i>RAkEL</i>	0.96	0.959	0.455
<i>MLkNN</i>	0.951	0.943	0.472

Our similarity search results are presented in Table 15. Again, different measures give quite different results in the overall evaluation, with similarity search being superior to all others on the macro-averaged F-measure. These mixed results indicate that we need to examine the classification results in more detail. We will examine results for each individual appliance in Section 5.4.

Table 15 Evaluation results of Similarity Search on REDD House 3

REDD 3	micro F-measure	Macro F-measure
Similarity Search	0.50	0.541

### 5.2.2 REDD, House 1

The proposed identification methods have been tested on seven appliances in House 1 of REDD. This is the only one of our datasets that lacks reactive power measurements, which made appliance detection more challenging. The results in Table 16 and Table 17 show the overall performance of multi-label classification in the time domain and wavelet domain, respectively. From the results, it can be concluded that  $MLkNN$  classification in time domain has better performance among multi-label classification methods.

Table 16 Evaluation results of Multi-label classification in Time domain on  
REDD House 1

	Accuracy	micro F-measure	Macro F-measure
RAkEL	0.495	0.587	0.393
MLkNN	0.788	0.776	0.619

Table 17 Evaluation results of Multi-label classification in Wavelet on REDD,  
House 1

	Accuracy	micro F-measure	Macro F-measure
RAkEL	0.606	0.763	0.430
MLkNN	0.652	0.597	0.524

The performance of our similarity search method has also been tested on house 1; these results are shown in Table 18. Comparing the results shows that MLkNN classification in time domain has better performance in this house.

Table 18 Evaluation results of Similarity Search on REDD, House 1

REDD 1	micro F-measure	Macro F-measure
Similarity Search	0.397	0.502

As the overall results for REDD shows, the performance of the proposed methods on real datasets is very different and complicated. It is even hard to



select one method as the best one for a dataset because different evaluation measures yield different conclusions. Again, it seems necessary to investigate our results in greater depth. However, we will first inquire whether creating local models based on a cluster analysis would significantly improve these results.

### **5.3 Clustering based Classification**

Impact of clustering to improve performance of multi-label classification has been evaluated in NILM problem. The idea is to use a clustering method to split data into small groups with (hopefully) similar appliances within the groups, and then build a local model for each cluster. In this section the performance of EM clustering along with multi-label classification in the time domain is evaluated on our four datasets.

Table 19 shows our evaluation of the proposed clustering method on dataset 1 when the number of clusters varies from 2 to 15. From Table 19, it can be understood that the best result on dataset 1 is gained when ML $k$ NN has been applied on 11 clusters.

Table 19 Evaluation results of multi-label classification on dataset 1

Cluster #	RAkEL			MLkNN		
	Accuracy	m. F-measure	M. F-measure	Accuracy	m. F-measure	M. F-measure
2	0.742	0.463	0.468	0.770	0.481	0.537
3	0.729	0.447	0.481	0.784	0.516	0.558
4	0.727	0.415	0.453	0.783	0.517	0.560
5	0.737	0.446	0.504	0.782	0.504	0.541
6	0.748	0.451	0.507	0.790	0.513	0.543
7	0.738	0.451	0.473	0.785	0.506	0.524
8	0.747	0.446	0.483	0.787	0.506	0.526
9	0.746	0.450	0.498	0.789	0.505	0.529
10	0.7612	0.48526	0.51375	0.787	0.504	0.531
11	0.745	0.432	0.483	0.79477	0.51484	0.53189
12	0.742	0.432	0.457	0.791	0.506	0.508
13	0.749	0.425	0.449	0.789	0.503	0.507
14	0.754	0.430	0.471	0.790	0.506	0.505
15	0.737	0.406	0.474	0.789	0.505	0.505

However, when we compare the clustering output with the results of the methods without clustering in the previous section, clustering does not uniformly improve the results for all measures. Once again, the results are mixed.

Table 20 Evaluation results of multi-label classification on dataset 2

Cluster #	RAkEL			MLkNN		
	Accuracy	m. F-measure	M. F-measure	Accuracy	m. F-measure	M. F-measure
2	0.912	0.810	0.849	0.729	0.555	0.520
3	0.944	0.869	0.875	0.728	0.545	0.544
4	0.936	0.853	0.869	0.724	0.557	0.513
5	0.911	0.807	0.831	0.758	0.542	0.511

Evaluation results of clustering on dataset 2 in Table 20 show that RAkEL has the best results on 3 clusters. However, these results are inferior to the results obtained without clustering.

The results for REDD house 3 are shown in Table 21. Results demonstrate that when training multi-label classifier in each cluster, the best results are obtained when there are just 5 groups. There is no difference in using RAkEL and MLkNN on this dataset and both of them have similar performances. Comparing the results with Table 13 shows that clustering has again not improved the performance of our methods.

Table 21 Evaluation results of classification method on House 3, REDD

Cluster #	RAkEL			MLkNN		
	Accuracy	micro F-measure	Macro F-measure	Accuracy	micro F-measure	Macro F-measure
5	0.908	0.911	0.442	0.905	0.909	0.412
10	0.846	0.865	0.395	0.827	0.853	0.358
15	0.826	0.849	0.337	0.853	0.870	0.352
20	0.812	0.841	0.326	0.876	0.884	0.348
25	0.824	0.848	0.3690	0.864	0.877	0.340
30	0.791	0.824	0.320	0.872	0.882	0.369
35	0.879	0.876	0.353	0.887	0.881	0.352
40	0.898	0.901	0.399	0.897	0.894	0.402
45	0.895	0.894	0.348	0.901	0.902	0.410
50	0.865	0.877	0.373	0.884	0.892	0.406
55	0.897	0.896	0.386	0.897	0.895	0.366
60	0.897	0.895	0.366	0.905	0.905	0.404
65	0.902	0.903	0.377	0.904	0.902	0.359
70	0.887	0.887	0.352	0.884	0.890	0.366
75	0.855	0.870	0.335	0.860	0.865	0.318
80	0.892	0.888	0.342	0.904	0.899	0.377
85	0.885	0.890	0.355	0.897	0.893	0.375
90	0.897	0.899	0.362	0.903	0.903	0.412
95	0.894	0.888	0.386	0.880	0.877	0.381
100	0.849	0.866	0.308	0.883	0.892	0.385

Table 22 shows the evaluation of multi-label classification along with clustering on REDD, house 1. Again, there is no improvement in results when using clustering.

