Non-Intrusive load monitoring and anomaly detection: On the importance of feature selection for supervised and unsupervised learning for sensor applications

Mohammad Khazaei

Department of Electronic and Electrical Engineering

University of Strathclyde, Glasgow

November , 2025

Abstract

Non-Intrusive Load Monitoring (NILM) provides a scalable and cost-effective means of disaggregating household energy consumption from aggregate smart meter data. In addition to disaggregation, NILM holds promise for detecting abnormal appliance behaviour, supporting predictive maintenance, enhancing energy efficiency, and improving household safety. However, several challenges hinder its practical deployment, including high computational cost, lack of transferability across households, sensitivity to low-frequency smart meter data, and limited frameworks for anomaly detection.

This thesis addresses these challenges through three main contributions. First, it presents a systematic investigation of feature selection for NILM. By analysing both spectrometry and smart meter datasets, it demonstrates that selecting the most informative features reduces computational complexity while improving classification accuracy. Comparative analyses of supervised and unsupervised algorithms—including Artificial Neural Networks (ANN), Decision Trees (DT), K-Means, and DBSCAN—highlight the role of feature selection in balancing performance and efficiency.

Second, the thesis evaluates load disaggregation using supervised (DT, KNN) and unsupervised (DBSCAN, UGSP) techniques. Detailed pre-processing (e.g., noise reduction, resampling, segmentation) and post-processing (e.g., power reconciliation, error correction) methods are integrated to improve robustness. A two-stage disaggregation strategy is proposed to enhance detection accuracy, and transfer learning experiments are conducted to assess generalisability across households, offering insights into scalability and dataset adaptability.

Third, the thesis develops a novel framework for NILM-based anomaly detection. A hybrid approach, combining Unsupervised Graph Signal Processing (UGSP) with a statistical rule-based method, is applied to fridge-freezers and washing machines in the REFIT dataset. By modelling ON-duration thresholds and operational cycles, the framework successfully identifies abnormal appliance behaviour without the need for intrusive submetering. The results show improved precision and recall in detecting anomalies, demonstrating the feasibility of NILM-driven predictive fault detection.

Overall, this work advances NILM from a disaggregation-focused task toward an anomaly-aware monitoring framework. It contributes: (i) a systematic evaluation of feature selection strategies, (ii) comparative benchmarking of supervised and unsupervised disaggregation methods, (iii) an assessment of cross-dataset transferability, and (iv) the development of a hybrid NILM-based anomaly detection approach. Together, these contributions provide a comprehensive framework that reduces computational overhead, improves robustness, and enables early detection of appliance faults. The findings support the integration of NILM into smart grid infrastructures, offering benefits for energy efficiency, user engagement, and residential safety.

Contents

Contents

List of Acronyms

Acronym	Full Form
AMI	Advanced Metering Infrastructure
ANN	Artificial Neural Network
APS	Appliance Power Signature
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DER	Distributed Energy Resources
DL	Deep Learning
DNO	Distribution Network Operator
DSM	Demand Side Management
DT	Decision Tree
DW	Dishwasher
FF	Fridge-Freezer
FN	False Negative
FP	False Positive
GBF	Bilateral Graph Filtering
GSP	Graph Signal Processing
HAN	Home Area Network
HMM	Hidden Markov Model
HV	High Voltage
IHD	In-Home Display
ITP	Incorrect True Positive
KNN	K-Nearest Neighbour
LV	Low Voltage
ML	Machine Learning
NAN	Neighbourhood Area Network
NILM	Non-Intrusive Load Monitoring
NOR	Norway dataset (used in NILM evaluation)
PR	Precision
RE	Recall
REDD	Reference Energy Disaggregation Dataset
REFIT	Reference Energy Forecasting and Intervention Dataset
RNN	Recurrent Neural Network
SVM	Support Vector Machine
TD	Tumble Dryer
TP	True Positive
UGSP	Unsupervised Graph Signal Processing
WAN	Wide Area Network
WD	Washing Dryer
WM	Washing Machine

List of Figures

List of Tables

Acknowledgements

I would like to sincerely thank Prof Lina Stankovic and Prof Vladimir Stankovic for their unwavering guidance and support over the years, both academically and personally, which has been instrumental in shaping me into a well-rounded PhD candidate. They introduced me to the exciting and innovative fields of energy management, data analysis, and machine learning. My gratitude also extends to all my colleagues at Strathclyde, as well as the reviewers and audiences, for their valuable advice and feedback.

The author was supported by the University of Strathclyde Faculty of Engineering Scholarship (2018-2021), and University of Strathclyde Department of Electronic and Electrical Engineering Scholarship (2018-2021).

Lastly, I want to express my heartfelt appreciation to my wife and family for their constant support and encouragement throughout the years.

Chapter 1

1 Introduction

1.1 Introduction

According to the quarterly report by UK government at the end of September 2024 , Across Great Britain, 37 million smart and advanced meters have been installed in homes and small businesses, representing 65% of all meters [1]. By 2025, the majority of UK households will have smart electricity meters, offering Distribution Network Operators access to new customer data streams at the Low Voltage (LV) network level. These smart meters will facilitate two-way communication between network operators and end users[2] the installation has some delay in [3] is mentioned that the government recently engaged with suppliers and other industry stakeholders to discuss its plans for 2024 and 2025. It has now set targets for suppliers to ensure that smart meters are installed in at least 74.5% of households and nearly 69% of small businesses by the end of 2025 and the main motivation for the widespread deployment of smart meters globally is to maximize the benefits of the smart grid. Smart meter data has proven to improve grid operations, support the maintenance of distribution networks, detect faults, identify non-technical losses, predict outages, enhance load forecasting, and enable demand response. It also boosts customer satisfaction by ensuring accurate billing and providing valuable energy feedback through Non-Intrusive Load Monitoring (NILM), which disaggregating the total household consumption down to the load level [4]. There are two main methods for obtaining information about appliances: Intrusive Load Monitoring (ILM) and Non-Intrusive Load Monitoring (NILM). ILM involves placing multiple sensors directly on appliances or at the plug level, while NILM relies on a single smart meter typically installed at the panel level. In the NILM approach, all electrical signatures from various appliances are combined into a single signal, and disaggregation techniques are used to extract portions of the original data. Understanding the current operational state of appliances can help implement energy-saving measures, such as turning off devices in standby mode, powering off appliances, or transitioning them to a more energy-efficient state[5].

Feature selection plays a vital role in both supervised and unsupervised learning algorithms used for sensor data classification. Effective feature selection can enhance the performance of machine learning models by reducing noise, improving interpretability, and minimizing computational cost, especially when applied to complex sensing systems such as photonic sensors or smart meters. In NILM and anomaly detection systems, identifying the most informative features from raw data is essential for achieving reliable appliance classification and timely anomaly identification.

1.2 Research Motivation

Non-Intrusive Load Monitoring (NILM) enables load disaggregation through a non-intrusive, software-based method. This computational approach separates the combined energy usage recorded by a single electricity meter into the individual loads of specific appliances. NILM presents a significant opportunity to capitalize on global investments in smart metering technology[6][7]. In addition to providing valuable energy feedback, NILM shows potential for the timely detection of malfunctioning appliances without the need for submetering, making it a promising solution[8][9]. Indeed, the BBC has reported that malfunctioning appliances, especially white goods, were responsible for almost 12,000 fires in Great Britain in just over 3 years[10] and the Scottish Fire and Rescue Service reported 340 fires in 2019 alone, caused by tumble dryers, washing machine, fridge-freezers and dishwashers[11]. As shown in [8] Appliance malfunctions can be identified when the NILM signature significantly deviates from the actual energy consumption, with the deviation aligning with a known anomaly signature. Furthermore, [8] indicates that appliance-level anomaly detection performs best for best performing NILM algorithms, but this was only verified for relatively complex supervised Combinatorial Optimization (CO) [12] and Factorial Hidden Markov Model (FHMM) [13] based NILM from the NILMTK [14], Latent Bayesian Melding [15], Super-state HMM (SSHMM) [13] and unsupervised Graph Signal Processing (GSP) [16] NILM algorithms and anomalies in electrical heater and freezer operation in the REFIT dataset [17] . Motivation for the contributions of this thesis, are as below:

(I) Find the best feature to minimise computational cost and maximise the performance of ML

(II)Determine suitability of NILM algorithms to quickly identify ON-OFF appliance cycles

(III) How we can use transferability in training and what is the effect of different data sets

(IV) How we can find abnormal or anomalous appliances in the home from aggregation data (aggregate smart meter)?

1.3 Research Contribution

This thesis extends earlier work within the research group in three important ways. Prior research by team, established the foundations of low-rate NILM using Graph Signal Processing (GSP) and pre-processing and explored preliminary feature analysis for classification. The work in this thesis builds upon these foundations through: (i) a systematic and cross-dataset evaluation of feature selection methods, which had not previously been undertaken in the group's NILM research; (ii) the development and benchmarking of a two-stage disaggregation strategy designed specifically for low-frequency smart meter environments; and (iii) the introduction of a hybrid NILM-based anomaly detection framework that integrates Unsupervised GSP (UGSP) with statistical post-processing methods. These contributions go beyond earlier work by addressing computational efficiency, transferability, and anomaly detection—areas that were previously only partly explored. Specifically, the thesis contributions are clarified below.

In order to address **how to identify the most effective features that minimize computational cost while maximizing the performance of machine learning algorithms**, **Chapter 3** investigates feature selection and classification techniques within NILM datasets. Artificial Neural Networks (ANN) are employed to evaluate the selected features, and the results are validated against ground truth data. A comparative analysis with Support Vector Machines (SVM) is also conducted. This chapter demonstrates the crucial role of feature selection in enhancing the efficiency and accuracy of NILM systems.

To explore **which NILM algorithms are most suitable for near real-time disaggregation and rapid appliance identification**, **Chapter 4** evaluates a range of supervised and unsupervised machine learning methods. Supervised algorithms such as Decision Trees (DT) and K-Nearest Neighbours (KNN) are compared with unsupervised techniques like DBSCAN and K-Means. The impact of various preprocessing approaches on classification performance is examined. Additionally, a two-step disaggregation strategy is proposed, in which appliances with similar load characteristics are initially grouped to improve detection accuracy and processing speed.

In examining **how transferability in training can be leveraged, and what the effects of using different datasets are**, **Chapter 4** further investigates the generalization capability of NILM models across multiple households. The study assesses how models trained on one dataset perform when applied to another, offering insights into the robustness and adaptability of NILM frameworks in diverse residential settings.

To answer **how anomalous or abnormal appliance behaviour can be detected using aggregate smart meter data**, **Chapter 5** focuses on anomaly detection in household appliances such as ovens, washing machines, dishwashers, and fridge-freezers. Unsupervised NILM techniques and rule-based methods are applied to identify deviations from typical usage patterns. This chapter demonstrates how NILM can support early fault detection, improve energy efficiency, and enhance safety in residential environments.

1.4 Publication List

This section provides a list of all published or submitted works related to this thesis

1. Khazaei, M., Stankovic, L., & Stankovic, V. (2019, December). ***Trends and challenges in smart metering analytics*****.**

Conference: 2019 MTMI International Conference on Emerging Issues in Business, Technology and Applied Sciences (pp. 111–117).
Detailed in: Chapter 3

CRediT author statement:
Mohammad Khazaei: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft& Project administration.
Lina Stankovic: Supervision, Writing – review & editing.
Vladimir Stankovic: Supervision, Resources, Writing – review & editing.

---

2. Herrero-Bermello, A., Li, J., Khazaei, M., Grinberg, Y., Velasco, A.V., Vachon, M., Cheben, P., Stankovic, L., Stankovic, V., Xu, D.X. and Schmid, J.H.. (2019). ***On-chip Fourier-transform spectrometers and machine learning: a new route to smart photonic sensors*****.**

Journal: *Optics Letters*, 44(23), 5840–5843.
Detailed in: Chapter 3

CRediT author statement:
Mohammad Khazaei: Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft (partial) for ANN feature selection and classification.
Herrero-Bermello, A., Li, J., Khazaei, M., Grinberg, Y., Velasco, A.V., Vachon, M., Cheben, P., Stankovic, L., Stankovic, V., Xu, D.X. and Schmid, J.H.: Conceptualization, Supervision, Resources, Methodology, Software, Validation, Formal analysis, Investigation, Writing – review & editing, Project administration.

---

3. Herrero-Bermello, A., Li, J., Khazaei, M., Grinberg, Y., Villafranca-Velasco, A., Vachon, M., Cheben, P., Stankovic, L., Stankovic, V., Xu, D.X. and Schmid, J.H. (2020, February). ***Smart on-chip Fourier-transform spectrometers harnessing machine learning algorithms*****.**

Conference: SPIE Photonics West OPTO.
Detailed in: Chapter 3

CRediT author statement:
Mohammad Khazaei: Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft (partial) for ANN feature selection and classification
Herrero-Bermello, A., Li, J., Grinberg, Y., Villafranca-Velasco, A., Vachon, M., Cheben, P., Stankovic, L., Stankovic, V., Xu, D.X. and Schmid, J.H.: Conceptualization, Supervision, Project administration, Resources, Methodology, Software, Validation, Formal analysis, Investigation ,Writing–review & editing.

4. Khazaei, M., Stankovic, L., & Stankovic, V. (2020, November). ***Evaluation of low-complexity supervised and unsupervised NILM methods and pre-processing for detection of multistate white goods*****.**

Conference: 5th International Workshop on Non-Intrusive Load Monitoring.
Detailed in: Chapter 4

CRediT author statement:
Mohammad Khazaei: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft & Project administration .
Lina Stankovic: Supervision, Writing – review & editing.
Vladimir Stankovic: Supervision, Resources, Writing – review & editing.

1.5 Thesis overview

This thesis addresses three key aspects of Non-Intrusive Load Monitoring (NILM): feature selection, load disaggregation, and anomaly detection. The overarching aim is to enhance NILM performance on low-frequency smart meter data by reducing computational complexity, improving transferability, and enabling early detection of appliance faults.

Chapter 2 provides the background and literature review. It introduces NILM, smart meter applications, electrical device characteristics, and feature extraction methods. The chapter also highlights existing NILM algorithms, challenges in scalability and generalisability, performance evaluation metrics, and anomaly detection approaches, identifying gaps that motivate this research.

Chapter 3 investigates feature selection and multi-class classification. Two distinct datasets—spectrometry data and electrical measurement data—are analysed using Artificial Neural Networks (ANN), Decision Trees (DT), K-Means, and DBSCAN. Results show that effective feature selection reduces redundancy, lowers computational cost, and improves classification accuracy.

Chapter 4 examines load disaggregation of smart meter data. Both supervised (DT, KNN) and unsupervised (DBSCAN, UGSP) algorithms are evaluated, with detailed consideration of pre-processing and post-processing stages. A two-stage disaggregation strategy is proposed, and transfer learning is explored to assess cross-household applicability, addressing limitations in generalisation across datasets.

Chapter 5 focuses on anomaly detection in appliances, with a particular emphasis on fridge-freezers. A hybrid framework combining unsupervised UGSP with a rule-based approach is developed to detect abnormal ON-durations and usage cycles. Case studies on the REFIT dataset demonstrate the framework's robustness and highlight its contribution to energy efficiency and household safety.

Chapter 6: Conclusions and Future Work

In summary, the thesis advances NILM from disaggregation-focused methods toward anomaly-aware monitoring, providing a comprehensive framework that supports both improved energy management and enhanced appliance fault detection in residential environments.

Chapter 2

2 Background and Literature Review

The global transition toward smarter and more sustainable energy systems has driven significant interest in Non-Intrusive Load Monitoring (NILM) as a cost-effective and scalable method for estimating individual appliance energy consumption from aggregate smart meter data. NILM eliminates the need for appliance-level submetering by applying advanced signal processing and machine learning techniques to infer the operational state and energy profile of household devices. The accurate identification of appliance-level consumption patterns has far-reaching implications for energy efficiency, demand response, predictive maintenance, and user feedback [12], [18].

At the core of NILM performance lies the process of feature extraction and selection. Features represent the distinct characteristics of power signals that can be linked to appliance operation, including steady-state power levels, switching events (ΔP), transient signatures, time-of-day usage, and duration. The richness of the feature set directly influences the accuracy of disaggregation, yet an overly complex feature space can hinder real-time processing and generalizability. Studies such as [18] and [19] have underscored the trade-off between model complexity and computational cost, particularly when NILM is deployed on edge devices or integrated with smart meters with limited processing capabilities. Efficient feature selection methods, including mutual information, correlation-based selection, and L1-regularization, have been employed to reduce redundancy and retain the most informative attributes, thereby enabling lightweight, real-time implementations.

Real-time or near real-time NILM is gaining increasing relevance as smart grid infrastructure demands timely and actionable insights. Traditional NILM algorithms such as Factorial Hidden Markov Models, while powerful, are typically computationally intensive and unsuitable for live processing [20] . In contrast, supervised learning approaches, such as Decision Trees, Support Vector Machines, and Neural Networks, have shown promise for quick and efficient classification once trained [21] . Additionally, event-based NILM frameworks that focus on detecting significant power transitions before invoking classification routines have demonstrated improved scalability and responsiveness [22] . These systems are particularly suited for embedded deployment, where responsiveness and energy efficiency are critical.

Another dimension of NILM research involves model generalizability and transferability across households and datasets. Due to variations in appliance brands, household behaviours, and regional grid conditions, NILM models trained on one dataset often underperform when applied to new environments. Transfer learning and domain adaptation techniques have emerged to address this limitation. For instance, models trained on widely used datasets like REFIT, UK-DALE, or REDD can be fine-tuned or adapted using unlabelled data from target homes to improve performance without the need for extensive retraining [23] . Cross-dataset evaluation remains a central concern, and studies emphasize the need for robust, domain-invariant features that can maintain predictive power across different settings [21] .

Beyond energy disaggregation, NILM has been extended to support appliance-level anomaly detection. Identifying abnormal behaviour in electrical appliances can inform users about potential faults, energy waste, or safety risks. This task is particularly challenging when relying solely on aggregate data, but recent advances have demonstrated that once disaggregated appliance profiles are obtained, they can be monitored for deviations in expected usage patterns. Methods such as statistical thresholding, clustering, and unsupervised learning—e.g., autoencoders and isolation forests—have been used to detect anomalies based on variations in operating duration, consumption levels, or usage frequency [12], [24] . Fridges with compressor faults or washing machines running outside normal hours are typical examples of anomalous operation that NILM systems aim to identify. Such applications offer valuable opportunities for predictive maintenance and improving residential energy safety and efficiency.

The following section highlights key aspects of smart meters and electrical appliances, focusing on their operational characteristics and relevant features for NILM applications

2.1 Smart Metering

Climate change, increased awareness of energy efficiency, evolving electricity markets, and the growing role of consumers as prosumers in microgrid generation are driving the adoption of Renewable Energy Resources (RES), Distributed Generation (DG), and Distributed Storage (DS). The transition to a decentralized electricity grid capable of efficiently managing numerous generation and storage devices lies at the core of the Smart Grid (SG). This evolution relies heavily on communication across all components of the grid. Smart meters serve as the interface between the grid and end-users (e.g., households) by facilitating communication between meters and suppliers, either directly or through neutral third parties[6].

The choice of communication technology depends on factors such as ease of installation, integration with existing infrastructure, economic considerations, technical requirements, and functionality. For instance, wireless communication may face challenges in remote areas, while Power Line Communication (PLC) is cost-effective for rural long-distance coverage but suffers from long data transmission latencies, limited bandwidth, and higher costs in urban areas [25] [26] . A mixed communication approach is often adopted; for example, in the Netherlands and Germany, 20% of meters use GPRS, while 80% rely on PLC [27].

Direct communication between smart meters and data collectors offers high bandwidth and throughput but struggles with long distances and integration with distribution automation devices. Consequently, a mesh network approach is widely used, organized into three hierarchical tiers[28]: Wide Area Network (WAN), Neighbourhood Area Network (NAN), and Home Area Network (HAN).

- HAN: Enables connectivity within a home or business environment, connecting the smart meter to devices such as in-home displays, gas meters, and sensors. Common technologies include IEEE 802.15.4 (ZigBee), IEEE 802.11 (Wi-Fi), and wired PLC. These networks have short coverage (1–10 meters) and low data rates in the range of kilobits per second.
- NAN: Aggregates energy usage data from multiple HANs to gateways, enhancing the number of smart meters that can connect to a base station. Communication between a base station and data concentrators often uses WiMAX/LTE, while RF 900 MHz links data concentrators to smart meters.
- WAN: Acts as the backbone for communication across distributed networks in the grid, using technologies like fibre optics, DSL, and cellular networks (e.g., 2G/3G/4G LTE, WiMAX). WANs cover long distances (10–100 km) and support high data rates (10–100 Mbps).

In the U.S., Advanced Metering Infrastructure (AMI) deployments often employ fibre optics and WiMAX/LTE for the AMI backbone, RF mesh networks for NANs, and ZigBee or Wi-Fi for HANs [29].

The widespread adoption of smart meters is driven by benefits such as accurate billing, enhanced customer experience, and improved Demand Side Management (DSM). Real-time energy feedback through In-Home Displays (IHDs) has been shown to reduce energy demand, while advanced analytics like Non-Intrusive Load Monitoring (NILM) enable appliance-level consumption insights, leading to additional energy savings of up to 4.5% [21]. At the grid level, smart meters improve operations, grid maintenance, low-voltage network forecasting, and infrastructure planning, making them valuable tools for energy suppliers and Distribution Network Operators (DNOs) to optimize network performance and electricity management.

Smart metering applications are examined in two categories: those focused on network operations from the perspective of Distribution Network Operators (DNOs), those addressing technical and non-technical losses from the energy provider's standpoint, and those cantered on Non-Intrusive Load Monitoring (NILM) for end-users.