Table 22 Evaluation multi-label classification along with clustering on REDD,  
House 1

Cluster #	RAkEL			MLkNN		
	Accuracy	micro F-measure	Macro F-measure	Accuracy	micro F-measure	Macro F-measure
5	0.744	0.807	0.510	0.726	0.733	0.554
10	0.741	0.793	0.508	0.740	0.758	0.559
15	0.735	0.790	0.588	0.756	0.768	0.570
20	0.733	0.788	0.574	0.756	0.746	0.571
25	0.734	0.787	0.562	0.748	0.774	0.570
30	0.745	0.804	0.569	0.752	0.761	0.573
35	0.739	0.738	0.517	0.763	0.764	0.557
40	0.748	0.804	0.585	0.757	0.766	0.576
45	0.733	0.786	0.549	0.764	0.753	0.571
50	0.743	0.794	0.608	0.764	0.766	0.572
55	0.739	0.789	0.598	0.759	0.772	0.557
60	0.751	0.804	0.602	0.767	0.787	0.566
65	0.746	0.796	0.578	0.763	0.747	0.554
70	0.746	0.796	0.605	0.763	0.748	0.557
75	0.754	0.808	0.584	0.762	0.762	0.547
80	0.743	0.742	0.550	0.762	0.747	0.528
85	0.753	0.753	0.568	0.761	0.745	0.553
90	0.750	0.745	0.567	0.760	0.742	0.528
95	0.752	0.752	0.575	0.763	0.745	0.554
100	0.747	0.738	0.560	0.759	0.741	0.537

## 5.4 Evaluation of the proposed methods on each appliance

In this section, for better understanding of the proposed methods' performance, we analyze our experimental results at the level of individual appliances. We are interested in determining which methods accurately detect more appliances, and especially the large appliances that consume the most power.

### 5.4.1 Simulated Datasets

Dataset 1 and 2 have just two registered appliances: Refrigerator and microwave. The performance of our methods on the refrigerator and microwave in dataset 1 has been shown in Table 23 and Table 24, respectively. It can be seen that the similarity search has better overall performance. Although RAKEL is the best algorithm in detecting the microwave in wavelet domain, it failed in detecting the refrigerator.

Table 23 Performance of identification methods for detecting Refrigerator in

Dataset 1

Features	Delay Embedding		Wavelet		Similarity Search
	RAkEL	MLkNN	RAkEL	MLkNN	
Precision	0.43	0.49	0	0.4	0.79
Recall	0.6	0.56	0	0.5	0.54
F-Measure	0.5	0.52	0	0.44	0.64
Specificity	0.8	0.86	1	0.81	0.89

Table 24 Performance of identification methods for detecting Microwave in  
Dataset 1

Features	Delay Embedding		Wavelet		
Method	RAkEL	MLkNN	RAkEL	MLkNN	Similarity Search
Precision	0.43	0.72	1	0.71	0.63
Recall	0.72	0.71	0.7	0.75	0.79
F-Measure	0.54	0.71	0.82	0.73	0.7
Specificity	0.98	0.99	1	0.99	1

Our appliance-level results for dataset 2 in are presented in Table 25 and Table 26. For this dataset, RAkEL classification in the wavelet domain is the best overall.

Table 25 Performance of identification methods for detecting Refrigerator in  
Dataset 2

Features	Delay Embedding		Wavelet		
Method	RAkEL	MLkNN	RAkEL	MLkNN	Similarity Search
Precision	0.88	0.63	0.99	1.0	0.95
Recall	0.91	0.97	1.0	1.0	0.9
F-Measure	0.89	0.77	1.0	1.0	0.93
Specificity	0.97	0.87	1.0	1.0	0.98



Table 26 Performance of identification methods for detecting Microwave in  
Dataset 2

Features	Delay Embedding		Wavelet		
Method	RAkEL	MLkNN	RAkEL	MLkNN	Similarity Search
Precision	1.0	0.13	1.0	0.19	0.83
Recall	0.72	0.85	0.75	0.81	0.95
F-Measure	0.84	0.23	0.85	0.31	0.89
Specificity	1.0	0.9	1.0	0.94	1.0

#### 5.4.2 REDD, House 3

To analyze the proposed methods, the performance of the time-domain classifiers, wavelet-domain classifiers, and similarity search for each appliance in REDD, house 3 is been shown in Table 27, Table 28, and Table 29, respectively.

Table 27 Multi label classification performance on REDD, House 3 in Time

Domain

	RAkEL				MLkNN			
	Precision	Recall	F-measure	Specificity	Precision	Recall	F-measure	Specificity
Electronics	1.0	1.0	1.0	0	1.0	1.0	1.0	0
Furnace	0.001	0.308	0.001	0.989	0	0	0	0.997
Washer dryer	0.997	0.902	0.947	1.0	0.998	0.981	0.989	1.0
Microwave	0.921	0.237	0.377	1.0	0.612	0.645	0.628	0.998
Bath GFI	0.468	0.947	0.627	0.992	0.294	0.802	0.430	0.986
Kitchen outlet	0.733	0.009	0.017	1.0	0.587	0.075	0.133	0.996

Table 28 Multi label classification performance on REDD, House 3 in Wavelet

domain

	RAkEL				MLkNN			
	Precision	Recall	F-measure	Specificity	Precision	Recall	F-measure	Specificity
Electronics	1.0	1.0	1.0	0	1.0	1.0	1.0	0
Furnace	0	0	0	1.0	0	0	0	0.99
Washer/Dryer	0.94	1.0	0.97	1.0	0.99	0.96	0.98	1.0
Microwave	0	0	0	1.0	0.08	0.71	0.15	0.96
Bath GFI	0.62	0.98	0.76	1.0	0.44	0.89	0.59	0.99
Kitchen Outlet	0	0	0	1.0	0.85	0.06	0.12	1.0

Table 29 DTW Similarity search result on REDD, House 3

Appliance	Precision	Recall	F-measure	Specificity
Electronics	0.155	0.848	0.261	1.0
Furnace	0.866	0.659	0.748	1.0
Washer/Dryer	0.962	0.855	0.905	0.971
Microwave	0.877	0.798	0.835	0.909
Bath GFI	0.80	0.163	0.27	0.983
Kitchen Outlet	0.232	0.22	0.226	0.969

As is obvious, performance of the methods for different appliances is different and it is hard to select one method as the best method. From the overall results in Section 5.2, macro f-measure of similarity search method which is the average of each appliance f-measure is better than other methods but its micro f-measure is not the best. Exploring the appliances in detail indicates the reason. The Electronics class is the main reason that micro f-measure of similarity search is much less than the classification method. The Electronics appliance/circuit is ON most of the time and has no unique signature most of the time, which confounds the similarity search; the classification methods seem more robust to the irregularity of this class.

The Furnace is a distinguishable appliance in power signal but performance of the multi-label classifiers in this method is very low and the method has essentially failed in detecting this appliance. Looking into the house 3 dataset, we found that measurements of the houses in REDD was done in the beginning of summer, therefore furnace rarely appears in the dataset. Furthermore

most of those few occurrences are in the early (training) portion of the dataset; during the period covered by the test set, the furnace has only been used for a few minutes.

Microwave and Bath gfi are very similar to each other and their main difference is the reactive power of microwave. The microwave is always mixed with other appliances, causing the classification algorithm to identify the mixed appliances also as the microwave during testing. The kitchen outlet is a circuit level measurement which includes three different appliances all with one label. Having a single label for several appliances with different distribution in the dataset has led to a poor performance.

### **5.4.3 REDD, House 1**

Performance of multi-label classification in time and wavelet domain, and the similarity search method on REDD, house 1 are presented in Table 30, Table 31 and Table 32 respectively. As it was expected from overall performance of the proposed methods on house 1,  $MLkNN$  has better performance in appliance detection compared to the others, except for the microwave.

Table 30 Multi label classification performance on REDD House 1 in time domain

	RAkEL				MLkNN			
	Precision	Recall	F-measure	Specificity	Precision	Recall	F-measure	Specificity
Oven	0	0	0	1.0	0.64	0.21	0.31	1.0
Refrigerator	0.89	0.95	0.92	0.96	0.94	0.95	0.94	0.98
Light	0.58	0.29	0.39	0.69	0.92	0.69	0.79	0.91
Microwave	0.51	0.54	0.52	0.99	0.05	0.55	0.09	0.87
Bath GFI	0	0	0	1.0	0.44	0.66	0.53	1.0
Outlet	0	0	0	1.0	0.71	0.84	0.77	1.0
Washer	0.86	0.99	0.92	1.0	0.94	0.85	0.90	1.0

Table 31 Multi label classification performance on REDD House 1 in Wavelet domain

	RAkEL				MLkNN			
	Precision	Recall	F-measure	Specificity	Precision	Recall	F-measure	Specificity
Oven	0.23	0.01	0.01	1.0	0.27	0.01	0.01	1.0
Refrigerator	0.92	0.94	0.93	0.97	0.49	0.95	0.65	0.64
Light	0.57	0.99	0.72	0.03	0.84	0.42	0.56	0.90
Microwave	0.48	0.62	0.54	0.99	0.54	0.54	0.54	0.99
Bath GFI	0	0	0	1.0	0.33	0.47	0.39	1.0
Outlet	0	0	0	1.0	0.85	0.57	0.68	1.0
Washer	0.67	0.99	0.80	0.99	0.73	0.99	0.84	1.0

Table 32 Similarity Search Performance on House 1

	Precision	Recall	F-measure	Specificity
Oven	0.228	0.236	0.232	0.993
Refrigerator	0.994	0.759	0.861	0.923
Light	0.587	0.009	0.018	0.411
Microwave	0.680	0.523	0.591	0.995
Bath GFI	0.981	0.222	0.362	0.997
Outlet	0.671	0.592	0.629	0.998
Washer	0.743	0.918	0.822	0.999

In the simulated datasets our conclusion was that similarity search is a useful method, and competitive with the multi-label classifiers. However, on REDD house 1 we can see that  $MLkNN$  has better performance in most of the appliances. The main difference of this dataset with the others is lack of reactive power data in this dataset. It means half of the features of house 1 are missing compared to the other datasets. Reactive power is important especially when a resistive appliance without reactive power is mixed with non-resistive appliances and features are not distinguishable.

## 5.5 Energy Error

We now compute the energy error for the various appliances in each dataset for each of our proposed methods. We present the results for dataset 1 in Table 33.

The smallest error for Refrigerator is achieved with  $MLkNN$  in time domain and for Microwave with similarity search.

Table 33 Energy error on Dataset 1

Energy Error	Delay Embedding		Wavelet		Similarity Search
	$RAkEL$	$MLkNN$	$RAkEL$	$MLkNN$	
Refrigerator	0.321	0.061	1.0	0.152	0.448
Microwave	0.976	0.163	0.199	0.214	0.010

Energy error results in Table 34 shows that for dataset 2  $RAkEL$  error in wavelet domain is smaller than other methods though similarity search error is the second best method.

Table 34 Energy error on Dataset 2

Energy Error	Delay Embedding		Wavelet		Similarity Search
	$RAkEL$	$MLkNN$	$RAkEL$	$MLkNN$	
Refrigerator	0.084	0.614	0.040	0.033	0.07
Microwave	0.174	6.281	0.170	3.374	0.025

Table 35 shows the energy error evaluation on house 3. It can be seen that best results are different for appliance to appliance and it is hard to select one method as the best one. However it can concluded from data that multi-label  $k$ -NN ( $MLkNN$ ) in time domain and similarity search have better results.

Table 35 Energy Error on REDD, House 3

Appliance	Delay Embedding		Wavelet		Similarity Search
	RAkEL	MLkNN	RAkEL	MLkNN	
Electronics	0.009	0.009	0.009	0.009	0.949
Furnace	0.954	0.456	0.322	0.469	0.412
Washer/Dryer	0.105	0.027	0.050	0.046	0.240
Microwave	0.759	0.012	1.0	6.267	0.306
Bath GFI	0.344	0.813	0.011	0.305	0.244
Kitchen Outlet	0.983	0.824	1.000	0.898	0.634

The results of house 1 are shown in Table 36. Again, it is hard to select one method or feature space as the best one for all appliances. Although performance of MLkNN on feature space for microwave is disappointing, overall it has better performance among the methods.