2.1.1 Network Topology and Maintenance

Measurements from smart meters can be used to derive network topology, enabling the identification of loading and voltage profiles, connectivity issues, and the impact of Distributed Energy Resources (DER) on the grid, which is particularly beneficial for High Voltage (HV) networks. Additionally, smart meter data can aid in forecasting future demand trends and management needs, as well as maintenance requirements. For instance, it can be utilized to calculate the operational duration of transformers [31].

2.1.2 Detecting Voltage Deviation

The analysis of smart meter events enhances the operation of low-voltage (LV) networks by identifying voltage deviations or outages. For instance, in [32], a series of strategies were implemented to manage large-scale measurement data effectively: (1) gathering, filtering, and sorting events; (2) isolating event types closely linked to Quality of Service (QoS), such as overvoltage, undervoltage, and neutral loss; and (3) incorporating measurements from distribution transformers at each substation to complement event analysis and improve the detection of critical hotspots in the LV network.

2.1.3 Network Line Outage and Fault Detection

Real-time analysis of multivariate smart meter data is employed to predict outages before they occur and to detect and isolate their location and severity after they happen [33]. The study concluded that identifying pre-outage conditions as an anomaly level and efficiently predicting them using smart meter data can enable utilities to anticipate outages, allowing for better maintenance planning and the prevention of costly disruptions. Similarly, defining post-outage behaviour as another anomaly level assists utilities in performing post-outage activities more effectively, such as pinpointing the outage location, assessing its severity, and ensuring faster restoration.

Descriptive analytics based on historical data for various cause-specific outages revealed no fully deterministic or consistent relationship between individual smart meter measurements and outage events. This underscores the need for advanced analytics frameworks capable of integrating and analysing multi-dimensional, multi-source smart meter data. Such frameworks must generate actionable insights while addressing critical challenges, such as handling missing data points, to support informed decision-making and enhance outage management[6].

2.1.4 Technical and Non-Technical losses

Energy theft detection [34] can be achieved using data from smart meters and power system state information, such as node voltages. When relying solely on smart meter data, detection methods can be categorized into supervised and unsupervised learning approaches:

- **Supervised Learning**:

  These methods require labelled training data to build classifiers that distinguish between normal and malicious consumption patterns. For example:

  - In [35], distribution transformer meters are used to first pinpoint areas with a high likelihood of energy theft, followed by anomaly detection to identify suspicious customers.
  - A top-down scheme uses a decision tree to estimate expected electricity use from features such as appliance count, occupancy, and outdoor temperature, followed by an SVM that labels customers as normal or malicious based on that estimate [36].
- **Unsupervised Learning**:

  Unsupervised methods are particularly valuable when labelled datasets are unavailable, as they do not require predefined training data. For instance:

  - An optimum-path forest clustering method models clusters as Gaussians; a load profile is flagged as anomalous when its distance to the cluster centre exceeds a threshold, performing competitively with k-means, BIRCH, affinity propagation, and Gaussian mixture models [37].

The practical benefits of smart meters in detecting energy theft are significant. For example, BC Hydro reported in 2013 that electricity theft accounted for approximately 7% of residential load before deploying smart meters [38]. With 1.9 million customers, they estimated an annual revenue increase of $802 million, attributing this improvement to smart meters' ability to enhance theft detection. To reduce non-technical losses, A data-fusion approach leverages smart-meter data and external databases to characterise customers' usage behaviour and location attributes [39]. These features are then used in a supervised machine learning algorithm, such as XGBoost, to model and evaluate non-technical losses [39]..

2.1.5 Load forecasting

Algorithms at the MV/LV network level must address the high variability in consumption, which is captured through smart meters providing granular time-interval data. This detailed data enables trend and cycle analysis, as well as time-of-day consumption patterns [40]. Additionally, time-interval-based consumption data allows for profiling consumer behaviour and correlating it with external factors like weather, geography, and demographic information.

One significant limitation of short-term forecasting models is their failure to incorporate detailed consumer consumption profiling. Using smart meter data, end-users can be grouped into categories based on distinct demand profiles, such as peak seasonal demand variations. Customer categorization, extensively studied, typically involves machine learning techniques, beginning with feature extraction and followed by clustering and classification. The methods used for feature extraction and classification depend on the resolution and volume of data available and may include techniques like Fast Fourier Transform (FFT), followed by Support Vector Machines (SVM), decision trees, or deep neural networks.

For instance, [41] reports on short-term forecasting using hourly load data from a Belgian grid substation, emphasizing the interdependence of forecasting and customer profiling. The study proposes a unified framework that integrates both processes. Initial modelling is based on seasonal time-series analysis using a periodic auto-regression model. Stationary properties derived from these models are then processed through K-means clustering to identify distinct customer profiles.

2.1.6 Load Disaggregation

Monitoring the energy consumption of individual appliances using dedicated sensors in homes or commercial buildings is often impractical and costly, particularly given the growing number of electrical devices in these environments. In contrast, energy disaggregation through Non-Intrusive Load Monitoring (NILM) provides a feasible alternative. This software-based, computational approach separates the total energy consumption measured by a single electricity meter into the usage of individual appliances.

While most NILM research has focused on high-frequency load measurements, the lower sampling rates offered by smart metering systems (ranging from 1 second to 1 hour) have spurred interest in developing low-rate NILM techniques. However, low-rate NILM presents unique challenges, such as noise from untracked appliances, signal transients that act as noise, load fluctuations, and the complexity of disaggregating loads in households that typically have more than 40 devices.[6]

2.2 Electrical Devices and Electrical Features

Understanding the electrical characteristics of household devices and selecting appropriate features are central to the effectiveness of Non-Intrusive Load Monitoring (NILM). Hart's pioneering work [12] introduced the concept of disaggregating total household energy consumption into individual appliance signatures using steady-state and transient power changes. Since then, various NILM review studies have categorized electrical features into steady-state, transient, and temporal attributes [18] , [26] . In low-frequency NILM—typically operating at a sampling rate of 1 Hz or lower, as seen in most smart metering infrastructures—transient details are often lost, making steady-state features (such as real power and ΔP) and temporal usage patterns (e.g., time-of-day, duration, and repetition) more critical [19] , [42]. Appliances differ in their load profiles, with simple on/off devices (like kettles) being easier to detect than multi-state or continuously variable devices (like washing machines or TVs) [18] . Advances in signal processing and feature selection techniques, including correlation analysis, mutual information, and graph-based models, continue to improve NILM accuracy and computational efficiency at low sampling rates [22] , ensuring the feasibility of disaggregation in real-world smart meter deployments.

2.2.1 Electrical Devices

Electrical devices are broadly categorized based on their operational characteristics, energy consumption patterns, and electrical signatures. Their classification significantly impacts the performance of NILM algorithms.

a. Categories of Electrical Devices

1. Resistive Loads: Devices like heaters, toasters, and incandescent bulbs primarily produce heat or light. These loads have a simple power signature with minimal variation over time, often represented as steady-state signals [12].
2. Inductive Loads: Appliances such as refrigerators, washing machines, and air conditioners operate with motors or compressors. These devices exhibit transient signals during startup and steady-state behaviour during operation [43].
3. Electronic Loads: Devices like computers, televisions, and LED lights incorporate power electronics, leading to complex waveforms with harmonic distortion [17].
4. Hybrid Loads: Certain appliances, such as washing machines and dishwashers, exhibit a mix of resistive, inductive, and electronic characteristics, making their disaggregation more challenging [21] .

b. Appliance Operating States

Devices operate in various states, including standby, active, and off modes. Multi-state appliances, like washing machines, undergo distinct operational cycles (e.g., washing, rinsing, spinning) that must be accurately identified for effective NILM [45].

c. Temporal Behaviour

Electrical devices have varying operational patterns. For instance:

- Periodic Usage: Refrigerators cycle on and off periodically [14].
- Scheduled Usage: Washing machines operate based on user-defined schedules [18]
- Random Usage: Lights and chargers may turn on or off at irregular intervals [12].

2.2.2 Electrical Features

Electrical features are the key attributes extracted from power signals to distinguish between devices. These features play a pivotal role in load disaggregation and are selected based on the type of devices and the NILM algorithm employed.

a. Fundamental Features

1. Active Power (P): The real power consumed by an appliance, measured in watts, is often the primary feature for NILM [17].
2. Reactive Power (Q): The power associated with magnetic fields in inductive devices, measured in vars, provides additional discriminative information [43].
3. Apparent Power (S): The combined magnitude of active and reactive power, offering insights into total power consumption [12].
4. Power Factor (PF): The ratio of active power to apparent power indicates the efficiency of an appliance [21] .

b. Time-Domain Features

1. Transient Signals: Startup and shutdown characteristics, such as spikes or dips in power, help identify inductive and hybrid loads [18] .
2. Steady-State Signals: Average power levels during normal operation provide a baseline for appliance identification [14].
3. Duty Cycle: The ratio of active time to total time for devices with cyclical behaviour, such as refrigerators [21] .

d. Advanced Features

1. Edge Detection: Sudden changes in power signals indicating state transitions, such as turning an appliance on or off [43].
2. Event Detection: Identifying specific operational events, such as the activation of a motor or compressor [21] .
3. Load Signature Profiling: Unique power signatures that combine multiple features (e.g., active power, harmonics, transients) to create a comprehensive profile for each appliance [18] .

2.2.3 Challenges and Opportunities

a. Challenges

Despite significant progress in NILM research, several challenges persist in accurately identifying appliances, particularly in low-frequency settings:

- **Feature Overlap**: Devices with similar power consumption profiles—such as microwaves and kettles—can lead to misclassification due to overlapping steady-state features [14], [22] .
- **Low-Power Devices**: Minor appliances or those in standby mode often generate signals that are obscured by larger concurrent loads, making them difficult to detect in aggregate data [21] , [46].
- **Dynamic Behaviour**: Appliances with variable operating states or unpredictable usage patterns (e.g., washing machines or HVAC systems) introduce complexities in modelling and classification [45], [47].
- **Noise and Interference**: Measurement noise, signal interference, and background loads reduce the signal-to-noise ratio, limiting the effectiveness of feature extraction and increasing false positives [17].

b. Opportunities
These challenges have also driven innovation in NILM algorithms and system design, particularly through the following opportunities:

- **Machine Learning**: Advanced models, particularly deep learning architectures, have demonstrated the ability to learn nonlinear and complex feature relationships, improving classification even under noisy or overlapping conditions [43].
- **Transfer Learning**: Cross-domain feature reuse—where models trained on one household or dataset are adapted to others—can improve generalizability and reduce the dependence on labelled data [18] .
- **Hybrid Models**: Integrating time-domain, frequency-domain, and contextual features can lead to more robust NILM systems capable of adapting to diverse appliance behaviours and sampling rates [21] , [48].

2.3 Non-intrusive load Monitoring

Research in load disaggregation originated in 1992 with the pioneering work of George W. Hart [12], who introduced the foundational concept of Non-Intrusive Load Monitoring (NILM). Since then, substantial progress has been made, with studies addressing various components and stages of NILM systems, including appliance modelling, feature extraction, classification, and evaluation.

An early line of work used state-based appliance models, similar to ILM. Using HMMs on the ACS-F2 dataset, appliance operating states were recognised with 97.78% overall accuracy, though refrigerators were often confused with ovens due to overlapping power signatures [5], [49]. This underscores HMMs' strength at modelling sequential behaviour and the challenge of separating devices with similar loads.

With growing interest in deep learning, CNNs were explored for appliance classification using switch-on transients: each detected event is windowed to the first 12 post-activation samples, assuming appliances exhibit distinctive transient signatures [50]. Training and testing used REDD (six homes, 1 Hz mains data) [51], targeting seven appliance types spanning on/off, multilevel, and variable-load classes.. The CNN model achieved an average accuracy of 82.43% and an F-score of 82.46%, demonstrating the effectiveness of transient-based classification. However, the approach required high sampling resolution, which may limit its deployment on low-rate smart meters.

To overcome the reliance on labelled data and high-frequency sampling, Zhao, Stankovic, and Stankovic [17] introduced a blind, low-rate NILM approach based on Graph Signal Processing (GSP). Their method does not require prior training and operates in three main stages: (i) determining event thresholds, (ii) clustering detected events, and (iii) feature matching. The algorithm is fully deterministic and uses time-series active power data exclusively, requiring expert input only during the final appliance labelling phase. Performance was validated on both the REDD dataset [51], downsampled to one-minute intervals, and the UK REFIT dataset [52]. The GSP-based method achieved over 75% accuracy for most appliances in the REDD dataset but underperformed for appliances like stoves, heaters, and lights. In the REFIT dataset, strong performance was observed for the kettle and refrigerator, while televisions were misclassified due to overlapping signatures with the freezer.

In general, NILM methodologies can be categorized into supervised and unsupervised techniques, with hybrid and semi-supervised methods also emerging. Supervised NILM relies on labelled training data and typically includes event detection (e.g., ON/OFF switching), feature extraction, and appliance classification using trained models. Recent supervised approaches have leveraged advanced methods such as GSP with Decision Trees [22] , Support Vector Machines (SVM) combined with K-means clustering [53], and Deep Neural Networks (DNN) [42], each offering different trade-offs in terms of computational efficiency and classification accuracy.

In contrast, unsupervised NILM techniques eliminate the need for labelled data and often focus on pattern similarity or clustering. Notable examples include Dynamic Time Warping (DTW) [54] for aligning time-series data and unsupervised GSP [17], which clusters appliance events based on signal characteristics. These methods are particularly valuable in practical settings where appliance-level labelling is not feasible, offering scalable solutions for real-world NILM deployment.

2.3.1 Machine Learning

Machine learning refers to computer algorithms that automatically improve through experience and is considered a subset of artificial intelligence. These algorithms create models based on sample data, referred to as "training data," to make predictions or decisions. Machine learning is extensively applied in NILM (Non-Intrusive Load Monitoring).

For example, Kamat utilized fuzzy pattern recognition theory to calculate the operating time of electrical equipment based on switching time information, ultimately completing the load disaggregation process [55]. Similarly, in 2005, Srinivasan et al. employed artificial neural networks to identify various devices, comparing the performance of Multi-Layer Perceptron (MLP), Support Vector Machines (SVM), and Radial Basis Function (RBF) networks [56].

This thesis investigates a range of machine learning techniques, encompassing both supervised and unsupervised approaches. The methods examined include Artificial Neural Networks (ANN), Decision Trees (DT), K-Nearest Neighbours (KNN), K-means clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and Graph Signal Processing (GSP). A concise overview of each technique is provided below.

1. **Supervised learning**
   - **Artificial neural networks (ANN)** are computational models inspired by biological neural networks, consisting of interconnected units or nodes known as artificial neurons. These neurons mimic the behaviour of biological neurons in the brain. Each connection between neurons, akin to synapses in a biological brain, transmits signals to other neurons. When a neuron receives a signal, it processes the input and may

transmit a signal to connected neurons. The transmitted signal is represented by a real number, and the output of a neuron is determined by applying a non-linear function to the sum of its inputs. The connections between neurons, referred to as edges, have adjustable weights that influence the strength of the transmitted signal. These weights are modified as the network learns, increasing or decreasing the signal's impact. Neurons can also have thresholds, allowing them to transmit signals only when the combined input exceeds a certain value. Typically, neurons are organized into layers, with each layer performing specific transformations on the input data. Signals flow through the network starting at the input layer, passing through hidden layers (potentially multiple times), and eventually reaching the output layer[57].

- **K-nearest Neighbours (k-NN)** classifier operates on the assumption that all instances can be represented as points in an n-dimensional space. During the learning phase, it retains all instances in memory. When classifying a new point, the k closest points to it are identified and used, with weights assigned based on their proximity, to determine the class of the new point. To improve accuracy, closer points are given higher weights [58].
- **Decision Tree (DT)** Classifier builds a tree structure based on instances to make classifications. The decision tree consists of two types of nodes:
  a) Root and internal nodes – these are associated with attributes.
  b) Leaf nodes – these represent classes.
  Each non-leaf node has an outgoing branch for every possible value of the attribute it represents. To classify a new instance using a decision tree, the process starts at the root node and moves through successive internal nodes until a leaf node is reached. At each node, a test is performed, and the result determines which branch to follow and which node to visit next. The class assigned to the instance is the one associated with the final leaf node [59].

2. **Unsupervised learning**

- **K-means** is an unsupervised machine learning algorithm used for clustering data points into distinct groups based on their features. It aims to minimize the distance between data points within a cluster while maximizing the separation between clusters[60].
- **Density-based spatial clustering of applications with noise (DBSCAN)** is a density-based clustering algorithm used to identify clusters of varying shapes and sizes in data, as well as outliers or noise. Unlike K-Means, DBSCAN does not require specifying the number of clusters in advance and can find arbitrarily shaped clusters[61]
- **Graph Signal Processing(GSP)**is a computational framework that generalizes classical signal processing techniques to data represented on irregular, non-Euclidean domains, such as graphs. In GSP, data points are treated as signals residing on the nodes of a graph, with edges representing relationships or connections between the nodes[62].

2.4 Gaps in Transferability and Scalability in NILM

Non-Intrusive Load Monitoring (NILM) has significant potential for energy disaggregation in smart grids, but its applicability in real-world scenarios is often limited by issues of transferability and scalability. Below are the key considerations and gaps:

2.4.1 Transferability

- **Dataset Dependence**: Most NILM algorithms are developed and evaluated using specific datasets, often collected under controlled conditions. These algorithms may struggle when applied to new datasets with different household settings, appliance types, or consumption patterns [21].
- **Regional Variations**: Energy consumption behaviour, appliance types, and power grid characteristics vary significantly across regions. Algorithms trained in one geographic location may perform poorly when applied elsewhere [63].
- **Generalization to Unknown Appliances**: Many NILM approaches rely on prior knowledge of appliance signatures, limiting their ability to disaggregate loads from previously unseen or new appliances [2].

2.4.2 Scalability

- **Computational Complexity**: Many NILM algorithms, particularly those using deep learning, require significant computational resources. Scaling these methods to larger datasets or networks with millions of users can be computationally expensive [64].
- **Data Volume**: High-resolution smart meter data generates massive volumes of information. Efficiently processing and storing this data becomes challenging as the number of households increases [65].
- **Distributed Deployment**: Implementing NILM at scale often requires deploying algorithms across distributed systems, such as smart meters at the edge. Many current methods are not optimized for edge computing environments [66].

2.4.3 Data Privacy and Security

- **User-Specific Models**: To achieve high accuracy, NILM often tailors models to individual households. This reliance on personalized data can raise privacy concerns and complicate deployment at scale [13].
- **Privacy-Preserving Methods**: While some techniques, such as federated learning, aim to preserve privacy, their integration into NILM systems remains in its infancy [67].

2.4.4 Robustness to Noise and Data Quality

- **Data Quality Variability**: Real-world smart meter data often contains noise, missing values, or inconsistencies. Algorithms designed for ideal datasets may fail to deliver accurate results under these conditions [65].
- **Interference from Unknown Appliances**: Noise introduced by untracked or unidentified appliances can significantly degrade NILM performance, particularly in large-scale deployments [14].

2.4.5 Integration with Emerging Technologies

- **IoT and Smart Home Devices**: NILM systems must adapt to increasingly complex household environments with numerous IoT and smart devices. These devices may introduce additional noise or non-standard consumption patterns [63].
- **Distributed Energy Resources (DERs)**: The rise of renewable energy sources and home energy storage systems adds complexity to load disaggregation tasks [66].

2.4.6 Benchmarking and Standardization

- **Lack of Standard Benchmarks**: NILM research lacks universally accepted benchmarks for evaluating scalability and transferability. This makes it difficult to compare algorithms and assess their real-world applicability [21].
- **Cross-Dataset Testing**: Few studies test algorithms on multiple datasets to demonstrate their robustness and transferability across different conditions [64].

2.5 Real-Time NILM

Real-time NILM involves the immediate or near-immediate disaggregation of total power consumption data into appliance-specific energy usage. This capability enables utilities and consumers to gain actionable insights into energy usage patterns, identify inefficiencies, and respond promptly to energy events.

2.5.1 Key Features of Real-Time NILM

1. **Low Latency Processing**:
   - Real-time NILM systems process incoming data with minimal delay, typically leveraging high-frequency or low-frequency smart meter data (e.g., 1-second or 1-minute intervals).
2. **Streaming Data Analytics**:
   - Unlike batch processing, real-time NILM requires continuous analysis of data streams from smart meters or energy monitors.
3. **Immediate Feedback**:
   - Provides real-time insights for applications such as energy feedback systems, fault detection, demand-side management, and load balancing.

2.5.2 Methods and Technologies Used in Real-Time NILM

1. **Machine Learning and AI:**
   - Algorithms such as Neural Networks, Hidden Markov Models (HMMs), and Decision Trees are tailored for real-time inference and classification.
   - Techniques like sparse matrix optimization (e.g., sparse Viterbi algorithms in HMMs) help achieve computational efficiency for real-time applications [16].
2. **Edge Computing:**
   - Deploying NILM on edge devices like smart meters, Raspberry Pi, or ARM processors minimizes the need for centralized data processing, reducing latency [68].
3. **Streaming Architectures:**
   - Frameworks like Apache Kafka or Apache Flink are used for handling continuous data streams in large-scale real-time NILM deployments.
4. **Efficient Hardware:**
   - Hardware optimizations ensure that NILM algorithms run on low-power devices without sacrificing accuracy. For example, lightweight implementations have been demonstrated on embedded systems like Raspberry Pi [69].

2.5.3 Challenges in Real-Time NILM

1. **Data Noise:**
   - Noise from unknown appliances or environmental factors can impact the accuracy of real-time disaggregation [21].
2. **Scalability:**
   - Processing data from thousands or millions of households in real time requires highly efficient algorithms and robust infrastructure [63].