Table 36 Energy Error on REDD, House 1

	Delay Embedding		Wavelet		Similarity Search
	RAkEL	MLkNN	RAkEL	MLkNN	
Oven	1.0	0.607	0.937	0.956	0.668
Refrigerator	0.070	0.019	0.059	1.023	0.084
Light	0.437	0.149	0.856	0.472	0.956
Microwave	0.031	10.413	0.434	0.110	0.116
Bath_GFI	1.00	0.443	1.0	0.510	0.684
Outlet	1.00	0.193	1.0	0.366	0.099
Washer	0.145	0.093	0.025	0.112	0.351



If an appliance has different mode of operation but only part of them which have high power level have been registered for similarity matching, then missing low power modes will appeared in point to point evaluation but will not be significant in energy error. These not-registered points are not important for energy breakdown purposes, but if the goal is control the overall demand, their registration is necessary.

One important issue to validate an appliance identification method is defining an acceptable value (i.e. %5) for maximum error. High error in energy calculation will disappoint customer about the results.

## **5.6 Comparison with Published Methods on REDD**

We now compare our proposed methods with the published methods that have been evaluated on REDD. Table 37 shows the results of implementing Factorial HMM to identify appliances in a home [60] AFAMAP convex optimization was used to estimate the appliances (states) from the model. Similarity search which has better performance among our evaluated method on REDD, house 3 is compared with FHMM. In comparison to [60] our similarity search is better overall.

Table 37 REDD, house 3 comparisons

Appliance	Proposed Similarity matching			Factorial HMM [60]		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Electronics	0.155	0.848	0.261	0.416	0.008	0.016
Furnace	0.866	0.659	0.748	0.917	0.708	0.799
Washer/Dryer	0.962	0.855	0.905	0.988	0.736	0.844
Microwave	0.877	0.798	0.835	0.975	0.661	0.788
Bath GFI	0.800	0.163	0.270	0.827	0.708	0.763
Kitchen Outlet	0.232	0.220	0.226	0.452	0.16	0.236

The method in [62] decomposes a set of appliance models into groups of appliances which have overlap in power signal and then disaggregates each group by Viterbi algorithm. This method is evaluated on house 1 of REDD and they have compared the F-measure of their method with Bayesian classifier. The results of F-measure of their method and our proposed method are shown in Table 38. They have implemented their method on certain appliances so just the common results are shown in this table. Their results are superior overall (although not uniformly) to our proposed methods.

Table 38 REDD, house 1 comparison

Appliance	Proposed MLkNN in time domain	Bayesian method in [62]	Proposed method in [62]
	F-measure	F-measure	F-measure
Oven	0.31	0.8	0.908
Refrigerator	0.94	0.859	0.831
Microwave	0.09	0.775	0.899
Bath GFI	0.53	0.753	0.927
Outlet	0.77	0.409	0.84

HMM along with EM clustering and Viterbi algorithm has been used to identify a few appliances from REDD in [64] and [63]. Their methods are unsupervised and a general database has been used to label the appliances after detection. The idea of their research is to develop the Factorial HMM method in [60]. They have evaluated their method by calculating the energy consumption error. Table 39 shows the results of the two published articles. For these three appliances RAkEL in time domain has better performance.

Table 39 Comparison of the energy error of proposed method with methods in [64] and [63]

	Proposed method	Results in [64]			Results in [63]		
Appliance	RAkEL	No training	Aggregate training	Sub - metered training	No training	Aggregate training	Sub-metered training
Refrigerator	0.07	0.550	0.150	0.140	0.380	0.210	0.550
Washer/Dryer	0.145	4.280	0.280	0.240	34.690	0.550	0.710
Microwave	0.031	0.540	0.220	0.100	0.630	0.530	0.380

## Chapter 6

### CONCLUSION

In this thesis two NILM methods are proposed: Multi-label classification, which can identify appliances at each sample instant, and a similarity matching method that identifies each candidate subsequence of data. Although the similarity search method is implemented on wavelet coefficients features, the proposed multi-label classification is evaluated in both time domain and wavelet domain. Feature sets in time domain are constructed using delay coordinate embedding.

In future work we will further develop the similarity search method. There are several challenges in practical NILM applications which the proposed database matching method can help to solve. Developing an unsupervised method with dynamic time warping measure to cluster the divided segments is one possible means to automatically build a signature database. Close to real time load identification and online feedback to the customer could also be accomplished with the proposed method. It can define new application for NILM. We could also train multi-label classifiers with simulated datasets similar to the intended real-world dataset.

## Chapter 7

### REFERENCES

- [1] G. o. Canada, "Energy Statistics Handbook," vol. Q4, ed, 2009, pp. Table 8.7-1.
- [2] S. Darby, "The effectiveness of feedback on energy consumption," Oxford, UK2006.
- [3] B. J. Fogg, *Persuasive Technology: Using Computers to Change what We Think and Do*. Amsterdam, Netherlands: Morgan Kaufmann Publishers, 2003.
- [4] A. Capasso, W. Grattieri, R. Lamedica, and A. Prudenzi, "A bottom-up approach to residential load modeling," *IEEE Transactions on Power Systems*, vol. 9, no. 2, pp. 957-964, 1994.
- [5] G. Kats and E. Capital, *Green building costs and financial benefits*. Westborough, MA: Massachusetts Technology Collaborative Boston, MA, 2003.
- [6] 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)*, San Diego, CA, 2011.
- [7] G. W. Hart, "Nonintrusive appliance load monitoring," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870-1891, 1992.
- [8] D. Ming, P. C. M. Meira, X. Wilsun, and W. Freitas, "An Event Window Based Load Monitoring Technique for Smart Meters," *IEEE Transactions on Smart Grid*, vol. 3, no. 2, pp. 787-796, 2012.
- [9] N. Barry and E. McQuade, "Temperature control using integer-cycle binary rate modulation of the AC mains," *IEEE Transactions on Industry Applications*, vol. 31, no. 5, pp. 965-969, 1995.
- [10] J. Liang, S. Ng, G. Kendall, and J. Cheng, "Load signature study—Part I: Basic concept, structure, and methodology," *IEEE Transactions on Power Delivery*, vol. 25, no. 2, pp. 551-560, 2010.
- [11] J. Liang, S. K. Ng, G. Kendall, and J. W. Cheng, "Load Signature Study—Part II: Disaggregation Framework, Simulation, and Applications," *IEEE Transactions on Power Delivery*, vol. 25, no. 2, pp. 561-569, 2010.
- [12] A. Marchiori, D. Hakkarinen, Q. Han, and L. Earle, "Circuit-Level Load Monitoring for Household Energy Management," *IEEE Pervasive Computing*, vol. 10, no. 1, pp. 40-48, Jan-Mar 2011.
- [13] M. Akbar and Z. A. Khan, "Modified nonintrusive appliance load monitoring for nonlinear devices," in *Proceedings of the 11th Ieee International MultitopicConference (Inmic 2007)*, Lahore, Pakistan, 2007, pp. 69-73.