3. **Accuracy:**
    - Achieving high accuracy in real-time NILM is challenging, especially with low-resolution data or overlapping appliance signatures [14].
4. **Hardware Constraints:**
    - Resource-limited devices may struggle to run computationally intensive NILM algorithms efficiently [16].

2.5.4 Applications of Real-Time NILM

1. **Energy Feedback:**
    - Provides users with immediate insights into their energy usage, enabling them to modify behaviour and reduce consumption.
2. **Fault Detection:**
    - Identifies appliance malfunctions or unusual energy usage patterns as they occur [69].
3. **Demand-Side Management:**
    - Helps utilities balance load and optimize energy distribution by understanding real-time consumption patterns [67].
4. **Smart Home Automation:**
    - Integrates with IoT systems to enable automated responses, such as turning off appliances during peak loads [66].

The advancements in real-time Non-Intrusive Load Monitoring (NILM) have been driven by significant contributions across multiple domains. Stephen Makonin's work on efficient HMM-based algorithms has set a benchmark for real-time implementation on embedded hardware, demonstrating practical and scalable solutions [16]. Similarly, the adoption of deep learning techniques, including CNNs and RNNs, has enhanced the accuracy and efficiency of energy disaggregation, enabling real-time applications in dynamic environments [69]. Moreover, open-source initiatives like NILMTK and Sparse NILM have provided accessible platforms for researchers and developers, fostering innovation and collaboration in NILM system development [21]. Together, these contributions have significantly advanced the field, paving the way for widespread deployment and integration of real-time NILM in modern energy management systems.

2.6 Performance Evaluation Metrics for NILM

Evaluating NILM systems requires metrics that assess accuracy, efficiency, scalability, and practicality in real-world scenarios. Here’s an expanded and detailed breakdown:

2.6.1 Disaggregation Accuracy Metrics

These metrics measure how effectively NILM systems separate total energy consumption into appliance-specific loads.

- **Mean Absolute Error (MAE)**:

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|y_t - \hat{y}_t| \tag{2.1}$$

Purpose**:** Captures the average absolute error in predicted appliance loads ($\hat{y}_t$) compared to actual loads (y_t)[69].

- **Root Mean Squared Error (RMSE):**

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2} \tag{2.2}$$

Purpose: Weighs larger errors more heavily, highlighting models sensitive to extreme disaggregation deviations [21].

- **Signal Aggregate Error (SAE):**

$$SAE = \frac{|\sum_{t=1}^{T} y_t - \sum_{t=1}^{T} \hat{y}_t|}{\sum_{t=1}^{T} y_t} \qquad (2.3)$$

Purpose: Evaluates whether the model accurately predicts the cumulative energy consumption of each appliance [13].

- **Normalized Disaggregation Error (NDE):**

$$SAE = \frac{\sum_{t=1}^{T} (y_t - \hat{y}_t)^2}{\sum_{t=1}^{T} {y_t}^2} \qquad (2.4)$$

- **Match Rate**

Match rate quantifies the similarity between the estimated and ground truth power signals over a time series. It is particularly useful in evaluating continuous power estimation accuracy. The match rate is defined as:

$$Match\ Rate = 1 - \frac{\sum_t |\hat{P}_t - P_t|}{\sum_t P_t} \qquad (2.5)$$

where:

- $\hat{P}_t$ is the estimated power consumption at time t,
- P_t is the true power consumption at time t,
- and the denominator normalizes the absolute error by the total actual energy.

A higher match rate indicates better performance, with a value close to 1 signifying near-perfect disaggregation.

- **Accuracy (ACC)**

Accuracy (ACC) is often used in event-based NILM to evaluate how well the algorithm classifies the ON/OFF status of appliances. It is defined as:

$$Accuracy(ACC) = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2.6)$$

where:

TP is the number of true positives (correct ON-state detections),

TN is the number of true negatives (correct OFF-state detections),

FP is the number of false positives (incorrect ON-state detections),

FN is the number of false negatives (missed ON-state detections).

This metric reflects the overall correctness of appliance state classification. However, in scenarios with class imbalance (e.g., where OFF states dominate), additional metrics such as **precision**, **recall**, and **F1-score** may offer more informative insights[70].

2.6.2 Classification Metrics

Event detection can be treated as a binary classification problem, where each sample is classified into one of two categories. The outcomes of this classification can be divided into four cases: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). These results are typically summarized in a confusion matrix, as illustrated in Table 2.1.

Table 2.1: Confusion Matrix

	Positive	Negative
Positive Actual	TP	FN
Negative Actual	FP	TN

In the context of Non-Intrusive Load Monitoring (NILM), the confusion matrix is a valuable tool for evaluating the performance of appliance classification algorithms. It illustrates the correspondence between actual appliance states and those predicted by the model, where each row represents the true state and each column corresponds to the predicted state. Correct classifications appear along the diagonal, while off-diagonal entries reflect errors such as false positives (predicting an appliance is ON when it is OFF) or false negatives (failing to detect an actual ON state). This distinction is particularly important in NILM applications, where certain appliances (e.g., high-energy devices) may have greater impact on system performance and energy feedback. By analysing the confusion matrix, researchers can better understand algorithmic biases and improve detection accuracy accordingly[22]

Precision evaluates the accuracy of positive predictions. It is defined as the probability that samples predicted to be positive are actually positive. In the context of appliance activity, Precision (PR) represents the proportion of correctly predicted active appliance states to the total number of predicted positive states. The formula for calculating Precision is provided below[69]:

$$PR = \frac{TP}{TP+FP} \qquad (2.7)$$

Recall measures the ability to identify positive samples from the original dataset. It is defined as the probability that a sample that is actually positive is correctly predicted as positive. In the context of appliance activity, Recall (RE) represents the proportion of correctly predicted active appliance states to the total number of actual positive states. The formula for calculating Recall is provided below:

$$RE = \frac{TP}{TP+FN} \qquad (2.8)$$

The F-Score is the weighted harmonic mean of Precision and Recall, designed to balance the trade-off between these two metrics. Precision and Recall often exhibit an inverse relationship—when one is higher, the other tends to be lower. The F-Score provides a single measure that combines both metrics. It is calculated as the harmonic mean of Precision and Recall. The formula for computing the F-Score is provided below[21][71]:

$$F\_score = \frac{2*PR*RE}{PR+RE} \qquad (2.9)$$

A more detailed NILM evaluation splits true positives into accurate true positives (ATP) and inaccurate true positives (ITP) [17], [71]. An ATP occurs when both the usage event and the appliance identity are correctly detected, while an ITP occurs when the usage event is detected but the appliance is misclassified. False positives (FP) are activations reported when none occurred, and false negatives (FN) are real activations that were missed. The table below summarises the differences among ATP, ITP, FP, and FN

Table 2.2 : NILM event-classification example showing ATP, ITP, FP, and FN (framework per [17], [71])

Event Detected	Actual Appliance	Predicted Appliance	Classification
ON event at 07:45	Kettle	Kettle	ATP
ON event at 08:10	Washing Machine	Dishwasher	ITP
ON event at 09:00	—	Kettle	FP
Missed ON event at 09:30	Microwave	—	FN

Precision (PR) measures how many of the detected events were correctly identified and is calculated as ATP divided by ATP plus FP. A high PR suggests the system produces fewer false alarms. Recall (RE), calculated as ATP over the sum of ATP, ITP, and FN, reflects the ability to correctly capture all relevant events, including accounting for mislabelling and missed detections. The F-measure (Fm) combines both PR and RE into a single score to provide a balanced evaluation of classification performance. Notably, Fm focuses exclusively on state detection (i.e., whether an appliance turned ON or OFF), rather than measuring the accuracy of energy consumption estimates. For evaluating how closely the predicted power values match the actual usage, disaggregation accuracy is used instead.[17][EU].

$$PR = \frac{ATP}{ATP+FP} \tag{2.10}$$

$$RE = \frac{ATP}{ATP+\mathrm{ITP}+FN} \tag{2.11}$$

$$FM = \frac{2*PR*RE}{PR+RE} \tag{2.12}$$

These refined metrics account for misclassification and missed detections. Note that Fm targets ON/OFF state detection only, while disaggregation accuracy focuses on estimated power values. In Chapter 3&4, the term F-score is used instead of Fm, following standard classification terminology.

2.6.3 Justification for Selecting F-Measure (Fm)

Evaluating NILM systems requires metrics that capture both correctness and application-specific priorities. Given this thesis's focus on anomaly detection, the contribution chapters adopt the F-measure (Fm) as the principal metric. This choice is supported by theory and prior work [17], [71], and is justified below against alternative metrics.

Relative Merits of F-Measure (Fm)

- **Balances Precision and Recall**:

  Fm is the harmonic mean of Precision (PR) and Recall (RE), offering a balanced evaluation between false alarms (FP) and missed detections (FN). In NILM-based anomaly detection, it is crucial not only to detect anomalies accurately (high Recall) but also to avoid frequent false alerts (high Precision) [72].

- **Adaptable to Multi-State Appliance Detection**:

  As shown in [17], the F-measure (Fm) is well suited to event-based NILM—where appliances can be multi-state and detection hinges on ON/OFF transitions—and remains sensitive to both misclassifications and omissions, which is critical for fault identification.

- **ATP/ITP-Aware Variant Improves Granularity**:

  The variant used in this thesis computes precision and recall with accurate true positives (ATP) and inaccurate true positives (ITP), offering finer granularity than a standard F1-score. This is especially valuable for anomaly detection, where assigning a detected event to the wrong appliance (ITP) can mislead diagnostics [17], [71].

- **Suitable for Imbalanced Class Distributions**:

  Appliance usage is often imbalanced—e.g., fridges generate frequent ON events while ovens are rarely active—so Accuracy (ACC) can be inflated by the dominant OFF class; the F-measure (Fm) mitigates this optimism by balancing precision and recall [12].

Table 2.3: Disadvantages of Fm and Why They Are Acceptable

Disadvantage	Implication	Justification for Use
Ignores True Negatives (TN)	May overlook performance on OFF-state detection	In anomaly detection, missed or false detections (FP, FN) are of higher operational concern
Focuses only on binary transitions	Does not reflect energy magnitude accuracy	This thesis uses Fm for event detection and supplements it with disaggregation metrics (e.g., SAE, MAE)
Sensitive to class imbalance	May yield low Fm for rare events despite detection	This challenge is embraced intentionally, since detecting rare events is the goal in anomaly detection

In this thesis, Chapters 5 introduce NILM-based methods for appliance-level anomaly detection, which require precise identification of abnormal appliance behaviour (e.g., unusual cycles, persistent operation, or consumption spikes). These events are typically sparse and high-risk, meaning:

- **Recall** is critical to detect all potential anomalies (minimize FN),
- **Precision** is critical to avoid unnecessary alarms that reduce user trust (minimize FP),
- And thus **Fm**, as a balanced and interpretable score, is ideal.

Adopting the ATP/ITP-aware F-measure (Fm) aligns with the goal of verifying the appliance's identity—not just its operational state—since mislabelling a faulty fridge as a microwave could prompt inappropriate or delayed action [17], [71].

While Fm does not capture every facet of NILM performance, it is the most relevant metric for the event-based, class-imbalanced, and safety-critical anomaly-detection setting considered here. Its adoption is consistent with prior work [17], [71] and aligns with the problem's demands, where balanced detection of sparse, high-impact events is paramount.

2.7 Anomaly detection

Faulty or malfunctioning electrical appliances pose significant risks to both safety and energy efficiency in residential buildings. According to recent reports by the BBC and the Scottish Fire and Rescue Service, malfunctioning white goods were linked to nearly 12,000 domestic fires in Great Britain over just three years, with incidents occurring almost daily in 2019 [30]. Faults such as refrigerant leaks, compressor failure, or user-related issues like leaving refrigerator doors open can lead to excessive energy consumption and, in some cases, permanent equipment damage [9].

As household electricity demand continues to rise—projected to increase by approximately 60% from 2017 to 2025 and 70% from 2025 to 2040—ensuring appliances operate efficiently and safely is increasingly critical [74]. Studies suggest that energy savings of up to 12% can be achieved through timely detection of anomalies, making fault detection a priority in energy management strategies [74].

Traditionally, anomaly detection has relied on sub-metering, where each appliance is individually monitored. While this offers detailed insights, it becomes impractical in modern households with more than 40 appliances due to cost, complexity, and scalability concerns [9], [75], [76], [77]. As a result, there is growing interest in Non-Intrusive Load Monitoring (NILM) as a more scalable solution. NILM enables the inference of appliance-level behaviour from aggregate smart meter data, eliminating the need for widespread submeter deployment.

Recent efforts have explored both unsupervised and supervised approaches to detect anomalies using NILM. These include statistical models of normal appliance behaviour, clustering methods, and classification algorithms. For example, some studies identify abnormal energy patterns based on the frequency or duration of ON/OFF cycles, while others classify consumption into normal and anomalous categories such as "excessive usage" or "usage during absence" [9], [78]. Despite the progress, NILM-based anomaly detection still faces challenges due to noise, feature overlap, and limited signal granularity in low-frequency smart meter data. This highlights the need for anomaly-aware NILM algorithms and robust post-processing techniques to improve accuracy and scalability in real-world environments.

However, existing research still leaves several important gaps that motivate the core contributions of this thesis. First, while feature engineering has been widely discussed in NILM, there is limited systematic evaluation of which features are most effective for different appliance types and operating conditions. This lack of consensus on robust feature-selection strategies motivates the work presented in Chapter 3, where comparative analyses of feature extraction and selection methods are conducted.

Second, although both supervised and unsupervised NILM algorithms have been developed, their relative performance in terms of accuracy, interpretability, and computational efficiency has not been fully established, particularly when applied to real-world smart-meter data. To address this, Chapter 4 provides a detailed comparison between supervised and unsupervised NILM algorithms, highlighting trade-offs in algorithmic complexity, generalisation ability, and robustness against data noise.

Finally, anomaly detection in household appliances remains underexplored, especially in the context of NILM outputs. Prior work often relies on rule-based thresholds or direct submetered data, which limits practical scalability. To overcome these limitations, Chapter 5 focuses on detecting anomalous behaviour in appliances such as fridge-freezers using NILM-derived signals, with an emphasis on statistical methods and unsupervised clustering approaches to improve reliability and safety insights.

Collectively, these three chapters build a comprehensive framework for anomaly-aware NILM by bridging gaps in feature selection, algorithmic evaluation, and post-processing for appliance anomaly detection.

2.8 Summary of Research Gaps and Link to Contribution Chapters

This chapter has reviewed the existing literature on Non-Intrusive Load Monitoring (NILM), with particular emphasis on feature selection, load disaggregation, and anomaly detection using smart-meter data. While significant progress has been achieved, the review highlights several unresolved gaps that limit the robustness, transferability, and real-world applicability of NILM systems, particularly under low-frequency smart-meter constraints.

First, although feature selection is widely recognised as a critical component of both supervised and unsupervised NILM frameworks, most existing studies rely on heuristic or dataset-specific feature choices. Feature evaluations are typically conducted within a single dataset and are often tailored to specific appliances or operating conditions. As a result, there is limited understanding of which features remain informative and robust across different datasets, households, and sampling rates. This lack of systematic, cross-dataset feature analysis restricts transferability and increases computational overhead in low-frequency smart-meter environments.

Second, the NILM literature presents a wide range of load disaggregation techniques based on supervised and unsupervised machine learning methods. However, comparative evaluations are frequently performed under heterogeneous assumptions, using different datasets, feature sets, pre-processing pipelines, and evaluation metrics. In particular, the impact of different pre-processing steps on disaggregation accuracy and robustness is not consistently analysed. Consequently, the relative suitability of NILM algorithms for low-frequency smart-meter data remains unclear. Trade-offs between algorithm performance, sensitivity to pre-processing choices, computational complexity, interpretability, and scalability are not systematically examined, despite their importance for practical deployment.

Third, although anomaly detection has been extensively studied in energy analytics, its integration with NILM remains limited. Many existing approaches treat anomaly detection as a standalone task applied to aggregate load profiles or rely on intrusive sub-metering. NILM-based anomaly detection studies are comparatively sparse and often depend on simple threshold-based rules or appliance-specific heuristics. While NILM can provide appliance-level information from low-frequency smart-meter data, there is a lack of systematic studies that exploit these outputs for anomaly detection using graph-based techniques. In particular, the comparative effectiveness of hybrid approaches that combine Unsupervised Graph Signal Processing (UGSP) with rule-based post-processing, versus fully unsupervised UGSP-based anomaly detection methods, remains underexplored. This gap is especially relevant in low-frequency settings, where overlapping appliance signatures, noise, and sparse event information complicate the reliable identification of abnormal behaviour.

These gaps collectively motivate the structure and focus of the subsequent chapters. Chapter 3 addresses the need for systematic and robust feature selection by evaluating feature relevance and performance across multiple datasets and learning paradigms. Chapter 4 responds to the lack of consistent algorithmic evaluation by providing a structured comparison of supervised and unsupervised NILM methods for load disaggregation under unified low-frequency conditions, with explicit analysis of the effects of different pre-processing strategies on algorithm performance and transferability across households. Chapter 5 builds upon NILM-derived appliance-level outputs and investigates anomaly detection by comparing a hybrid UGSP + rule-based framework with a fully unsupervised UGSP-based anomaly detection approach, thereby clarifying the role of UGSP in detecting abnormal appliance behaviour.

By addressing the limitations identified in this chapter, the following contribution chapters collectively advance NILM from a disaggregation-focused task toward a more robust, transferable, and anomaly-aware monitoring framework suitable for real-world smart-meter deployments.

Chapter 3

2 Feature selection and multi-class classification within a dataset

3.1 introduction

Feature selection and classification are essential processes in data analysis that enhance the performance of predictive models and identifying key patterns in a dataset. The data available for mining is typically raw and may exist in various formats due to its collection from different sources. Before applying data mining algorithms, several preprocessing steps are necessary [79]:

1. Data Integration:
    - When data is sourced from multiple origins, integration is required to ensure consistency. This involves resolving discrepancies in attribute names or values across datasets from different sources.
2. Data Cleaning:
    - This step focuses on detecting and correcting errors in the data, such as filling in missing values and addressing inconsistencies. Various data cleaning techniques are discussed in [80, 80].
3. Attribute Selection:
    - Not all attributes are relevant to the mining process. Attribute selection involves identifying and retaining a subset of attributes that are most relevant for the analysis, improving the efficiency of the mining process [82].

- **Feature selection**

Datasets often contain numerous irrelevant attributes that can negatively impact the mining process. Removing these attributes is essential, as many mining algorithms struggle to perform effectively with a large number of features. Therefore, applying feature selection techniques prior to using any mining algorithm is crucial. The primary goals of feature selection are to prevent overfitting, enhance model performance, and develop faster and more cost-efficient models. A comprehensive assessment of feature selection methods for NILM appliance classification was reported in [83]. They reviewed a wide array of features reported in the literature, categorizing them into steady-state features—such as active power, reactive power, and power factor—and transient features like transient root mean square current and transient duration. The authors emphasized that while steady-state features are straightforward to compute, they may not sufficiently distinguish appliances with similar power ratings. Conversely, transient features can enhance classification performance but often require high-frequency sampling and more complex hardware setups. To address these challenges, the study proposed a systematic feature elimination process aimed at identifying the most effective feature set for appliance classification. This process was validated on a large benchmark dataset, demonstrating improved classification accuracy across various appliances compared to using all features or randomly selected subsets.[83]

- **Classification**

The classification task is a supervised learning technique where each instance is assigned to a specific class, determined by the value of a target attribute, often referred to as the class attribute. This target attribute takes categorical values, each representing a class. An example comprises two components: a set of predictor attribute values and a target attribute value. The predictor attributes are used to forecast the target attribute's value, and they must be relevant for accurately predicting an instance's class.

The classification process divides the dataset into two distinct and exhaustive subsets: the training set and the test set. Correspondingly, classification occurs in two phases:

1. **Training Phase:**

   During training, the algorithm uses the training set, where both the predictor and target attribute values are available for all instances. This information is used to construct a classification model, which encodes the relationship between predictor attributes and classes. This model captures the knowledge necessary for predicting an instance's class based on its predictor attribute values.

2. **Testing Phase:**

   In the testing phase, the classification model is evaluated using the test set, where the class values are hidden from the algorithm during prediction. Once a prediction is made, the algorithm can access the actual class value to assess its accuracy.

The primary objective of a classification algorithm is to maximize the predictive accuracy of the model when applied to unseen examples in the test set.

In this thesis, various classification techniques are used, including Artificial Neural Networks (ANN), Decision Trees (DT), Density-Based Spatial Clustering of Applications with Noise (DBSCAN), K-Means Clustering, and K-Nearest Neighbours (K-NN).

3.2 Contributions over the state of the art

Feature selection has been widely recognised as a critical component of Non-Intrusive Load Monitoring (NILM), as it directly affects classification accuracy, robustness, and computational efficiency. Existing studies typically focus on a limited set of features evaluated on a single dataset or within a single learning paradigm, most commonly supervised classification. High-frequency features, transient signatures, and harmonic-based descriptors are often emphasised, despite their limited availability in smart-meter data. Within the research group, earlier work investigated low-rate NILM using Graph Signal Processing (GSP) and explored preliminary feature analysis for appliance classification; however, these studies did not include a systematic or cross-dataset comparison of feature selection strategies.

Despite the extensive literature, several limitations remain. First, there is a lack of systematic evaluation of feature selection methods across multiple datasets, particularly under low-frequency smart-meter constraints. Second, most studies focus exclusively on either supervised or unsupervised learning, with little attention to feature robustness across both paradigms. Third, computational efficiency and deployability on low-power devices are often treated as secondary considerations. These gaps limit the practical applicability of many proposed NILM solutions.