- [14] D. Jung, A. Savvides, and A. Bamis, "Tracking Appliance Usage Information in Residential Settings Using Off-the-Shelf Low-Frequency Meters," in *2012 49th Acm/Edac/Ieee Design Automation Conference (Dac)*, San Francisco, CA, 2012, pp. 163-168.
- [15] K. D. Lee, S. B. Leeb, L. K. Norford, P. R. Armstrong, J. Holloway, and S. R. Shaw, "Estimation of variable-speed-drive power consumption from harmonic content," *IEEE Transactions on Energy Conversion*, vol. 20, no. 3, pp. 566-574, Sep 2005.
- [16] J. Paris, Z. Remscrim, L. K. Douglas, S. B. Leeb, R. W. Cox, M. S. T. Galvin, M. S. G. Coe, and L. J. R. Haag, *Scalability of Non-Intrusive Load Monitoring for Shipboard Applications*. Cambridge, MA: Sea Grant College Program, Massachusetts Institute of Technology, 2009.
- [17] R. W. Cox, P. L. Bennett, 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*, Arlington, Virginia, 2007, pp. 523-530.
- [18] M. E. Berges, E. Goldman, H. S. Matthews, and L. Soibelman, "Enhancing electricity audits in residential buildings with nonintrusive load monitoring," *Journal of industrial ecology*, vol. 14, no. 5, pp. 844-858, 2010.
- [19] M. Fitta, S. Biza, M. Lehtonen, T. Nieminen, and G. Jacucci, "Exploring Techniques for Monitoring Electric Power Consumption in Households," in *UBICOMM 2010, The Fourth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies*, Florence, Italy, 2010, pp. 471-477.
- [20] M. Figueiredo, A. de Almeida, and B. Ribeiro, "Home electrical signal disaggregation for non-intrusive load monitoring (NILM) systems," *Neurocomputing*, vol. 96, pp. 66-73, Nov 1 2012.
- [21] M. Weiss, A. Helfenstein, F. Mattern, and T. Staake, "Leveraging smart meter data to recognize home appliances," in *2012 IEEE International Conference on Pervasive Computing and Communications (Percom)*, Lugano, Switzerland, 2012, pp. 190-197.
- [22] M. Berges, E. Goldman, H. S. Matthews, L. Soibelman, and K. Anderson, "User-Centered Nonintrusive Electricity Load Monitoring for Residential Buildings," *Journal of Computing in Civil Engineering*, vol. 25, no. 6, pp. 471-480, Nov-Dec 2011.
- [23] 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 *UbiComp 2010: Proceedings of the 2010 Acm Conference on Ubiquitous Computing*, Copenhagen, Denmark, 2010, pp. 139-148.
- [24] M. B. Figueiredo, A. De Almeida, and B. Ribeiro, "An experimental study on electrical signature identification of non-intrusive load monitoring (nilm) systems," in *Adaptive and Natural Computing Algorithms*, ed: Springer, 2011, pp. 31-40.

- [25] K. Chahine, K. E. Drissi, C. Pasquier, K. Kerroum, C. Faure, T. Jouannet, and M. Michou, "Electric Load Disaggregation in Smart Metering Using a Novel Feature Extraction Method and Supervised Classification," in *Impact of Integrated Clean Energy on the Future of the Mediterranean Environment*, Beirut, Lebanon, 2011, pp. 627-632.
- [26] A. Prudenzi, "A neuron nets based procedure for identifying domestic appliances pattern-of-use from energy recordings at meter panel.," in *2002 IEEE Power Engineering Society Winter Meeting*, New York, NY, 2002, pp. 941-946.
- [27] H. H. Chang, K. L. Chen, Y. P. Tsai, and W. J. Lee, "A New Measurement Method for Power Signatures of Nonintrusive Demand Monitoring and Load Identification," *IEEE Transactions on Industry Applications*, vol. 48, no. 2, pp. 764-771, Mar-Apr 2012.
- [28] Y.-C. Su, K.-L. Lian, and H.-H. Chang, "Feature Selection of Non-intrusive Load Monitoring System Using STFT and Wavelet Transform," in *2011 IEEE 8th International Conference on e-Business Engineering (ICEBE)*, Beijing, China, 2011, pp. 293-298.
- [29] D. Srinivasan, W. S. Ng, and A. C. Liew, "Neural-network-based signature recognition for harmonic source identification," *IEEE Transactions on Power Delivery*, vol. 21, no. 1, pp. 398-405, Jan 2006.
- [30] A. G. Ruzzelli, C. Nicolas, A. Schoofs, and G. M. O'Hare, "Real-time recognition and profiling of appliances through a single electricity sensor," in *2010 7th Annual IEEE Communications Society Conference on Sensor Mesh and Ad Hoc Communications and Networks (SECON)*, Boston, Massachusetts, 2010, pp. 1-9.
- [31] Y.-H. Lin and M.-S. Tsai, "A novel feature extraction method for the development of nonintrusive load monitoring system based on BP-ANN," in *2010 International Symposium on Computer Communication Control and Automation (3CA)*, Tainan, Taiwan, 2010, pp. 215-218.
- [32] J. G. Roos, I. E. Lane, E. C. Botha, and G. P. Hancke, "Using Neural Networks for Non-Intrusive Monitoring of Industrial Electrical Loads," in *Advanced Technologies in I & M, 10th Anniversary, Imtc/94*, 1994, pp. 1115-1118.
- [33] J. Froehlich, E. Larson, S. Gupta, G. Cohn, M. S. Reynolds, and S. N. Patel, "Disaggregated End-Use Energy Sensing for the Smart Grid," *IEEE Pervasive Computing*, vol. 10, no. 1, pp. 28-39, Jan-Mar 2011.
- [34] 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 - (Nominated for the best paper award)," in *UbiComp 2007: Ubiquitous Computing*, Innsbruck, Austria, 2007, pp. 271-288.
- [35] Y.-H. Lin and M.-S. Tsai, "Applications of hierarchical support vector machines for identifying load operation in nonintrusive load monitoring systems," in *2011 9th World Congress on Intelligent Control and Automation (WCICA)*, Taipei, Taiwan, 2011, pp. 688-693.