This chapter addresses these limitations by presenting a comprehensive and systematic evaluation of feature selection strategies for NILM across multiple public datasets. The contribution lies in analysing feature performance jointly for supervised and unsupervised algorithms, with explicit consideration of computational complexity and transferability. The results identify a compact and robust feature set suitable for low-frequency smart meters, directly supporting the thesis-wide objective of practical, deployable NILM solutions. This contribution aligns with the first thesis contribution outlined in Chapter 1, which emphasises systematic feature evaluation beyond prior work within the research group.

3.2 ML for two different and distinct datasets, both with the same problem of multi-class classification

ANN is used to classify first data sets that is spectrometry data and has been compared with SVM (that modelled by other) and ANN, DT, K-Means and DBSCAN are used for disaggregation kettle and oven from electrical data in REFIT House 17 and REDD House1.

3.2.1 Dataset 1 (spectrometry) feature selection and classification approaches

(Herrero-Bermello, A., Li, J., Khazaei, M., Grinberg, Y., Velasco, A. V., Vachon, M., ... & Schmid, J. H. (2019). On-chip Fourier-transform spectrometers and machine learning: a new route to smart photonic sensors. Optics letters, 44(23), 5840-5843.)[84]

Miniaturized silicon photonics spectrometers hold significant promise for use in diverse areas such as biomedical diagnostics, environmental monitoring, astrophysics, and telecommunications [85]. However, traditional spectrometer designs—such as arrayed waveguide gratings (AWGs) [86], waveguide echelle gratings [87], and cascaded micro-ring resonators [88]—often suffer from performance limitations. These stem from their sensitivity to fabrication imperfections, environmental fluctuations, signal-to-noise ratio (SNR) challenges, and restricted optical throughput.

Spatial heterodyne Fourier-transform (SHFT) spectrometers [89–90] have emerged as a promising solution, as they incorporate the advantages of conventional Fourier-transform (FT) spectrometers [91], including high SNR and throughput, while eliminating the need for mechanical components or active heating systems [92]. SHFT systems use arrays of Mach–Zehnder interferometers (MZIs) with varying optical path differences (OPDs) to generate spatial interferograms that are used to reconstruct the input spectrum [89]. Their structure supports passive calibration, which can partially compensate for fabrication-related variations [89, 92].

Nevertheless, a significant limitation of SHFT spectrometers lies in the thermal sensitivity of silicon waveguides. The thermo-optic coefficients are approximately 1.8×10^{-4} K^{-1} for TE and 1.2×10^{-4} K^{-1} for TM modes at 1.55 μm [93], making them highly susceptible to temperature-induced changes. Even small fluctuations (as little as 0.1°C) can alter the OPD, resulting in spectral distortion or drift . While a resolution of 42 pm [89] is sufficient for detecting gas-phase signatures [94], this thermal dependence poses a serious challenge for stable, real-world deployment of SHFT systems.

Some approaches have attempted to correct for temperature effects through temperature-dependent calibration schemes [93]. However, these methods typically require additional measurements and system complexity, limiting their practicality in field conditions. An alternative is offered by machine learning (ML), which excels at extracting meaningful patterns from complex and noisy datasets. ML has already demonstrated success in fields such as facial recognition [95], genomics [96], and traffic analysis [44], and it presents a compelling approach for addressing the thermal instability of SHFT outputs.

This thesis proposes a novel ML-driven framework for spectral recognition using SHFT spectrometers. Rather than reconstructing the input spectrum using temperature-calibrated transformation matrices, the system is trained to directly recognize specific absorption features from the raw interferogram. While some ML approaches have been used to improve matrix inversion in spectral reconstruction [73], they have not addressed thermal robustness explicitly. The proposed method uses supervised learning to associate interferogram patterns with corresponding spectral labels. Training is conducted over a 10°C temperature range using known spectra, and the model is evaluated on unseen data to assess its ability to generalize across varying thermal conditions.

Supervised transfer learning was implemented using artificial neural networks (ANNs), due to their capacity to handle high-dimensional input, minimal training complexity, and strong performance with limited data. Prior to training, a normalization step(pre-processing) was applied to scale the power signal to a standard range (e.g., 0 to 1), enhancing the stability and efficiency of the learning process [100]. The classification task involved four spectral classes—three with distinct absorption lines at 1549.8 nm, 1550.0 nm, and 1550.2 nm, and one reference with no absorption. Although simplified, this presents a challenging problem for Fourier-transform (FT) systems, which typically exhibit global responses to spectral variations. Each class consisted of 4655 samples, split evenly between training and testing sets. ANN models were trained using both Broyden–Fletcher–Goldfarb–Shanno (BFGS) and Levenberg–Marquardt backpropagation algorithms, with hidden layer sizes ranging from 10 to 50 neurons. Feature selection was conducted to determine which of the 28 interferogram values had the most significant impact on classification accuracy.

For the artificial neural network (ANN), training was performed using Broyden-Fletcher-Goldfarb-Shanno (BFGS) quasi-Newton backpropagation and Levenberg-Marquardt backpropagation. The number of hidden layers used in the experiments ranged from 10 to 50, with further optimization achieved through feature selection. This involved identifying which of the 28 optical power values contributed most significantly to accurately classifying the input spectra.

Classification accuracy in machine learning (ML) is commonly evaluated using the **F1 score**, which combines precision and recall metrics into a single value. Precision measures the ratio of correctly predicted positive observations to the total predicted positives (formula 2.10), while recall measures the ratio of correctly predicted positives to all actual positives (formula 2.11). The F1 score for each class is then calculated as formula 2.12 , with the results averaged across all four classes.

The results compared with those obtained using a support vector machine (SVM) with a Gaussian kernel that has been modelled by other.

In the initial experiment, classification was performed using the full vector of 28 spectrometer outputs, resulting in an F1 score accuracy of 77.6% for the SVM and 77.8% for the ANN, tested on a dataset of 2327 × 4 data points. The best ANN performance was achieved using Levenberg–Marquardt backpropagation with 35 hidden layers.

Subsequently, feature selection was applied to identify and exclude less informative measurements, improving classification accuracy. The improved results with feature selection demonstrated F1 score accuracies of 81.3% for the ANN and 82.5% for the SVM, evaluated on the same testing dataset of 2327 × 4 data points. As expected the confusion matrices reveal notable misclassification between Class 3 and Class 2, and between Class 2 and Class 1, highlighting strong similarities between these classes. However, Class 4 experienced minimal misclassification.

3.2.2 Dataset 2 (electrical measurements) feature selection and classification

To demonstrate the operation and performance of NILM, two supervised methods—Artificial Neural Network (ANN) and Decision Tree (DT)—and two unsupervised methods—DBScan and K-Means—were designed and evaluated to disaggregate two appliances: a kettle and an oven. The methodologies for ANN, DT, K-Means, and DBScan are presented as flowcharts in Figures 3.1, 3.2, 3.3(a), and 3.3(b), respectively. This study primarily focuses on disaggregating the oven, as it has received limited attention in NILM literature compared to other appliances like refrigerators, washing machines, dishwashers, toasters, microwaves, and kettles, despite being one of the major consumers of household electricity. The limited focus on ovens is largely due to their rare inclusion in submeters, their absence from most public electrical measurement datasets, and the resulting lack of ground truth data available to the research community. The kettle is also included in the study as a benchmark, given its heating element is similar to that of an oven but with a distinct electrical signature. Power and ON-duration were prioritised due to their robustness under low-frequency sampling, strong discriminative ability for major appliance types, and consistent availability across all NILM datasets. Unlike transient-based features, which require high-frequency signals that smart meters do not provide, power and temporal duration remain reliable even at 1-minute resolution. These features also directly support anomaly detection, where abnormal ON-durations or unexpected power levels are key indicators of faulty behaviour. Therefore, the focus on these two features is both methodologically justified and aligned with the operational constraints of smart metering infrastructure.

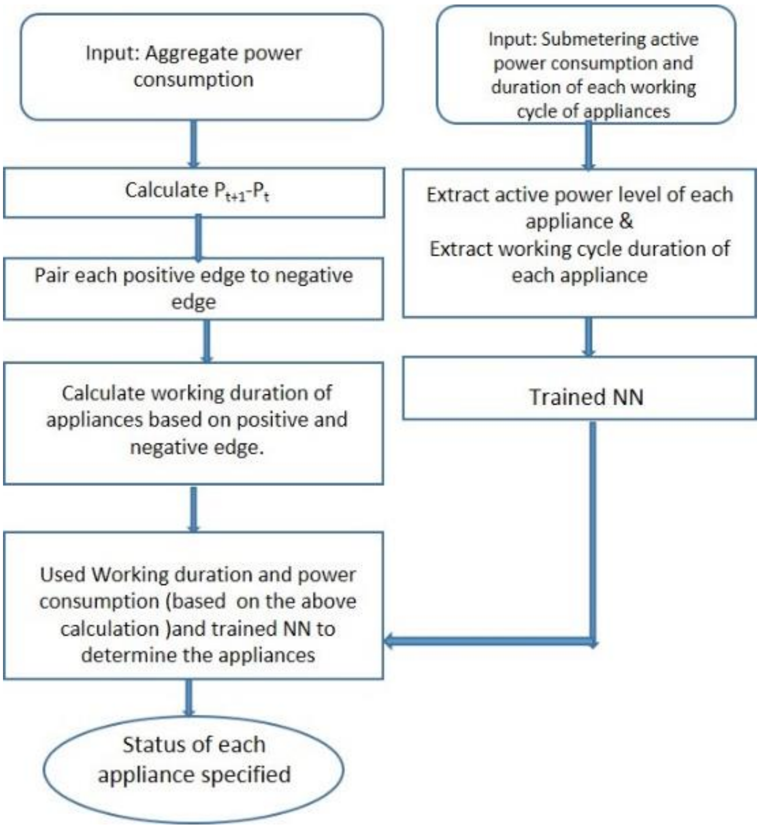

Input: Aggregate power consumption
Calculate P_{t+1}-P_t
Pair each positive edge to negative edge
Calculate working duration of appliances based on positive and negative edge.
Used Working duration and power consumption (based on the above calculation)and trained NN to determine the appliances
Status of each appliance specified
Input: Submetering active power consumption and duration of each working cycle of appliances
Extract active power level of each appliance &
Extract working cycle duration of each appliance
Trained NN

Figure 3.1: Flow chart of proposed supervised algorithms, each showing testing on the left and training on the right for: ANN, P_t denotes a power measurement at time instance t.

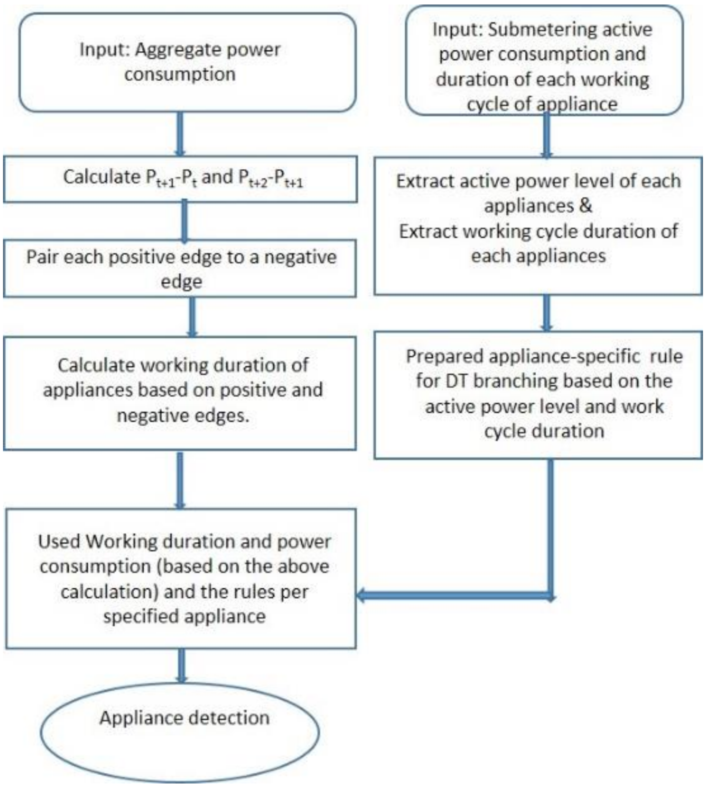
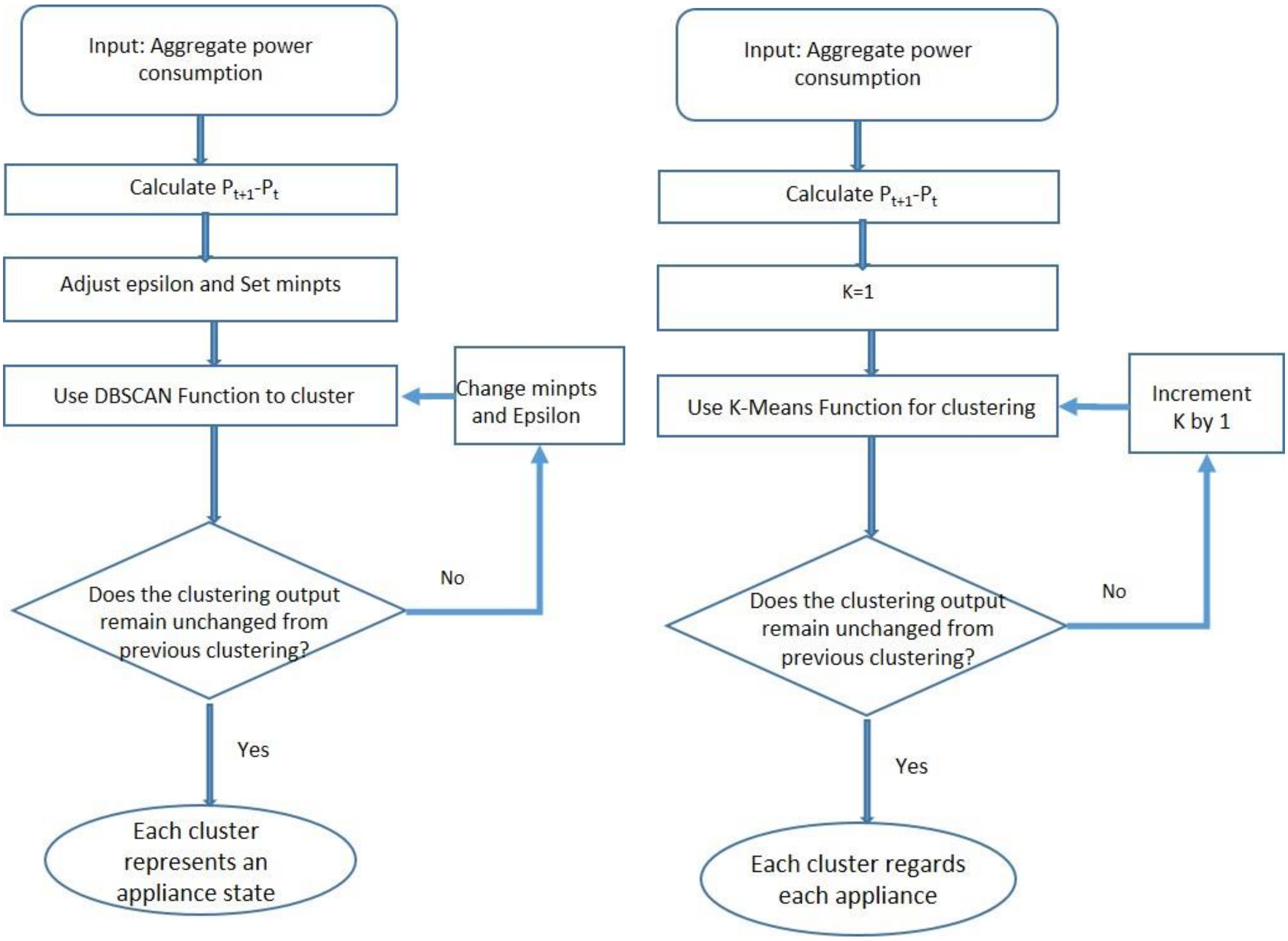

Input: Aggregate power consumption
Calculate P_{t+1}-P_t and P_{t+2}-P_{t+1}
Pair each positive edge to a negative edge
Calculate working duration of appliances based on positive and negative edges.
Used Working duration and power consumption (based on the above calculation) and the rules per specified appliance
Appliance detection
Input: Submetering active power consumption and duration of each working cycle of appliance
Extract active power level of each appliances & Extract working cycle duration of each appliances
Prepared appliance-specific rule for DT branching based on the active power level and work cycle duration

Figure 3.2: Flow chart of proposed supervised algorithms, each showing testing on the left and training on the right for: DT. P_t denotes a power measurement at time instance *t*.

The performance of the proposed algorithms demonstrated by using a dataset containing aggregate and sub-metering active power for 3 datasets: REFIT H17 [52], REDD H1 [51] and a house in Norway (NOR), all for a month's duration. Sampling rate for aggregate power are 10 secs, 1 sec and 1 minute for REFIT, REDD and NOR, respectively. For the supervised algorithms, 80% of the sub-metering data used to train the model and perform on testing of the remaining 20% of labelled aggregate data. Features used were active power and duration of each ON-state. There are multiple states in an oven signature due to the thermostat. There were 28 and 5 oven activations in NOR and REDD datasets, and 17 and 261 kettle activations for NOR and REFIT datasets.

Results are shown in Table 3.1. For training ANN, Levenberg-Marquardt backpropagation with 25 hidden layers to train the kettle model was used, and for the oven model, the conjugate gradient backpropagation with 50 hidden layers provides the best results. With the K-means unsupervised algorithms, the best results were obtained with k=10, indicating 10 distinct clusters. Input is the change in power level, i.e., $\Delta P = P_{t+1} - P_t$. With DBSCAN, using the same ΔP feature for clustering, and best results with Epsilon=10 and Minpts=3 was obtained.


Input: Aggregate power consumption
Calculate P_{t+1}-P_t
Adjust epsilon and Set minpts
Use DBSCAN Function to cluster
Change minpts and Epsilon
Does the clustering output remain unchanged from previous clustering?
No
Yes
Each cluster represents an appliance state
Input: Aggregate power consumption
Calculate P_{t+1}-P_t
K=1
Use K-Means Function for clustering
Increment K by 1
Does the clustering output remain unchanged from previous clustering?
No
Yes
Each cluster regards each appliance

(b) **(a)**

Figure 3.3: Flow chart of proposed unsupervised algorithms: a) K-Means, b) DBSCAN, where Epsilon: A scalar value for Epsilon-neighborhood threshold and Minpts: A scalar value for minimum points in Epsilon-neighborhood that holds the core-point condition. P_t denotes a power measurement at time instance t.

Table 3.1: F1-score accuracy in % for NILM of 2 appliances for 2 supervised and 2 unsupervised ML algorithms

Algorithms		Kettle		Oven	
		REFIT	NOR	NOR	REDD
Supervised	ANN	73	**84**	45	57
	DT	**92**	61	57	**67**
Unsupervised	K-Means	76	51	79	66
	DBSCAN	**92**	46	**85**	60

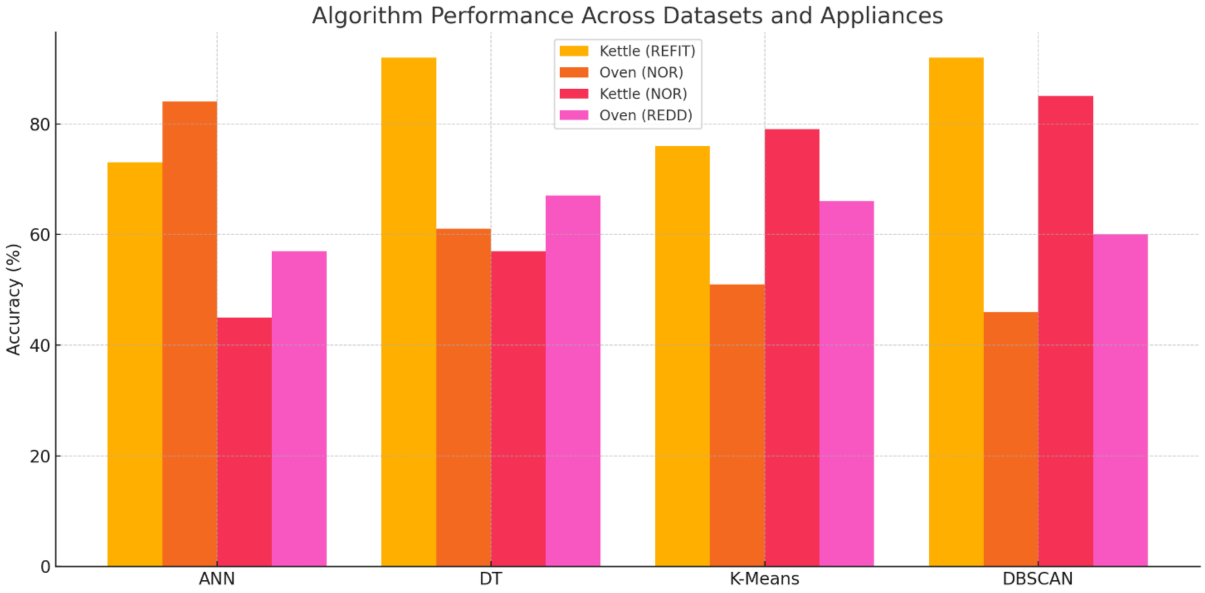


Figure 3.4: F1-score accuracy in % for NILM of 2 appliances for 2 supervised and 2 unsupervised ML algorithms

The best performing disaggregation algorithm for the oven is the unsupervised DBSCAN for 1 min sampled NOR dataset at 85% accuracy. Best results for disaggregation of kettle are observed with supervised ANN for NOR dataset and DT for REFIT datasets. Supervised algorithms, ANN and DT, disaggregate the kettle better than the oven in general since the kettle has a very distinct always-on profile for all activations, unlike the oven whose electrical signature does not have a distinct profile, i.e., the variations among the many activations of the oven for active power and duration vary too much due to opening/closing oven variability. DT performs better with the relatively higher sampling rate datasets (REFIT and REDD) than the NOR dataset. Results obtained are in line with previous studies, with 65% accuracy for the Oven for the REDD dataset obtained with unsupervised GSP in [17] and 94% obtained with DT for kettle in the REFIT dataset in [22]

Here ΔP and Δt used as a feature for training and classification with DT and ANN and for K-means and DBSCAN, only ΔP considered as a feature for classification. Feature selection has a direct impact on classification performance, as illustrated in Table 3.2 and Figure 3.5. The effect of different feature sets on the disaggregation of the fridge-freezer using the Decision Tree (DT) algorithm was evaluated and compared. The best performance is belonging to the situation that ΔP and P considered as a feature.