- [36] J. Z. Kolter, S. Batra, and A. Y. Ng, "Energy disaggregation via discriminative sparse coding," in *Neural Information Processing Systems*, Vancouver, 2010, pp. 1153-1161.
- [37] Y. H. Lin, M. S. Tsai, and C. S. Chen, "Applications of Fuzzy Classification with Fuzzy C-Means Clustering and Optimization Strategies for Load Identification in NILM Systems," in *IEEE International Conference on Fuzzy Systems (Fuzz 2011)*, Grand Hyatt Taipei, Taiwan, 2011, pp. 859-866.
- [38] Y. H. Lin and M. S. Tsai, "Application of Neuro-Fuzzy Pattern Recognition for Non-intrusive Appliance Load Monitoring in Electricity Energy Conservation," *2012 Ieee International Conference on Fuzzy Systems (Fuzz-Ieee)*, 2012.
- [39] M. L. Marceau and R. Zmeureanu, "Nonintrusive load disaggregation computer program to estimate the energy consumption of major end uses in residential buildings," *Energy Conversion and Management*, vol. 41, no. 13, pp. 1389-1403, Sep 2000.
- [40] 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 *APEC 2006: Twenty-First Annual IEEE Applied Power Electronics Conference and Exposition*, Dallas, Texas, 2006, pp. 1751-1757.
- [41] S. Drenker and A. Kader, "Nonintrusive monitoring of electric loads," *IEEE Computer Applications in Power*, vol. 12, no. 4, pp. 47-51, Oct 1999.
- [42] F. Sultanem, "Using Appliance Signatures for Monitoring Residential Loads at Meter Panel Level," *IEEE Transactions on Power Delivery*, vol. 6, no. 4, pp. 1380-1385, Oct 1991.
- [43] R. Breed and J. Delport, "Non-intrusive load monitoring of residential appliances," in *Proceedings of the 7th Domestic Use of Energy Conference*, 2001.
- [44] H. Murata and T. Onoda, "Applying Kernel based Subspace Classification to a non-intrusive monitoring for household electric appliances," in *Artificial Neural Networks, (ICANN 2001)*, Vienna, Austria, 2001, pp. 692-698.
- [45] G. Nasierding and A. Z. Kouzani, "Comparative evaluation of multi-label classification methods," in *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, Sichuan, China, 2012, pp. 679-683.
- [46] D. Jung and A. Savvides, "Estimating building consumption breakdowns using on/off state sensing and incremental sub-meter deployment," in *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, Zurich, Switzerland, 2010, pp. 225-238.
- [47] K. Suzuki, S. Inagaki, T. Suzuki, H. Nakamura, and K. Ito, "Nonintrusive Appliance Load Monitoring Based on Integer Programming," in *2008*

- Proceedings of Sice Annual Conference*, Fukuoka, Japan, 2008, pp. 2626-2631.
- [48] Y. Kim, T. Schmid, Z. M. Charbiwala, and M. B. Srivastava, "ViridiScope: design and implementation of a fine grained power monitoring system for homes," in *Proceedings of the 11th international conference on Ubiquitous computing*, Orlando, FL, 2009, pp. 245-254.
- [49] Y. X. Yu, P. Li, and C. L. Zhao, "Non-Intrusive Method for On-Line Power Load Decomposition," in *2008 China International Conference on Electricity Distribution*, Guangzhou, China, 2009, pp. 100-107.
- [50] T. Zia, D. Bruckner, and A. Zaidi, "A Hidden Markov Model Based Procedure for Identifying Household Electric Loads," *Iecon 2011: 37th Annual Conference on Ieee Industrial Electronics Society*, 2011.
- [51] S. Frank, L. G. Polese, E. Rader, M. Sheppy, and J. Smith, "Extracting Operating Modes from Building Electrical Load Data," in *2011 IEEE Green Technologies Conference (IEEE-Green)*, Baton Rouge, Louisiana, 2011, pp. 1-6.
- [52] H. Goncalves, A. Ocneanu, M. Berges, and R. Fan, "Unsupervised disaggregation of appliances using aggregated consumption data," in *The 1st KDD Workshop on Data Mining Applications in Sustainability (SustKDD)*, Beijing, China, 2011.
- [53] S. K. Ng, J. Liang, and J. W. Cheng, "Automatic appliance load signature identification by statistical clustering," in *8th International Conference on Advances in Power System Control, Operation and Management (APSCOM 2009)*, Manchester, UK, 2009.
- [54] D. C. Bergman, D. Jin, J. P. Juen, N. Tanaka, C. A. Gunter, and A. K. Wright, "Nonintrusive Load-Shed Verification," *IEEE Pervasive Computing*, vol. 10, no. 1, pp. 49-57, Jan-Mar 2011.
- [55] D. C. Bergman, D. Jin, J. P. Juen, N. Tanaka, C. A. Gunter, and A. K. Wright, "Distributed non-intrusive load monitoring," in *2011 IEEE PES Innovative Smart Grid Technologies (ISGT)*, Anaheim, CA, 2011, pp. 1-8.
- [56] M. Baranski and A. Voss, "Genetic algorithm for pattern detection in NIALM systems," in *2004 IEEE International Conference on Systems, Man & Cybernetics*, The Hague, Netherlands, 2004, pp. 3462-3468.
- [57] H. Y. Lam, G. S. K. Fung, and W. K. Lee, "A novel method to construct taxonomy of electrical appliances based on load signatures," *IEEE Transactions on Consumer Electronics*, vol. 53, no. 2, pp. 653-660, May 2007.
- [58] Z. Wang and G. Zheng, "Residential Appliances Identification and Monitoring by a Nonintrusive Method," *IEEE Transactions on Smart Grid*, vol. 3, no. 1, pp. 80-92, 2012.
- [59] Z. Y. Wang and G. L. Zheng, "The Application of Mean-shift Cluster in Residential Appliance Identification," in *2011 30th Chinese Control Conference (Ccc)*, Yantai, China, 2011, pp. 3111-3114.
- [60] J. Z. Kolter and T. Jaakkola, "Approximate Inference in Additive Factorial HMMs with Application to Energy Disaggregation," in *International*