3.2.3 Feature Selection & Time sequence

Feature selection is a pivotal step in the design of Non-Intrusive Load Monitoring (NILM) systems. Accurate feature selection enhances the system's ability to distinguish between appliances, reduces computational overhead, and ensures model generalizability. Steady-state features are derived from stable power levels and are useful for low-frequency data. These features capture the steady and predictable power consumption behaviour of appliances.

- Real Power (P): The active power consumed by an appliance (in watts). Differentiates between devices based on their typical power consumption [21].
- Reactive Power (Q): Indicates the inductive or capacitive nature of the appliance load. Helps identify appliances with motors or heating elements [97].

- Apparent Power (S): The combined effect of real and reactive power. Useful for distinguishing appliances with similar real power but different reactive power.
- Power Factor: Ratio of real power to apparent power, indicating efficiency. Appliances like fans and air conditioners have characteristic power factors.

Event-based features focus on changes in the power signal over time. These features are particularly effective for capturing appliance usage patterns and transitions.

- On/Off Transitions: Sudden changes in power levels that indicate the activation or deactivation of an appliance. For instance, lights and heaters exhibit sharp transitions.
- Duration of Events: Time an appliance remains in a specific state. Washing machines and refrigerators exhibit characteristic durations for cycles or cooling intervals.
- Transition Magnitudes: Differences in power before and after events. For example, the magnitude of the power drop can identify a specific appliance being turned off.
- Timing Information: Time-of-day or periodic usage patterns (e.g., lights used in the evening or coffee makers in the morning).

Feature selection is a cornerstone of NILM system design, as it directly influences the accuracy, interpretability, and efficiency of the disaggregation process ..In [6], ΔP and Δt were used as features for training and classification with DT and ANN, while only ΔP was considered as a feature for classification with K-Means and DBSCAN. The selection of features significantly affects classification performance. Table 3.2 and figure 3.5 presents the effect of different features on the disaggregation of a fridge-freezer using the DT algorithm, comparing their performance. The best results were achieved when both ΔP and P were used as features. Here, Pt represents the power measurement at time instance t, , $\Delta P=P_{t+1}-P_t$ and Pa is $P_{t+a\,\Delta T}$, where ΔT is the time interval between two consecutive samples.

Table 3.2: Performance for Fridge-freezer disaggregation by DT algorithm with different feature

Data	Sample rate	Function	appliances	PR	RE	FM	Acc	feature
Refit house 2	63-64 sec	DT	Fridge-freezer	0.41	0.95	0.57		Time & ΔP
Refit house 2	63-64 sec	DT	Fridge-freezer	0.53	0.08	0.14		ΔP
Refit house 2	63-64 sec	DT	Fridge-freezer	0.66	0.89	0.76	0.68	ΔP& P
Refit house 2	63-64 sec	DT	Fridge-freezer	0.40	0.81	0.54		Time & ΔP1, ΔP2
Refit house 2	63-64 sec	DT	Fridge-freezer	0.66	0.85	0.74		Time & P
Refit house 2	63-64 sec	DT	Fridge-freezer	0.66	0.78	0.72		P
Refit house 2	63-64 sec	DT	Fridge-freezer	0.66	0.87	0.74	0.69	Time & ΔP1, P
Refit house 2	63-64 sec	DT	Fridge-freezer	0.73	0.71	0.72	0.67	P,P-1,P-2,P-3,P-4,P-5,P-6,P-7,P-8,P-9,p-10,P-11,P-12,p-13,P-14, P-15/Pmax

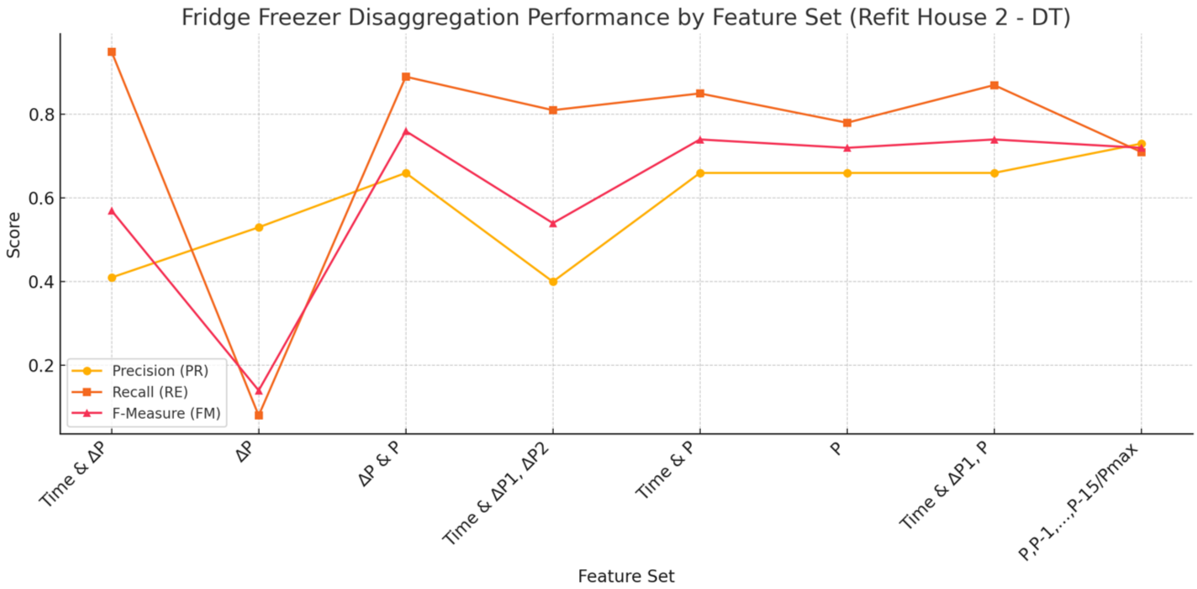


Figure 3.5: Performance for Fridge-Freezer disaggregation by DT algorithm with different feature

Table 3.3: Performance for Toaster disaggregation by DT algorithm with different feature

Data	Sample rate	Function	appliances	PR	RE	FM	Acc	feature
Refit house 2	63-64 sec	DT	toaster	0.45	0.17	0.25	0.45	P
Refit house 2	63-64 sec	DT	toaster	0.36	0.23	0.28	0.36	ΔP1,P
Refit house 2	63-64 sec	DT	toaster	0.92	0.21	0.34	0.58	Time & ΔP1,P
Refit house 2	63-64 sec	DT	toaster	0.46	0.24	0.32	0.44	P1,P2
Refit house 2	63-64 sec	DT	toaster	0.58	0.21	0.31	0.50	P1,P2,P3
Refit house 2	63-64 sec	DT	toaster	0.88	0.28	0.43	0.60	P1,P2,P3,P4
Refit house 2	63-64 sec	DT	toaster	0.72	0.40	0.51	0.58	P1,P2,P3,P4,P5
Refit house 2	63-64 sec	DT	toaster	0.73	0.41	0.53	0.58	P1,P2,P3,P4,P5,P6
Refit house 2	63-64 sec	DT	toaster	0.82	0.42	0.55	0.62	P1,P2,P3,P4,P5,P6,P7,P8

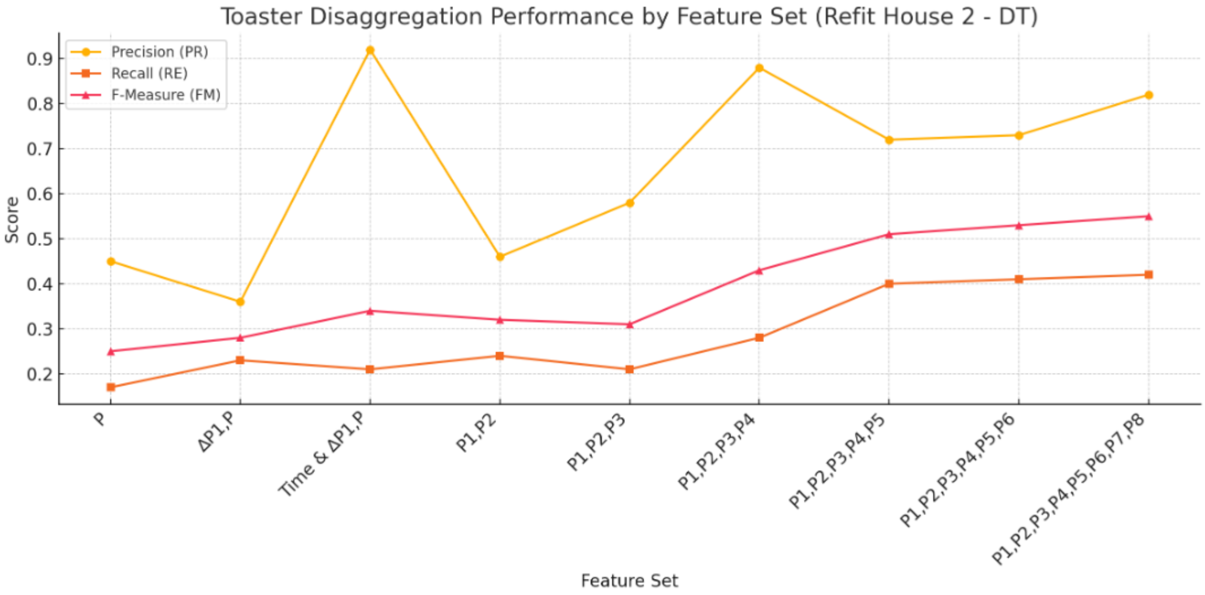


Figure 3.6: Performance for Toaster disaggregation by DT algorithm with different feature

Table 3.3 and figure 3.6 presents the performance of toaster disaggregation using the DT algorithm. It can be observed that when the features include a longer sequence of active power measurements, the results improve. This is because incorporating more sequences of active power provides a clearer and more distinct representation of each appliance's duty cycle, thereby enhancing the overall performance.

Table 3.4: Performance for Kettle disaggregation by DT algorithm with different feature

Data	Sample rate	Function	appliance	PR	RE	FM	Acc	feature
Refit house 2	63-64 sec	DT	Kettle	0.72	0.59	0.65	0.67	Time & ΔP1,P
Refit house 2	63-64 sec	DT	Kettle	0.53	0.72	0.62	0.54	ΔP1,P
Refit house 2	63-64 sec	DT	Kettle	0.76	0.85	0.80	0.73	P,P-1,P-2,P-3,P-4,P-5,P-6,P-7,P-8,P-9
Refit house 2	63-64 sec	DT	Kettle	0.76	0.88	0.82	0.80	P,P-1,P-2,P-3,P-4,P-5,P-6,P-7,P-8,P-9,p-10
Refit house 2	63-64 sec	DT	Kettle	0.81	0.89	**0.85**	**0.83**	P,P-1,P-2,P-3,P-4,P-5,P-6,P-7,P-8,P-9,p-10,P-11
Refit house 2	63-64 sec	DT	Kettle	0.80	0.89	0.84	0.83	P,P-1,P-2,P-3,P-4,P-5,P-6,P-7,P-8,P-9,p-10,P-11,P-12
Refit house 2	63-64 sec	DT	Kettle	0.79	0.90	0.84	0.82	P,P-1,P-2,P-3,P-4,P-5,P-6,P-7,P-8,P-9,p-10,P-11,P-12,p-13
Refit house 2	63-64 sec	DT	Kettle	0.78	0.84	0.82	0.80	P,P-1,P-2,P-3,P-4,P-5,P-6,P-7,P-8,P-9,p-10,P-11,P-12,p-13,P-14

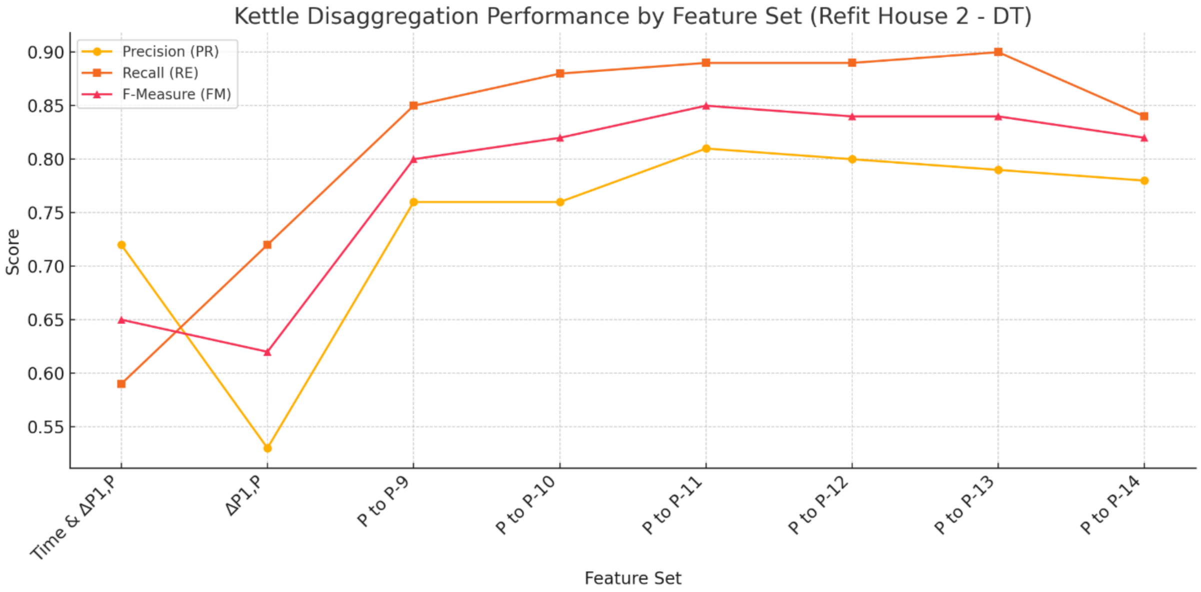


Figure 3.7: Performance for Kettle disaggregation by DT algorithm with different feature

Table 3.4 and figure 3.7 shows the performance of kettle disaggregation using the DT algorithm. It can be observed that when the features include a longer sequence of active power measurements, the performance improves. However, increasing the number of active power samples beyond 11 leads to a decline in disaggregation performance. This indicates that the window size has a significant impact on the effectiveness of the disaggregation.

3.3 SUMMARY

Feature selection plays a pivotal role in the performance of NILM algorithms, directly influencing accuracy, computational complexity, and data storage requirements. While incorporating additional features or extending the length of active power sequences can enhance the distinctiveness of appliance signatures, excessive feature expansion may lead to diminishing returns or even degrade performance, particularly when window sizes become overly large. This underscores the importance of striking a balance between classification accuracy, computational efficiency, and storage demands. Optimally selected features ensure that NILM algorithms remain both effective and resource-efficient, avoiding unnecessary overhead while maintaining high disaggregation performance.

This chapter systematically investigated the role of feature engineering and selection in NILM. A range of candidate features—such as power levels, operating durations, and temporal dynamics—were extracted and evaluated using statistical and machine learning criteria. Results demonstrated that feature choice has a substantial impact on disaggregation accuracy and robustness, especially in scenarios involving overlapping appliance signatures and noisy, low-frequency smart meter data. The key contribution of this chapter lies in providing a principled evaluation of feature extraction and selection strategies, identifying which features most reliably support appliance identification. This addresses a critical gap in the literature, where features have often been applied inconsistently or without rigorous justification, and establishes a foundation for the algorithmic developments explored in the following chapters.

Building on these results, Chapter 4 presents a comparative analysis of supervised and unsupervised NILM algorithms, examining how the identified feature sets affect classification accuracy, clustering quality, and computational efficiency across different algorithmic paradigms.

Chapter 4

3 Load Disaggregation of Electrical Measurements from Smart Meter Data

(Khazaei, M., Stankovic, L., & Stankovic, V. (2020, November). Evaluation of low-complexity supervised and unsupervised NILM methods and pre-processing for detection of multistate white goods. In 5th International Workshop on Non-Intrusive Load Monitoring[30])

Non-intrusive load monitoring (NILM) provides a computational, software-based solution to disaggregate the total energy consumption recorded by a single electricity meter into individual appliance loads. This approach leverages global investments in smart metering infrastructure [6], [7]. Beyond its primary application of delivering meaningful energy feedback, NILM offers significant potential for detecting malfunctioning appliances without the need for submetering [8], [9]. Notably, the BBC reported that malfunctioning appliances, particularly white goods, caused nearly 12,000 fires in Great Britain over three years [10], while the Scottish Fire and Rescue Service attributed 340 fires in 2019 to tumble dryers, washing machines, fridge-freezers, and dishwashers [11].

As highlighted in [8], appliance malfunctions can be identified when the NILM signature deviates substantially from expected energy consumption patterns, with these deviations aligning with known anomaly signatures. Furthermore, [8] emphasizes that appliance-level anomaly detection achieves optimal performance when using high-performing NILM algorithms. This was validated through complex supervised methods such as Combinatorial Optimization (CO) [12] and Factorial Hidden Markov Model (FHMM) [13], implemented via the NILMTK framework [14], as well as through Latent Bayesian Melding [15], Super-state HMM (SSHMM) [16], and unsupervised Graph Signal Processing (GSP) [17]. These algorithms successfully detected anomalies in electrical heater and freezer operations within the REFIT dataset [52].

Motivated by NILM's demonstrated potential to detect appliance malfunctions [8], [9], requirements for effective NILM algorithms shown as below:

(i) the ability to perform near real-time disaggregation to quickly identify faulty appliances,

(ii) low computational complexity to allow deployment on small, on-site devices rather than relying on cloud-based processing, and

(iii) scalability to handle diverse buildings, including those without labelled training data.

To evaluate NILM algorithms with strong disaggregation performance, one method from each of the following categories has been selected: supervised algorithms, which are known for their excellent performance on the specific households they are trained on but have limited transferability to new households [42], and unsupervised algorithms, which offer slightly lower performance but demonstrate greater robustness across datasets without labelled training data. Additionally, NILM algorithms often benefit from improved performance when combined with appropriate pre-processing of input meter data and post-processing of NILM outputs [22] .

4.1 Contributions over the state of the art

Most NILM disaggregation approaches reported in the literature rely on high-frequency measurements or appliance-level data, limiting their applicability to real-world smart-meter environments.

Supervised learning methods often demonstrate high accuracy but suffer from poor transferability across households, while unsupervised methods are typically evaluated on narrow scenarios with limited benchmarking. Prior work within the research group demonstrated the feasibility of low-rate NILM using GSP-based approaches; however, comprehensive comparisons between supervised and unsupervised techniques at smart-meter resolution were limited.

Key gaps in the state of the art include the absence of systematic benchmarking of supervised and unsupervised NILM methods under low-frequency sampling constraints and limited investigation of transferability across households. In addition, existing studies rarely explore structured disaggregation strategies, such as multi-stage approaches, that could mitigate appliance overlap and classification ambiguity in low-resolution data.

This chapter advances the state of the art by providing a detailed benchmarking of supervised and unsupervised NILM algorithms operating on low-frequency smart-meter data. A novel two-stage disaggregation strategy is introduced to address appliance overlap and ambiguity, improving robustness under realistic conditions. Furthermore, the chapter explicitly evaluates cross-household transferability, an aspect often overlooked in prior studies. These contributions correspond directly to the second thesis contribution defined in Chapter 1, which focuses on low-frequency disaggregation performance and generalisability.

4.2 Pre-Processing

Pre-processing is a critical stage in NILM systems, aimed at transforming raw aggregate power measurements into high-quality inputs for effective disaggregation. Its primary goals are to remove noise and outliers, enhance signal smoothness, align sampling rates, and extract informative features that facilitate the separation of appliance-level consumption from the total household load. Several complementary techniques are commonly used in NILM for pre-processing. Noise and outlier removal is typically achieved through median filtering, which smooths the raw power signal while preserving significant events. The choice of window size depends on signal granularity; for instance, a 5-minute window is effective for the REFIT dataset [52], whereas a 3-minute window performs well for REDD [51]. To further enhance signal smoothness while maintaining structural integrity, bilateral graph filtering (GBF) [99] or low-pass filtering [12] may be applied. These filters remove high-frequency noise and ensure piecewise continuity, which is essential for accurate detection of appliance transitions. Next, event detection is carried out to identify power transitions corresponding to appliance ON/OFF operations. Techniques such as derivative-based edge detection [97] or Canny edge detection [98] are often used to locate these transitions. Edge sharpening [99] can then be applied to merge consecutive rising or falling edges caused by gradual switching events, producing clearer and more distinct state changes. In cases where data from different sensors or datasets differ in temporal resolution, resampling ensures consistency. High-frequency data may be down-sampled to match the resolution of NILM algorithms (e.g., from 1 Hz to 15-minute intervals [14]), while low-frequency data may be up-sampled [21] to improve temporal precision for event detection. Normalization and scaling are also essential to improve model convergence. Normalization maps power values to a fixed range (e.g., 0 to 1) [100], while standardization centers the signal around zero mean and unit variance [101]. These steps stabilize training across different households and appliances. Further, segmentation divides the power signal into meaningful windows for analysis. Depending on the application, this can be based on fixed-time intervals (e.g., 1, 15, or 60 minutes [102]) or event-based triggers such as detected ON/OFF transitions [103]. Within each segment, key features are extracted in time, frequency, or temporal domains. These include statistical descriptors (mean, variance [45]), spectral components (dominant frequencies [104]), and periodic usage patterns [70]. To ensure data integrity, data cleaning and synchronization are performed. Missing samples can be reconstructed via interpolation or imputation [106], and timestamps aligned across multiple meters to ensure temporal consistency

[107]. Additional refinements such as power normalization [108] and phase alignment [109] are used in multi-phase systems to balance variations in household load levels. Overall, pre-processing substantially improves NILM performance by delivering clean, smooth, and well-structured input signals. It enhances event detection, increases classification accuracy, and reduces computational noise. In this study, following median filtering, bilateral graph filtering (GBF) [22] is employed to maintain piecewise smoothness in the power signal, and edge sharpening is applied to merge unclear consecutive transitions, ensuring precise event localization in the time-series data.

4.3 Post-Processing

Post-processing is a crucial stage in Non-Intrusive Load Monitoring (NILM), serving to improve the accuracy, interpretability, and robustness of disaggregation results. It encompasses several refinement steps aimed at smoothing noise, validating appliance states, reconciling power, refining operational behaviour, and detecting anomalies. To reduce short-term fluctuations and random measurement noise, temporal smoothing techniques such as moving averages or Savitzky–Golay filters [70] are applied to appliance-level power signals. This helps prevent false ON/OFF detections that may arise from minor variations in the power trace, such as those often observed in lighting systems. In addition, outlier detection and removal procedures [21] are employed to identify and eliminate spurious spikes in the disaggregated signals. These spikes may result from transient events or misclassification and are filtered out when they do not align with the expected behaviour of known appliances. Following the smoothing stage, state validation and correction are implemented to ensure the temporal and logical consistency of detected appliance states. Duration enforcement [110] guarantees that an appliance remains in a given state for a physically reasonable period; for instance, the operating phase of a washing machine cannot last for only a few seconds. Transition rule validation [111] further ensures realistic power transitions by incorporating domain knowledge. For example, a heater is expected to change power levels gradually rather than instantaneously switching from high to off.

Another essential step is power reconciliation, which ensures that the total of the disaggregated appliance powers remains consistent with the aggregate measured power [112]. This balance can be achieved by proportionally rescaling the individual appliance signals or assigning the residual discrepancy to an “unknown” category. Residual power [22] may also be allocated to always-on or standby appliances, thereby accounting for background consumption that cannot be attributed to specific devices.

Refinement of appliance behaviour can also be performed using clustering and statistical analysis techniques [113]. These methods allow the identification of multiple operating modes within a single appliance, such as distinguishing between different power levels of a microwave oven. For cyclic appliances like refrigerators or air conditioners, cycle consistency enforcement [114] ensures that their temporal operation patterns follow the expected duty cycles over time.

The effectiveness of these post-processing steps is typically evaluated using standard performance metrics such as the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and F1-score [115]. Energy summarisation over specific time intervals (daily, weekly, or monthly) [116] provides a higher-level understanding of consumption trends, which can support user feedback and energy-saving recommendations. Furthermore, anomaly detection mechanisms [105] are applied to identify abnormal or inefficient appliance behaviour. An example includes detecting an increase in refrigerator power consumption, which may suggest reduced efficiency or a developing fault.

Overall, post-processing substantially improves NILM outcomes by correcting errors in disaggregation, enhancing interpretability, and mitigating noise and misclassification effects. In the context of this study, the multi-state appliances under consideration have similar operating power levels but differ in their duty cycles. By incorporating expert knowledge of typical duty cycles, rising and falling ΔP events are grouped within the duration of the maximum expected duty cycle, thereby improving event matching and anomaly identification accuracy.

4.4 Algorithms Under Consideration

This study considers several machine learning algorithms that have been widely used for appliance load disaggregation, namely Decision Tree (DT), Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and K-Nearest Neighbour (KNN). Each method offers distinct advantages in terms of complexity, training requirements, and adaptability to limited feature availability.

The Decision Tree (DT) algorithm is a low-complexity supervised approach that can be trained using a relatively small labelled dataset. As demonstrated in [22] and [118], DT-based NILM can effectively use the difference between two consecutive active power measurements (ΔP) as the primary input feature. In the present study, the algorithm is further enhanced by incorporating both active power (P) and ΔP as training features, thereby improving the discrimination capability between appliances with overlapping power characteristics.

The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm is an unsupervised clustering technique that requires only two parameters: the neighbourhood radius (ε) and the minimum number of points (minPts) required to form a dense cluster. Previous work [119] has shown that DBSCAN is suitable for load disaggregation tasks, achieving approximately 81% classification accuracy when applied to the Eco dataset, downsampled to one-minute resolution, for distinguishing between two refrigerator loads. In this study, ΔP is employed as the primary feature for DBSCAN, given its robustness in capturing switching events.

The K-Nearest Neighbour (KNN) algorithm is a supervised classification method that relies on a labelled training dataset. The optimal number of neighbours, K, is determined using a validation subset corresponding to 60% of the labelled data. The algorithm classifies a new observation by examining the K closest examples in the feature space. Prior research [117] demonstrated that KNN performs effectively for disaggregating appliances such as dishwashers and clothes dryers using the AMPds2 dataset, achieving a classification accuracy of 95% when both active and reactive power were available, compared to 73% when only active power was used. Since reactive power data are typically unavailable from commercial smart meters, this study instead employs active power (P) and ΔP as features, maintaining a balance between simplicity and accuracy. The decision to prioritise power and on-duration as the primary features in this study is driven by both practical and methodological considerations. Power is one of the most consistently measured variables across all smart meter types and remains reliable even at low sampling rates, making it suitable for feature extraction under constrained conditions. Similarly, on-duration provides a stable temporal characteristic that captures key behavioural patterns of many household appliances. Alternative features—such as harmonic distortion, waveform shape, or transient signatures—require higher-resolution measurements that exceed the capabilities of low-power smart meters. By focusing on power and on-duration, the proposed method ensures compatibility with real-world datasets while maintaining computational simplicity and robustness.

Overall, these algorithms collectively represent a diverse set of approaches—ranging from supervised to unsupervised methods—providing a comprehensive evaluation framework for NILM performance under different feature availability and appliance characteristics.

4.5 Algorithmic Framework for Disaggregation

One of the main challenges in Non-Intrusive Load Monitoring (NILM) is the high similarity in power consumption patterns among different household appliances, which often leads to overlapping or non-discriminative features. To address this issue, the disaggregation process in this study is implemented in two stages.

In the first stage, appliances with similar power change characteristics (ΔP), such as the dishwasher and washing machine, are initially grouped and treated as a single composite category during algorithmic disaggregation. This grouping reduces confusion between devices that exhibit comparable power signatures and simplifies the classification process.

In the second stage, a finer disaggregation is applied within each composite group to separate individual appliances. This hierarchical approach improves accuracy by progressively refining classification boundaries. The outputs from this stage are subsequently processed through post-processing steps, as detailed in Section 4.2, to correct misclassifications and refine the estimated power profiles.

Given the focus on practical NILM applications using low-resolution smart-meter data—typically active power readings sampled at 1- to 60-second intervals—this study prioritizes low-complexity algorithms that maintain high interpretability and computational efficiency. Accordingly, three core algorithms are investigated:

- Decision Tree (DT) and K-Nearest Neighbour (K-NN) for supervised NILM, which rely on labelled training data to learn appliance-specific patterns;
- Density-Based Spatial Clustering of Applications with Noise (DBSCAN) for unsupervised NILM, which groups events based on feature similarity without requiring prior labels.

In addition, the effects of pre-processing (noise reduction, event detection, and feature extraction) and post-processing (error correction and rule-based refinement) are systematically examined to assess their influence on overall disaggregation performance.

The overall workflow integrating these three algorithms and the two-stage disaggregation approach is illustrated in Figure 4.1.

The algorithms shown in figure 4.1 utilize active power readings from two open-access datasets, both downsampled to a 1-minute resolution to align with standard smart meter data intervals: (1) House 1 of the REDD dataset [51], and (2) Houses 2 and 3 of the REFIT dataset [52]. These specific houses were selected due to their comprehensive submetered appliance-level data, clear labelling, and inclusion of diverse high-energy devices—making them well-suited for evaluating both supervised and unsupervised NILM algorithms under varied operational conditions.

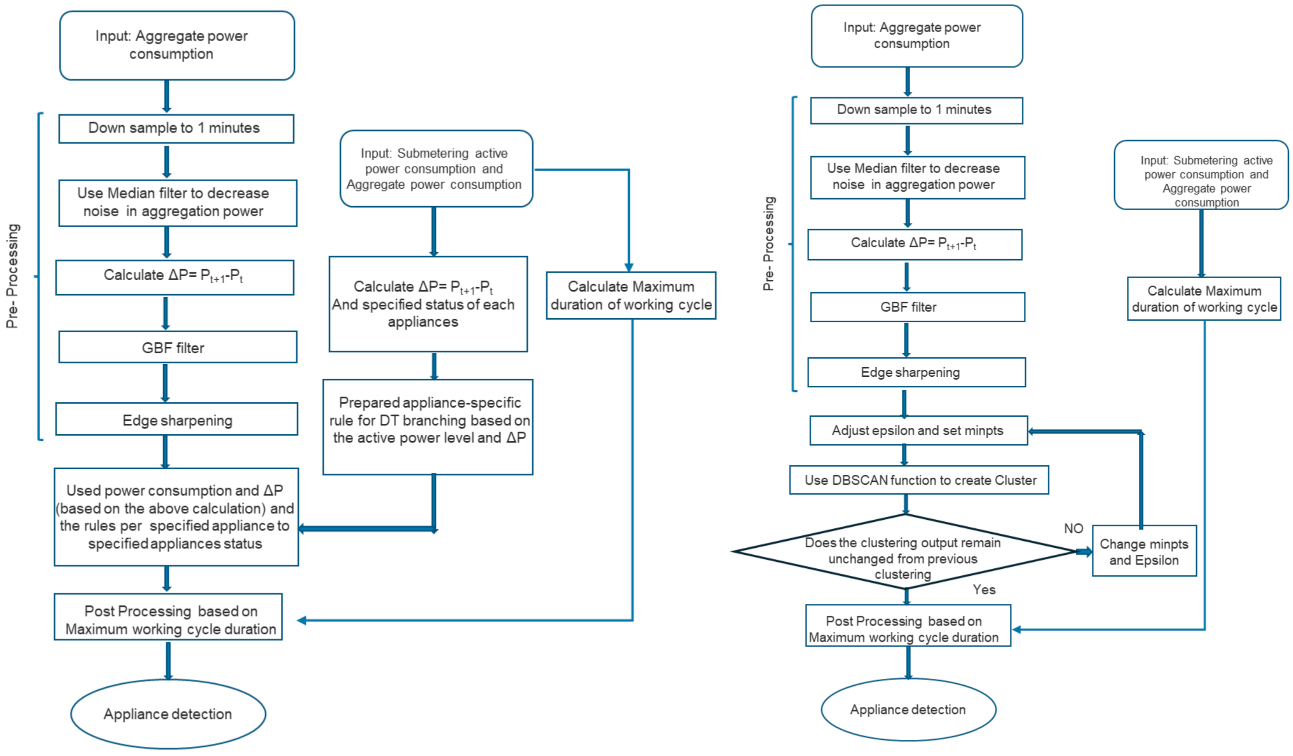
House 1 from the REDD dataset provides detailed short-term monitoring with well-annotated power events, particularly useful for event-based feature extraction and unsupervised learning approaches. In contrast, Houses 2 and 3 of the REFIT dataset offer long-term recordings, capturing realistic appliance usage patterns and enabling robust model training and temporal generalization, particularly advantageous for supervised learning models.

The analysis focuses on appliances that are cyclic in nature and commonly found in residential settings, yet challenging to disaggregate due to overlapping consumption patterns. These include the dishwasher (DW: Tables 4.1–4.7), washer-dryer (WD: Tables 4.8–4.10), washing machine (Tables 4.14–4.16), and tumble dryer (TD: Tables 4.17–4.19).

For all results presented, testing was carried out over a continuous one-month period for each dataset: REDD (30/04/2011 – 30/05/2011) and REFIT (01/10/2014 – 31/10/2014). Evaluation metrics included F-Score, as well as Precision (PR), Recall (RE), and Accuracy (Acc), in line with established NILM evaluation practices [17]. In each table, the best-performing results are highlighted in bold for clarity.

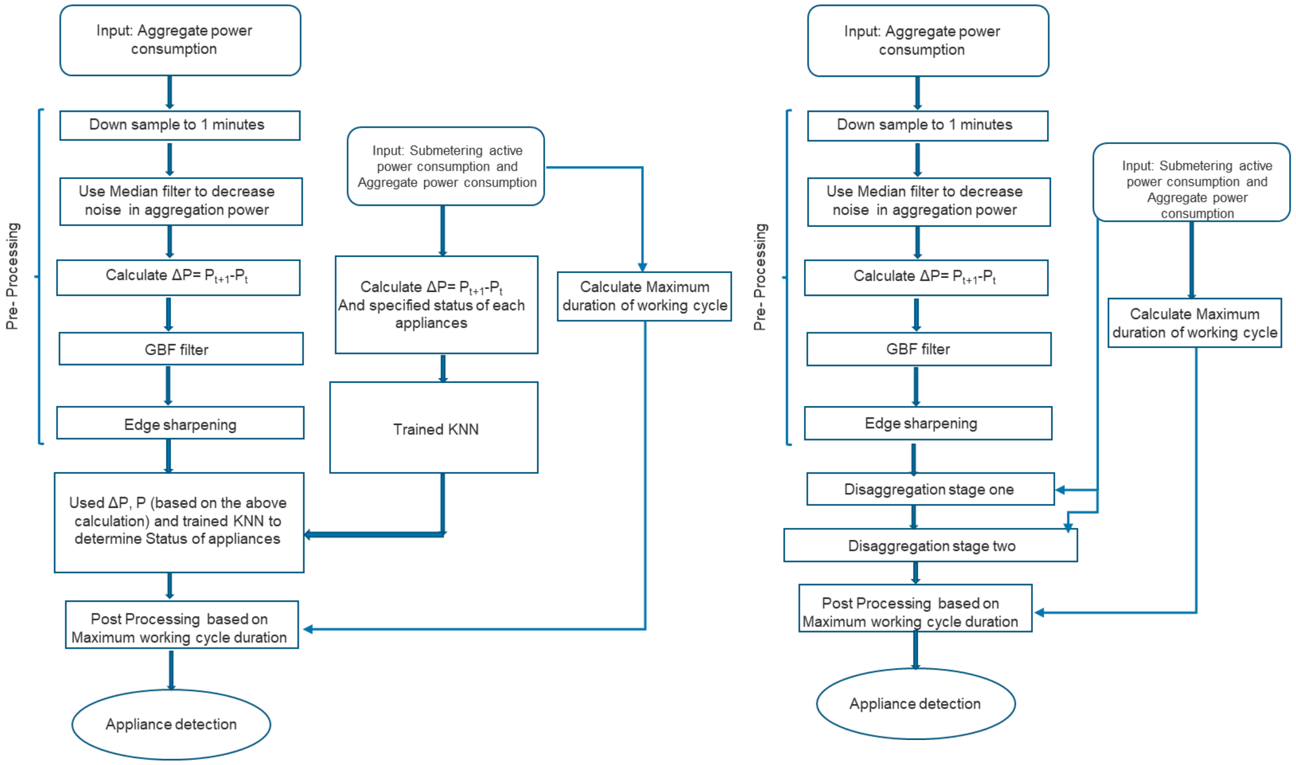
To enhance algorithmic performance, various pre-processing methods were applied prior to NILM execution, including normalization, delta feature computation, and temporal windowing. All algorithms were also followed by a post-processing stage, described in Section 4.2, which was applied uniformly to improve disaggregation accuracy and reduce false detections.

The experimental framework was implemented in MATLAB, employing built-in toolbox implementations for selected machine learning algorithms, including Artificial Neural Networks (ANN), Decision Trees (DT), K-Nearest Neighbours (k-NN), K-means clustering, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). All algorithm parameters were explicitly defined and controlled by the author, while data preprocessing, feature extraction, NILM workflow integration, post-processing, and performance evaluation were implemented consistently across datasets. This ensured methodological transparency, reproducibility, and fair comparison between algorithms.


Input: Aggregate power consumption
Pre- Processing
Down sample to 1 minutes
Use Median filter to decrease noise in aggregation power
Calculate ΔP= P_{t+1}-P_t
GBF filter
Edge sharpening
Input: Submetering active power consumption and Aggregate power consumption
Calculate ΔP= P_{t+1}-P_t And specified status of each appliances
Prepared appliance-specific rule for DT branching based on the active power level and ΔP
Calculate Maximum duration of working cycle
Used power consumption and ΔP (based on the above calculation) and the rules per specified appliance to specified appliances status
Post Processing based on Maximum working cycle duration
Appliance detection
Input: Aggregate power consumption
Pre- Processing
Down sample to 1 minutes
Use Median filter to decrease noise in aggregation power
Calculate ΔP= P_{t+1}-P_t
GBF filter
Edge sharpening
Input: Submetering active power consumption and Aggregate power consumption
Calculate Maximum duration of working cycle
Adjust epsilon and set minpts
Use DBSCAN function to create Cluster
Does the clustering output remain unchanged from previous clustering
NO
Change minpts and Epsilon
Yes
Post Processing based on Maximum working cycle duration
Appliance detection

(a)

(b)


Input: Aggregate power consumption
Down sample to 1 minutes
Use Median filter to decrease noise in aggregation power
Calculate ΔP= P_{t+1}-P_t
GBF filter
Edge sharpening
Pre- Processing
Input: Submetering active power consumption and Aggregate power consumption
Calculate ΔP= P_{t+1}-P_t And specified status of each appliances
Trained KNN
Calculate Maximum duration of working cycle
Used ΔP, P (based on the above calculation) and trained KNN to determine Status of appliances
Post Processing based on Maximum working cycle duration
Appliance detection
Input: Aggregate power consumption
Down sample to 1 minutes
Use Median filter to decrease noise in aggregation power
Calculate ΔP= P_{t+1}-P_t
GBF filter
Edge sharpening
Pre- Processing
Input: Submetering active power consumption and Aggregate power consumption
Calculate Maximum duration of working cycle
Disaggregation stage one
Disaggregation stage two
Post Processing based on Maximum working cycle duration
Appliance detection

(C)

(d)

Figure 4.1 : Flowchart for DT algorithm(a), DBSCAN algorithm (b), KNN algorithm(c) and 2 stage disaggregation (d)

Table 4.1: DW performance for REDD House 1 with DT

	PR	RE	F-Score	Acc
No pre-processing	0.65	**0.84**	**0.73**	0.63
Median filtering	0.59	0.83	0.69	0.61
Edge sharpening	**0.67**	0.72	0.69	0.63
Median filtering + edge sharpening	**0.67**	0.63	0.65	0.61
Median filtering + GBF + Edge sharpening	0.64	0.66	0.64	0.60
Benchmark of pre-processing with DT [15]			0.57	0.58
Benchmark of pre-processing [15] with SGSP			0.63	**0.72**

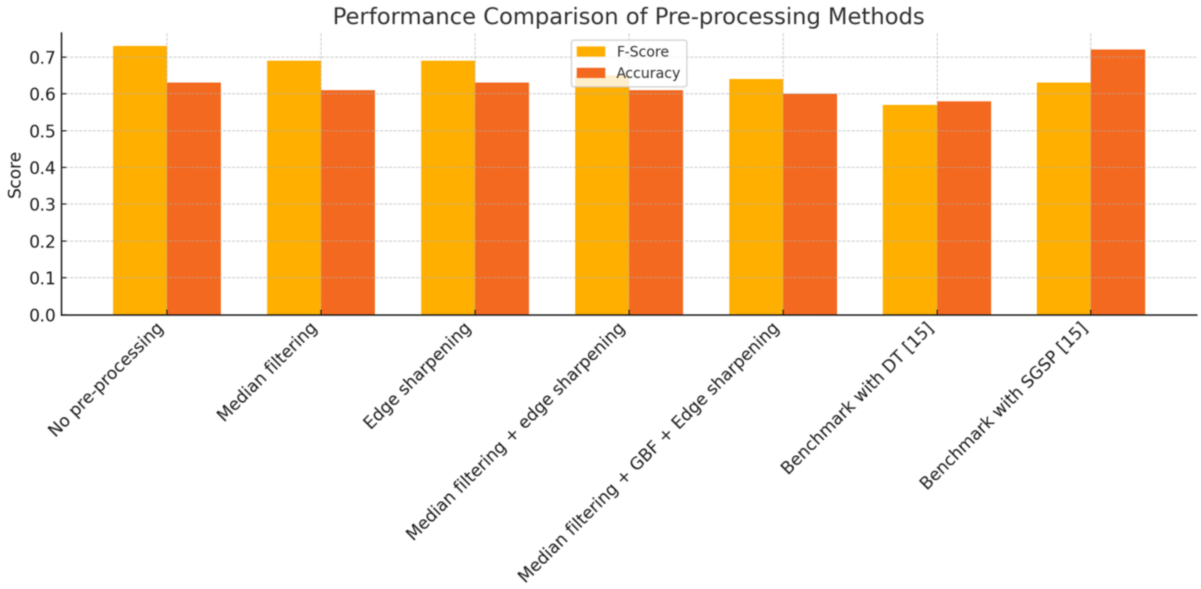


Figure 4.2: DW performance for REDD House 1 with DT

Figure 4.2 and Table 4.1 illustrate how various pre-processing techniques influence F-Score and accuracy in NILM when using a Decision Tree classifier. Among the tested configurations, the no pre-processing case achieved the highest F-Score of 0.73, although its accuracy remained moderate at 0.63. The edge-sharpening approach provided a good balance between the two metrics, maintaining an F-Score of 0.69 while matching the highest accuracy of 0.63. The benchmark SGSP method [15] attained the highest overall accuracy of 0.72, but this came with a moderate F-Score of 0.63. Finally, combining multiple pre-processing steps—such as median filtering, GBF, and edge sharpening—did not yield superior performance compared to the simpler configurations, indicating that additional complexity in the pre-processing pipeline does not necessarily translate to better classification outcomes.