- Conference on Artificial Intelligence and Statistics*, La Palma, Canary Islands, 2012, pp. 1472-1482.
- [61] H. S. Kim, "Unsupervised disaggregation of low frequency power measurements," PhD, University of Illinois, 2012.
- [62] M. Zeifman, "Disaggregation of Home Energy Display Data Using Probabilistic Approach," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 1, pp. 23-31, Feb 2012.
- [63] O. Parson, S. Ghosh, M. Weal, and A. Rogers, "Non-Intrusive Load Monitoring Using Prior Models of General Appliance Types," in *26th AAAI Conference on Artificial Intelligence*, Toronto, 2012.
- [64] O. Parson, S. Ghosh, M. Weal, and A. Rogers, "Using hidden markov models for iterative non-intrusive appliance monitoring," in *Neural Information Processing Systems workshop on Machine Learning for Sustainability*, Sierra Nevada, Spain, 2011.
- [65] M. Zeifman and K. Roth, "Viterbi algorithm with sparse transitions (VAST) for nonintrusive load monitoring," in *2011 IEEE Symposium on Computational Intelligence Applications In Smart Grid (CIASG)*, Paris, France, 2011, pp. 1-8.
- [66] M. Dong, P. C. M. Meira, W. Xu, and C. Y. Chung, "Non-Intrusive Signature Extraction for Major Residential Loads," *IEEE Transactions on Smart Grid*, vol. 4, no. 3, pp. 1421-1430, 2013.
- [67] M. B. Figueiredo, A. de Almeida, B. Ribeiro, and A. Martins, "Extracting Features from an Electrical Signal of a Non-Intrusive Load Monitoring System," in *Intelligent Data Engineering and Automated Learning (Ideal 2010)*, Paisley, Scotland, 2010, pp. 210-217.
- [68] H. S. Kim, "Unsupervised disaggregation of low frequency power measurements," University of Illinois, 2012.
- [69] Z. Xing, J. Pei, and E. Keogh, "A brief survey on sequence classification," *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 1, pp. 40-48, 2010.
- [70] E. Keogh and S. Kasetty, "On the need for time series data mining benchmarks: a survey and empirical demonstration," *Data Mining and Knowledge Discovery*, vol. 7, no. 4, pp. 349-371, 2003.
- [71] C. A. Ratanamahatana and E. Keogh, "Everything you know about dynamic time warping is wrong," in *Third Workshop on Mining Temporal and Sequential Data*, New York, NY, 2004, pp. 22-25.
- [72] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 43-49, 1978.
- [73] S. Chan, B. Kao, C. L. Yip, and M. Tang, "Mining emerging substrings," in *Proceedings Eighth International Conference on Database Systems for Advanced Applications*, Kyoto, Japan, 2003, pp. 119-126.
- [74] J. Nerbonne, W. Heeringa, and P. Kleiweg, "Comparison and classification of dialects," in *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, Stroudsburg, PA, 1999, pp. 281-282.

- [75] J. Aach and G. M. Church, "Aligning gene expression time series with time warping algorithms," *Bioinformatics*, vol. 17, no. 6, pp. 495-508, 2001.
- [76] Z. Bar-Joseph, G. Gerber, D. K. Gifford, T. S. Jaakkola, and I. Simon, "A new approach to analyzing gene expression time series data," in *Proceedings of the sixth annual international conference on Computational biology*, Washington, DC, 2002, pp. 39-48.
- [77] C. A. Ratanamahatana and E. Keogh, "Making time-series classification more accurate using learned constraints," in *Proceedings of SIAM international conference on data mining*, Buena Vista, FL, 2004, pp. 11-22.
- [78] F. Kupzog, T. Zia, and A. A. Zaidi, "Automatic electric load identification in self-configuring microgrids," in *AFRICON 2009*, Nairobi, Kenya, 2009, pp. 1-5.
- [79] C. Valens, "A really friendly guide to wavelets," *C. Valens@ mindless.com*, vol. 2004, 1999.
- [80] H. Chang, K. Lian, Y. Su, and W. Lee, "Power Spectrum-Based Wavelet Transform for Non-Intrusive Demand Monitoring and Load Identification," *IEEE Transactions on Industry Applications*, 2014.
- [81] S. Giri, M. Bergés, and A. Rowe, "Towards automated appliance recognition using an EMF sensor in NILM platforms," *Advanced Engineering Informatics*, 2013.
- [82] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining multi-label data," in *Data mining and knowledge discovery handbook*, ed: Springer, 2010, pp. 667-685.
- [83] S. Diplaris, G. Tsoumakas, P. A. Mitkas, and I. Vlahavas, "Protein classification with multiple algorithms," in *Advances in Informatics*, ed: Springer, 2005, pp. 448-456.
- [84] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in *Advances in neural information processing systems (NIPS)*, Vancouver, 2001, pp. 681-687.
- [85] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757-1771, 2004.
- [86] Z.-H. Zhou and M.-L. Zhang, "Multi-instance multi-label learning with application to scene classification," in *Advances in Neural Information Processing Systems*, Vancouver, 2006, pp. 1609-1616.
- [87] X.-S. Hua and G.-J. Qi, "Online multi-label active annotation: towards large-scale content-based video search," in *Proceedings of the 16th ACM international conference on Multimedia*, Vancouver, 2008, pp. 141-150.
- [88] X. Luo and A. N. Zincir-Heywood, "Evaluation of two systems on multi-class multi-label document classification," in *Foundations of Intelligent Systems*, ed: Springer, 2005, pp. 161-169.