Table 4.2: DW performance for REDD House 1 with KNN

	PR	RE	F-Score	Acc
No pre-processing	0.65	**0.83**	**0.73**	**0.63**
Median filtering	0.65	0.81	0.72	**0.63**
Edge sharpening	**0.70**	0.61	0.66	0.62
Median filtering + edge sharpening	0.67	0.65	0.66	0.62
Median filtering + GBF + Edge sharpening	0.48	0.72	0.57	0.50

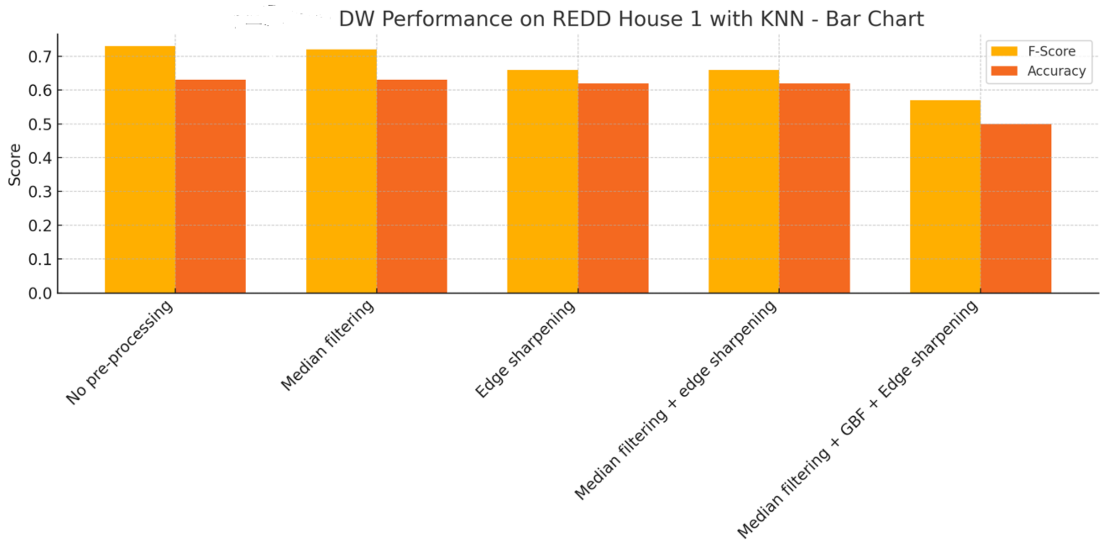


Figure 4.3: DW performance for REDD House 1 with KNN

Figure 4.3 and Table 4.2 present the impact of different pre-processing methods on dishwasher disaggregation using the KNN algorithm applied to REDD House 1. The results show that both the no pre-processing and median filtering configurations achieve the highest performance, with F-Scores and accuracies of approximately 0.73 and 0.63, respectively. In contrast, edge sharpening alone delivers moderate results, while its combination with GBF leads to a notable decline in performance, indicating that excessive or overlapping smoothing techniques can distort signal characteristics and reduce the classifier’s ability to accurately identify appliance events.

Table 4.3: DW performance for REDD House 1 with DBSCAN

	PR	RE	F-Score	Acc
No pre-processing	0.46	**0.65**	**0.54**	0.52
Median filtering	0.47	0.11	0.18	0.51
Edge sharpening	**0.59**	0.14	0.23	0.52
Median filtering + edge sharpening	0.58	0.23	0.33	**0.54**
Median filtering + GBF + Edge sharpening	0.47	**0.65**	0.34	0.51

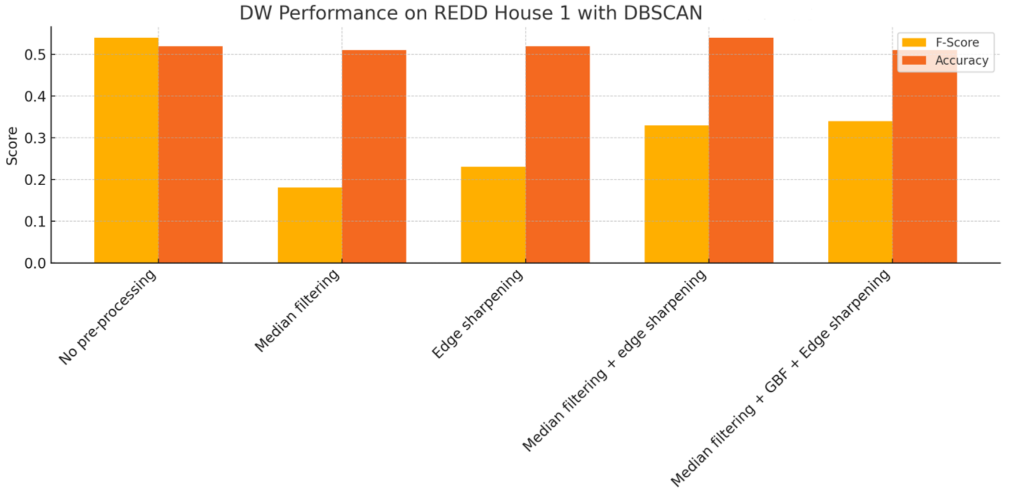


Figure 4.4: DW performance for REDD House 1 with DBSCAN

Table 4.3 and Figure 4.4 illustrate the dishwasher (DW) disaggregation performance using the DBSCAN algorithm on REDD House 1 under various pre-processing configurations. The results indicate that the no pre-processing case provides the best overall balance, achieving an F-Score of 0.54 and an accuracy of 0.52. Applying edge sharpening enhances precision but reduces recall, leading to a lower overall F-Score. When median filtering is combined with edge sharpening, accuracy improves slightly to 0.54, although the F-Score remains relatively low. These findings suggest that DBSCAN, as an unsupervised clustering method, is particularly sensitive to excessive or aggressive pre-processing, and therefore performs best when applied to raw or only lightly filtered signals that preserve the natural structure of appliance-level variations.

Table 4.4: DW performance for REFIT House 2 with DT

	PR	RE	F-Score	Acc
No pre-processing	0.63	**0.90**	0.74	0.67
Median filtering	**0.73**	0.79	0.76	0.73
Edge sharpening	0.53	0.58	0.56	0.53
Median filtering + edge sharpening	0.63	0.65	0.64	0.62
Median filtering + GBF + Edge sharpening	0.53	0.72	0.61	0.54
Median filtering + 2-step DT	0.71	0.82	**0.77**	**0.88**
Benchmark of pre-processing with DT [15]			0.73	0.61
Benchmark of pre-processing with SGSP [15]			0.73	0.67

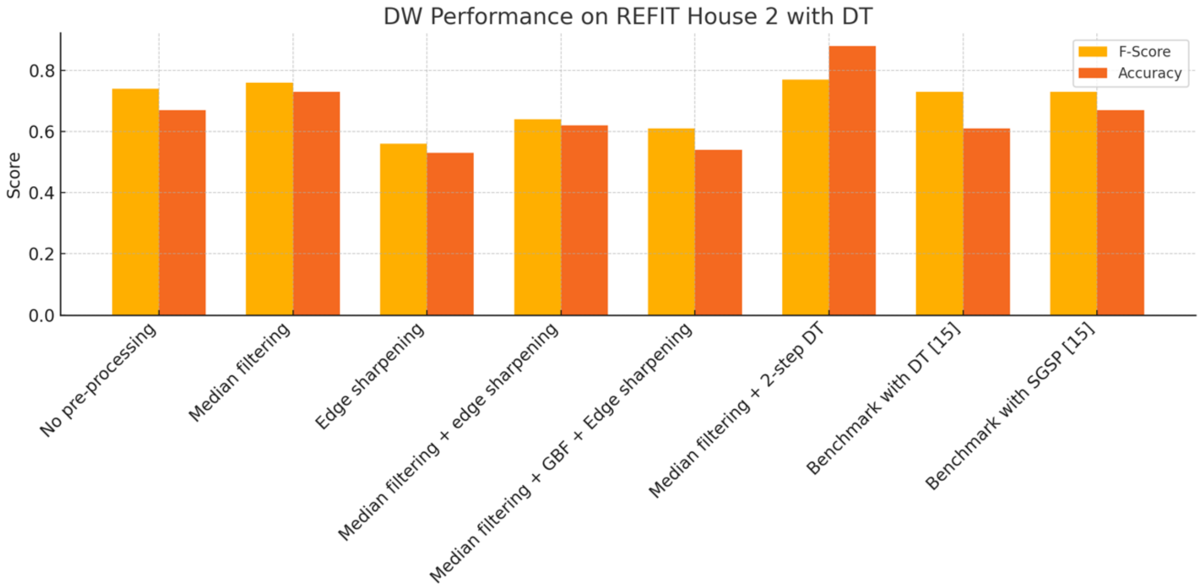


Figure 4.5: DW performance for REFIT House 2 with DT

Figure 4.5 and Table 4.4 present the dishwasher (DW) disaggregation performance using the Decision Tree (DT) algorithm on REFIT House 2 under different pre-processing strategies. The results show that the best performance is obtained with the “Median filtering + 2-step DT” configuration, achieving an F-Score of 0.77 and a notably high accuracy of 0.88. Median filtering alone also performs strongly, with an F-Score of 0.76 and an accuracy of 0.73, demonstrating its effectiveness in enhancing signal quality without excessive smoothing. In contrast, combinations involving edge sharpening or GBF yield lower performance, suggesting that these additional processing steps provide limited benefit for this particular appliance and dataset.

Tables 4.1 and 4.4 demonstrate that the inclusion of **active power (P)** as a feature significantly improves both **classification** and **disaggregation performance** for the **dishwasher** and **washing machine**, when using the **Decision Tree (DT)** classifier. This improvement is evident when compared to baseline approaches presented in [17]. Moreover, the application of the **proposed 2-step DT** yields **additional performance gains** across all evaluation metrics, particularly enhancing disaggregation accuracy. Furthermore, as highlighted in **Tables 4.1, 4.2, and 4.4**, the classification performance for the dishwasher (DW) surpasses that of the **best-performing supervised Graph Signal Processing (GSP)** algorithm with pre-processing reported in [17]. These results confirm the effectiveness of both the **feature selection strategy** (i.e., inclusion of P) and the **2-step classification architecture**, establishing clear advantages over prior state-of-the-art NILM methods.

Table 4.5: DW performance for REFIT House 2 with KNN

	PR	RE	F-Score	Acc
No pre-processing	**0.62**	**0.87**	**0.72**	**0.65**
Median filtering	**0.62**	0.83	0.70	0.64
Edge sharpening	0.44	0.72	0.55	0.42
Median filtering + edge sharpening	0.48	0.77	0.59	0.46
Median filtering + GBF + Edge sharpening	**0.62**	0.72	0.67	0.63

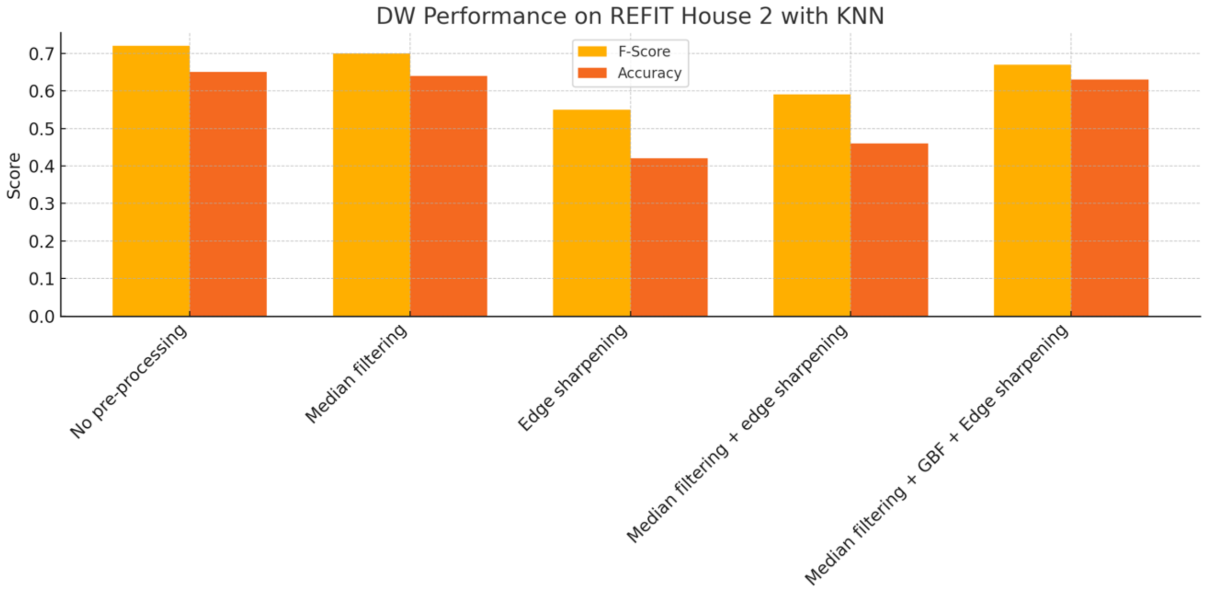


Figure 4.6: DW performance for REFIT House 2 with KNN

Figure 4.6 & table 4.5 visualizing the **DW (Dishwasher) performance on REFIT House 2 using the KNN algorithm** under various pre-processing methods.

The results clearly demonstrate the performance trade-offs across different signal treatment strategies. When no pre-processing is applied, the system achieves the highest overall performance, with an F-Score of 0.72 and an accuracy of 0.65, indicating that the raw signal retains sufficient discriminatory information for effective classification. In contrast, the combination of median filtering, GBF, and edge sharpening yields a slight improvement in accuracy (0.63) but results in a lower F-Score, suggesting that excessive smoothing can weaken class separability. Furthermore, applying edge sharpening alone leads to a notable decline in both accuracy and F-Score, confirming that overly complex or aggressive pre-processing steps may distort important signal characteristics and reduce the effectiveness of KNN-based classification.

Table 4.6: DW performance for REFIT House 2 with DBSCAN

	PR	RE	F-Score	Acc
No pre-processing	0.44	0.63	0.52	0.42
Median filtering	0.48	**0.92**	**0.63**	0.45
Edge sharpening	0.34	0.29	0.31	0.37
Median filtering + edge sharpening	**0.51**	0.47	0.49	**0.51**
Median filtering + GBF + Edge sharpening	0.47	0.39	0.43	0.48

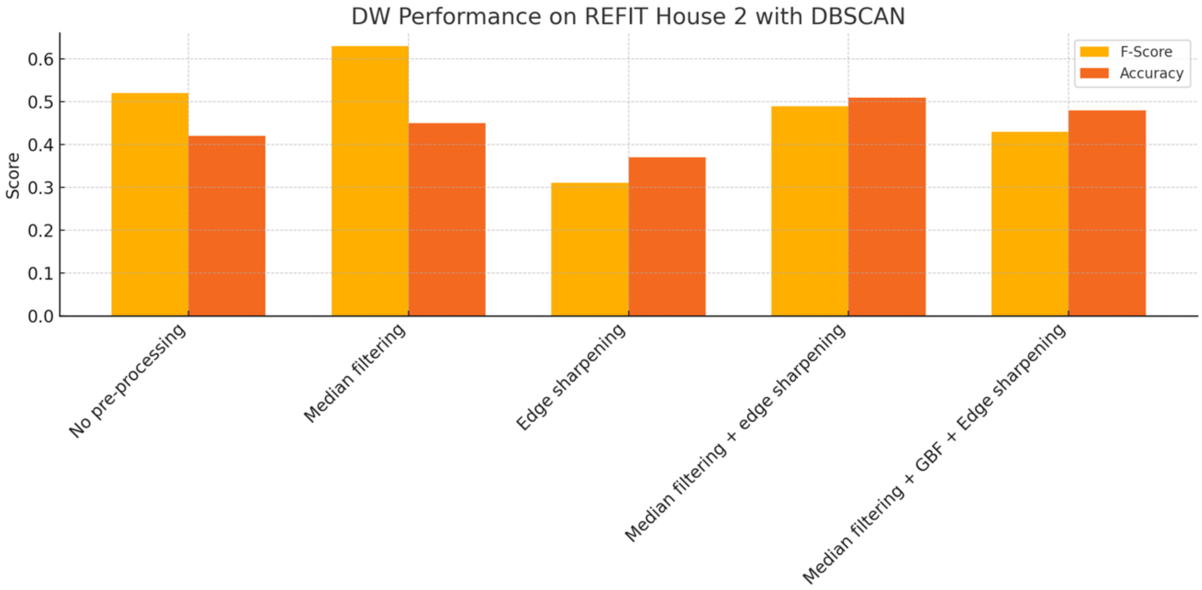


Figure 4.7: DW performance for REFIT House 2 with DBSCAN

Figure 4.7 & table 4.6 showing **dishwasher (DW) disaggregation performance on REFIT House 2 using the DBSCAN algorithm** with different pre-processing methods.

Key Takeaways:

- **Median filtering** offers the best performance (F-Score: 0.63), thanks to its ability to smooth noise without distorting the event structure. Despite a **high recall (0.92)**, the **precision (0.48)** indicates some false positives.

- **No pre-processing** yields moderate performance (F-Score: 0.52, Accuracy: 0.42), showing that DBSCAN can operate without filtering but struggles to separate events clearly.

- **Edge sharpening** significantly degrades both F-Score (0.31) and Accuracy (0.37), likely due to artificial fluctuations being misclassified as appliance events.

- **Combined filtering methods** (“Median + GBF + Edge”) lead to inconsistent gains, suggesting that **unsupervised clustering like DBSCAN is highly sensitive to input signal structure**.

Table 4.7: WD performance for REDD House 1 with DT

	PR	RE	F-Score	Acc
No pre-processing	**0.94**	**0.98**	**0.96**	**0.96**
Median filtering	0.81	0.84	0.83	0.82
Edge sharpening	0.72	0.67	0.69	0.70
Median filtering + edge sharpening	0.80	0.88	0.84	0.83
Median filtering + GBF + Edge sharpening	0.68	0.94	0.79	0.74
Benchmark of DT [5] (no pre-processing)			0.88	

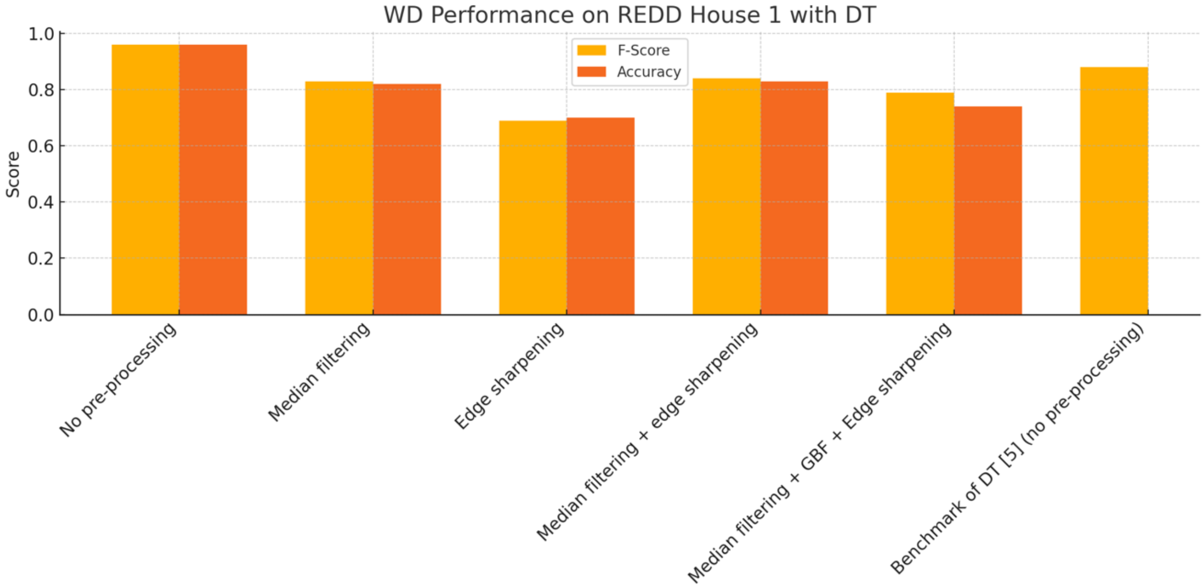


Figure 4.8: WD performance for REDD House 1 with DT

Figure 4.8 and Table 4.7 display the washer-dryer (WD) disaggregation performance on REDD House 1 using the Decision Tree (DT) algorithm under different pre-processing configurations. The results indicate that the no pre-processing approach achieves the highest performance, with an F-Score and accuracy both reaching 0.96, demonstrating that the DT algorithm can effectively classify WD events directly from raw data. Among the pre-processed variants, the combination of median filtering and edge sharpening provides the best results, achieving an F-Score of 0.84 and an accuracy of 0.83, indicating a modest improvement from selective smoothing and edge enhancement. In contrast, edge sharpening alone yields the weakest performance, with an F-Score of 0.69 and an accuracy of 0.70, likely due to excessive noise amplification that obscures clear event transitions.