- [89] A. McCallum, "Multi-label text classification with a mixture model trained by EM," in *AAAI'99 Workshop on Text Learning*, Madison, Wisconsin, 1999, pp. 1-7.
- [90] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2038-2048, 2007.
- [91] K. Basu, V. Debusschere, and S. Bacha, "Load identification from power recordings at meter panel in residential households," in *2012 XXth International Conference on Electrical Machines (ICEM)*, France, 2012, pp. 2098-2104.
- [92] M. Zeifman and K. Roth, "Nonintrusive appliance load monitoring: Review and outlook," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 1, pp. 76-84, 2011.
- [93] H. Kantz and T. Schreiber, *Nonlinear time series analysis*, 2nd ed. vol. 7. New York: Cambridge University Press, 2003.
- [94] T. Sauer, J. A. Yorke, and M. Casdagli, "Embedology," *Journal of Statistical Physics*, vol. 65, no. 3, pp. 579-616, 1991.
- [95] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical systems and turbulence, Warwick 1980*, Warwick, UK, 1981, pp. 366-381.
- [96] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, 2004.
- [97] K. Neusser, "Title," unpublished|.
- [98] H. S. Kim, R. Eykholt, and J. Salas, "Nonlinear dynamics, delay times, and embedding windows," *Physica D: Nonlinear Phenomena*, vol. 127, no. 1, pp. 48-60, 1999.
- [99] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3-55, 2001.
- [100] A. M. Fraser and H. L. Swinney, "Independent coordinates for strange attractors from mutual information," *Physical review A*, vol. 33, no. 2, p. 1134, 1986.
- [101] S. S. Haykin, *Neural networks and learning machines*, 3 ed. New York: Prentice Hall 2009.
- [102] S. Kullback, *Information theory and statistics*. Mineola, N.Y: Courier Dover Publications, 1997.
- [103] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig, "The mutual information: detecting and evaluating dependencies between variables," *Bioinformatics*, vol. 18, no. suppl 2, pp. S231-S240, 2002.
- [104] M. Zhu, "Feature Extraction and Dimension Reduction with Applications to Classification and the Analysis of Co-occurrence Data," stanford university, 2001.
- [105] Y. Y. Tang, *Wavelet theory approach to pattern recognition*, 2nd ed. vol. 74. Singapore: World Scientific, 2009.
- [106] R. X. Gao and R. Yan, *Wavelets: Theory and Applications for Manufacturing*. New York: Springer, 2010.

- [107] J. M. Fadili and E. T. Bullmore, "Wavelet-based approaches for multiple hypothesis testing in activation mapping of functional magnetic resonance images of the human brain," in *SPIE's 48th Annual Meeting Optical Science and Technology*, 2003, pp. 405-416.
- [108] Y. Liu, "Dimensionality reduction and main component extraction of mass spectrometry cancer data," *Knowledge-Based Systems*, vol. 26, pp. 207-215, 2012.
- [109] C. Torrence and G. P. Compo, "A practical guide to wavelet analysis," *Bulletin of the American Meteorological society*, vol. 79, no. 1, pp. 61-78, 1998.
- [110] Q. L. Tao Li, Shenghuo Zhu, Mitsunori Ogihara, "A survey on wavelet applications in data mining," *SIGKDD Explorations*, vol. 4, no. 2, pp. 49-68, 2003.
- [111] X. Z. Tan Minsheng, Wang Lei, "Voronoi Tessellation based Haar Wavelet Data Compression for Sensor Network," *International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM*, pp. 1-4, 2006.
- [112] C. L. Chentan Gupta, Song Wang, Abhay Mehta, "Non-Dyadic Haar Wavelets for Streaming and Sensor Data," *26th International Conference on Data Engineering (ICDE), IEEE*, pp. 569-580, 2010.
- [113] A. W.-c. F. Franky Kin-Pong Chan, Clement Yu, "Haar Wavelets for Efficient Similarity search of Time Series: With and Without Time Wrapping," *IEEE Transaction on Knowledge and Data Engineering*, vol. 15, no. 3, p. 20, 2003.
- [114] A. N. Su Chen, "Dynamic Nonuniform Data Aproximation in Databases with Haar Wavelet," *Journal of Computer*, vol. 2, no. 8, pp. 67-76, 2007.
- [115] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 3, no. 3, pp. 1-13, 2007.
- [116] G. Tsoumakas and I. Vlahavas, "Random k-Labelsets: An Ensemble Method for Multilabel Classification," in *ECML*, 2007.
- [117] G. Tsoumakas and I. Vlahavas, "Random k-labelsets: An ensemble method for multilabel classification," in *Machine Learning: ECML 2007*, ed: Springer, 2007, pp. 406-417.
- [118] M. H. DeGroot, *Optimal statistical decisions* vol. 82: Wiley-Interscience, 2005.
- [119] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, 2012.
- [120] S. Chu, E. J. Keogh, D. Hart, and M. J. Pazzani, "Iterative Deepening Dynamic Time Warping for Time Series," in *SIAM International Conference on Data Mining*, Chicago. Arlington, 2002.
- [121] R. Agrawal, C. Faloutsos, and A. Swami, "Efficient similarity search in sequence databases," in *Foundations of Data Organization and Algorithms*, ed: Springer, 1993, pp. 69-84.

- [122] M. Vlachos, M. Hadjieleftheriou, D. Gunopulos, and E. Keogh, "Indexing multi-dimensional time-series with support for multiple distance measures," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, Washington, DC, 2003, pp. 216-225.
- [123] M. Müller, *Information retrieval for music and motion*: Springer, 2007.
- [124] E. J. Keogh and M. J. Pazzani, "Scaling up dynamic time warping to massive datasets," in *Principles of Data Mining and Knowledge Discovery*, ed: Springer, 1999, pp. 1-11.
- [125] D. J. Berndt and J. Clifford, "Using Dynamic Time Warping to Find Patterns in Time Series," in *KDD workshop*, 1994, pp. 359-370.
- [126] R. Hegger, H. Kantz, and T. Schreiber, "Practical implementation of nonlinear time series methods: The TISEAN package," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 9, no. 2, pp. 413-435, 1999.
- [127] G. Tsoumakas, I. Katakis, and I. Vlahavas, *Mining multi-label data*. New York: Springer US, 2010.
- [128] R. Rak, L. Kurgan, and M. Reformat, "Multi-label associative classification of medical documents from medline," in *Fourth International Conference on Machine Learning and Applications*, Los Angeles, CA, 2005, p. 8 pp.
- [129] M. S. Ahmed, L. Khan, N. Oza, and M. Rajeswari, "Multi-Label ASRS Dataset Classification Using Semi-Supervised Subspace Clustering," in *CIDU 2010: Conference on Intelligent Data Understanding*, San Francisco, 2010.
- [130] L. S. Davis, "A survey on edge detection methods," *Computer Graphics and Image Processing*, vol. 4, no. 3, pp. 248-270, 1975.
- [131] R. N. Strickland and H. I. Hahn, "Wavelet transform methods for object detection and recovery," *IEEE Transactions on Image Processing*, vol. 6, no. 5, pp. 724-735, 1997.