Table 4.8: WD performance for REDD House 1 with KNN

	PR	RE	F-Score	Acc
No pre-processing	**0.91**	**0.98**	**0.94**	**0.94**
Median filtering	0.23	0.61	0.33	0.22
Edge sharpening	0.23	0.61	0.33	0.22
Median filtering + edge sharpening	0.20	0.51	0.29	0.25
Median filtering + GBF + Edge sharpening	0.47	0.49	0.48	0.47

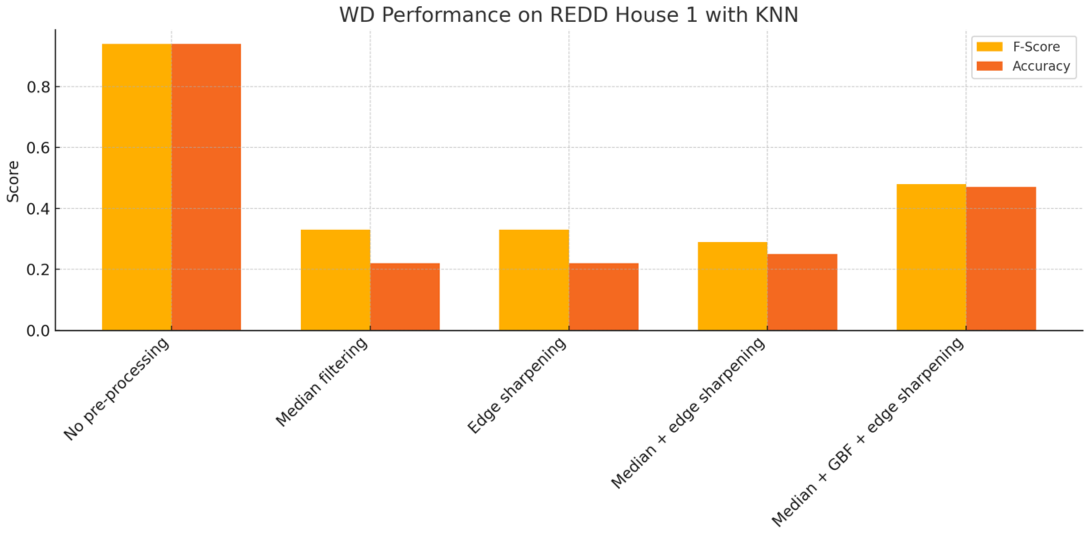


Figure 4.9: WD performance for REDD House 1 with KNN

Figure 4.9 and Table 4.8 compare the F-Score and accuracy for washer-dryer (WD) disaggregation performance on REDD House 1 using the KNN algorithm under different pre-processing techniques. The results reinforce the conclusions drawn in Table 4.8, showing that the raw data (no pre-processing) condition yields the best classification performance overall. In contrast, most filtering methods substantially degrade the effectiveness of KNN, with F-Scores dropping below 0.35, indicating that excessive smoothing can obscure meaningful signal variations critical for distance-based classification. Among the processed configurations, the Median + GBF + Edge sharpening combination produces the best relative performance; however, its results still fall significantly short of those obtained using unprocessed data, highlighting the sensitivity of KNN to pre-processing distortions.

Table 4.9: WD performance for REDD House 1 with DBSCAN

	PR	RE	F-Score	Acc
No pre-processing	**0.63**	**0.71**	**0.67**	**0.64**
Median filtering	0.06	0.02	0.03	0.32
Edge sharpening	0.54	0.22	0.31	0.51
Median filtering + edge sharpening	0.44	0.08	0.14	0.49
Median filtering + GBF + Edge sharpening	0.54	0.10	0.16	0.51

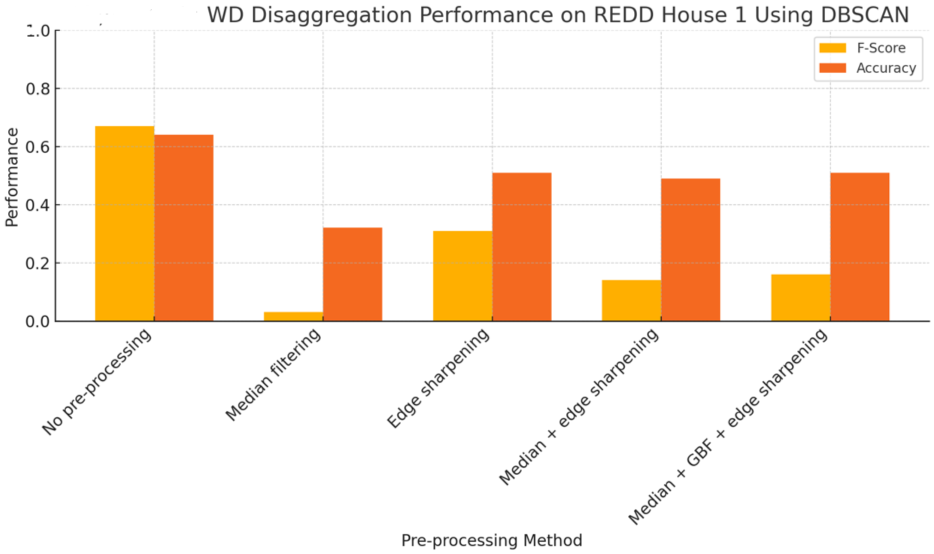


Figure 4..10 : WD performance for REDD House 1 with DBSCAN

Figure 4.10 and Table 4.9 illustrate the washer-dryer (WD) disaggregation performance on REDD House 1 using the DBSCAN algorithm, comparing F-Score and accuracy across various pre-processing strategies. The results show that the no pre-processing configuration achieves the highest F-Score (0.67) and accuracy (0.64), indicating that unsupervised DBSCAN performs best when the raw signal structure is preserved. In contrast, the application of median filtering and other combined filtering techniques leads to a substantial reduction in performance, with F-Scores dropping to as low as 0.03. Although methods such as edge sharpening and Median + GBF + Edge sharpening provide slight improvements compared to heavily filtered signals, their performance remains considerably lower than the unprocessed baseline. These findings underscore DBSCAN's sensitivity to excessive smoothing and highlight the importance of retaining natural signal variability for effective clustering.

Table 4.10: WM performance for REFIT House 2 with DT

	PR	RE	F-Score	Acc
No pre-processing	0.18	0.44	0.26	0.18
Median filtering	0.27	0.54	0.36	0.07
Edge sharpening	0.08	0.11	0.09	0.02
Median filtering + edge sharpening	0.32	0.39	0.35	0.31
Median filtering + GBF + Edge sharpening	0.29	0.30	0.29	0.18
Median filtering+2-step DT	**0.48**	**0.56**	**0.52**	**0.52**
Benchmark of DT [5] (no pre-processing)			0.36	

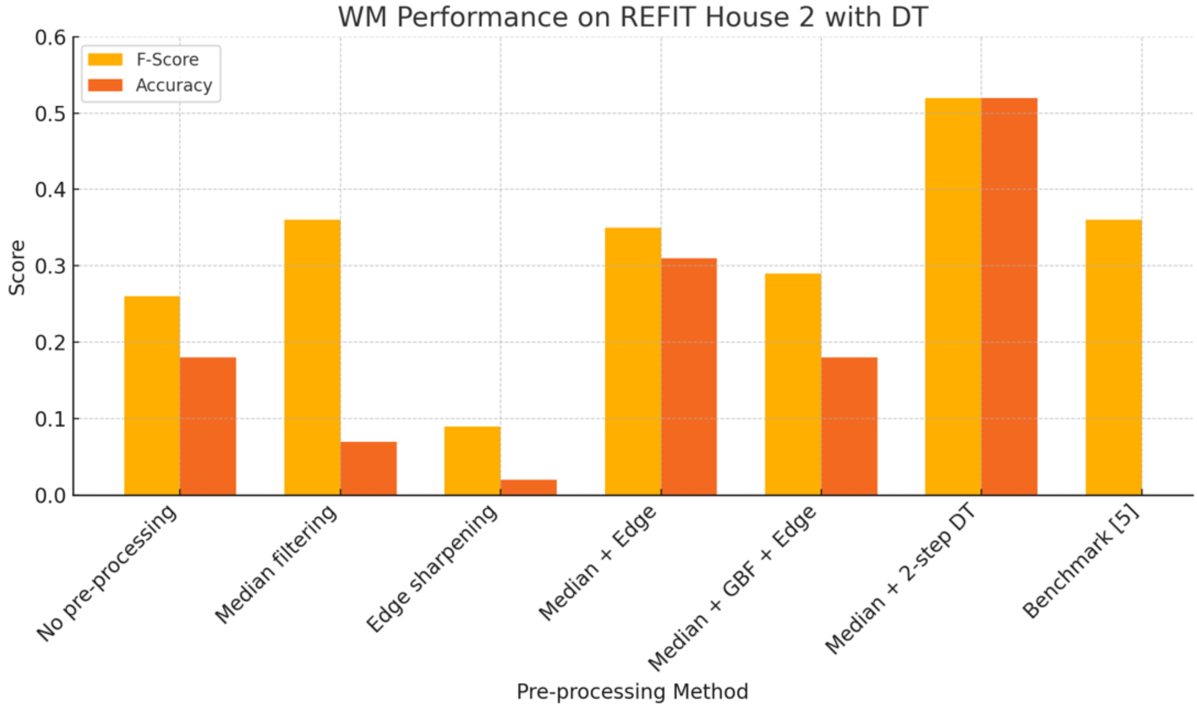


Figure 4.11: WM performance for REFIT House 2 with DT

Figure 4.11 and Table 4.10 present a visual comparison of the F-Score and accuracy for washing machine (WM) disaggregation performance on REFIT House 2 using the Decision Tree (DT) algorithm under different pre-processing configurations. The results indicate that the Median filtering + 2-step DT approach delivers the best classification performance, achieving both an F-Score and accuracy of 0.52, demonstrating its effectiveness in handling appliances with complex and variable load profiles such as washing machines. In contrast, the no pre-processing configuration results in limited performance, with an F-Score of 0.26, suggesting that raw data lacks sufficient clarity for accurate classification. Most basic filtering techniques, particularly edge sharpening, perform poorly—with F-Scores as low as 0.09—as they tend to obscure important appliance transition events. Although combinations such as Median + Edge and Median + GBF + Edge sharpening yield moderate improvements, their results still fall short of the structured and consistent gains achieved through the 2-step DT pipeline, highlighting the value of layered processing in improving NILM performance.

Table 4.11: WM performance for REFIT House 2 with KNN

	PR	RE	F-Score	Acc
No pre-processing	0.15	0.16	0.16	0.20
Median filtering	0.20	**0.35**	0.25	0.07
Edge sharpening	0.09	0.17	0.12	0.11
Median filtering + edge sharpening	0.26	0.40	0.32	0.20
Median filtering + GBF + Edge sharpening	**0.27**	0.33	0.30	**0.27**

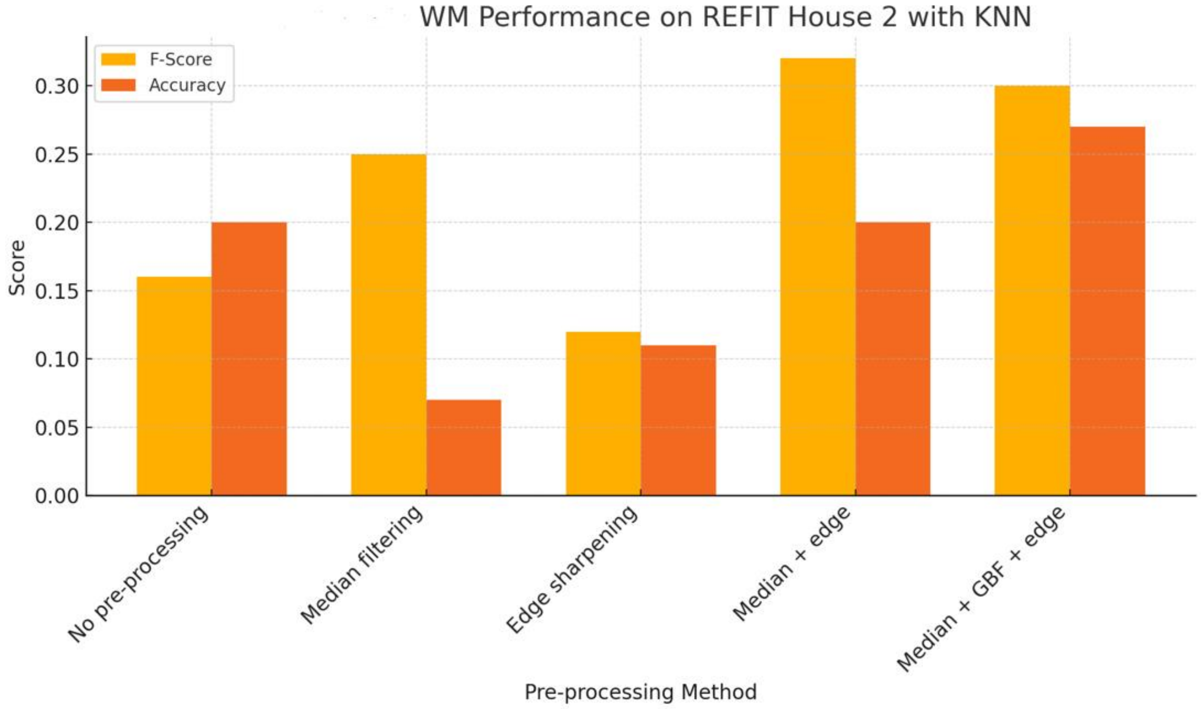


Figure 4.12: WM performance for REFIT House 2 with KNN

Figure 4.12 and Table 4.11 present the F-Score and accuracy for washing machine (WM) disaggregation performance on REFIT House 2 using the KNN algorithm under different pre-processing methods. The results confirm the trends observed in the corresponding performance table. The no pre-processing configuration produces the lowest performance, with an F-Score of 0.16 and an accuracy of 0.20, indicating that KNN struggles to classify washing machine loads effectively when using raw data. The combination of Median filtering and Edge sharpening achieves the highest F-Score of 0.32, although accuracy remains modest at 0.20. Adding GBF to this configuration—resulting in Median + GBF + Edge sharpening—slightly improves accuracy to 0.27, but the F-Score remains limited at 0.30, suggesting that while pre-processing can enhance certain aspects of performance, KNN's ability to distinguish complex appliance cycles such as washing machines remains constrained.

Table 4.12: WM performance for REFIT House 2, DBSCAN

	PR	RE	F-Score	Acc
No pre-processing	**0.12**	**0.76**	**0.21**	0.04
Median filtering	0.11	0.56	0.19	0.07
Edge sharpening	0.07	0.42	0.13	0.04
Median filtering + edge sharpening	0.10	0.57	0.18	**0.24**
Median filtering + GBF + Edge sharpening	**0.12**	0.52	0.16	**0.24**

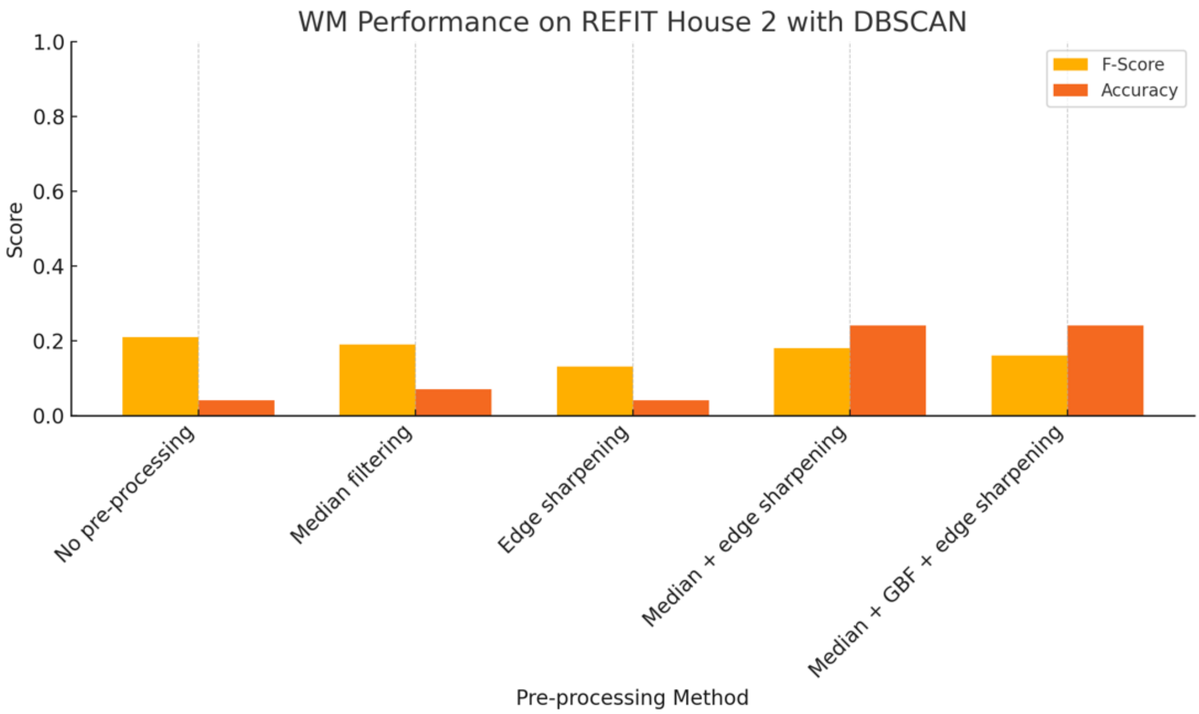


Figure 4.13: WM performance for REFIT House 2, DBSCAN

Figure 4.13 and Table 4.12 illustrate the F-Score and accuracy for washing machine (WM) disaggregation performance on REFIT House 2 using the DBSCAN algorithm under various pre-processing configurations. The results show that the no pre-processing case produces a low F-Score of 0.21 and an accuracy of 0.04, indicating that DBSCAN struggles to reliably identify washing machine activity when applied directly to raw data. Both Median filtering and Edge sharpening individually reduce performance, yielding F-Scores of 0.19 or lower. However, combining multiple filtering methods, such as Median + Edge sharpening or Median + GBF + Edge sharpening, leads to a slight improvement in accuracy, reaching up to 0.24, although F-Scores remain below 0.20. These findings suggest that while limited pre-processing can marginally enhance clustering stability, the unsupervised DBSCAN algorithm remains less effective for complex, variable-load appliances like washing machines.

Table 4.13: TD performance for REFIT House 3 with DT

	PR	RE	F-Score	Acc
No pre-processing	**0.34**	**0.78**	**0.47**	0.10
Median filtering	0.32	0.52	0.40	**0.21**
Edge sharpening	0.23	0.30	0.26	0.14
Median filtering + edge sharpening	0.23	0.43	0.30	0.01
Median filtering + GBF + Edge sharpening	0.25	0.44	0.32	0.06

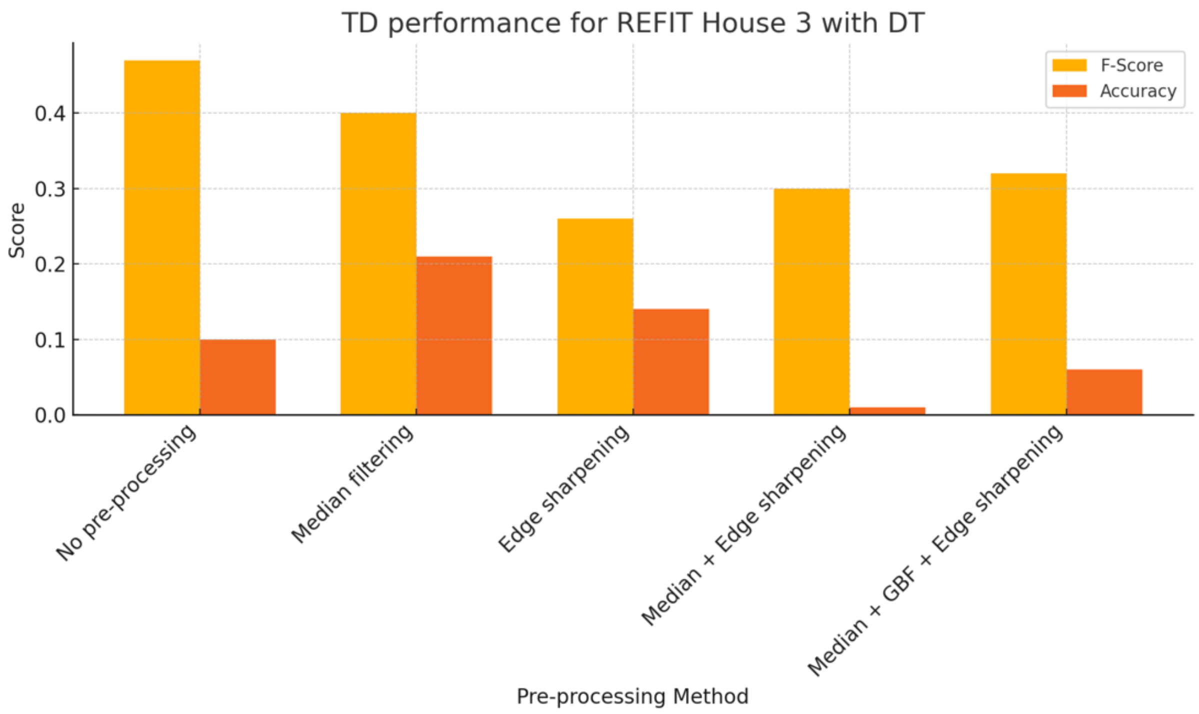


Figure 4.14: TD performance for REFIT House 3 with DT

Figure 4.14 and Table 4.15 illustrate the F-Score and accuracy for tumble dryer (TD) disaggregation performance on REFIT House 3 using the Decision Tree (DT) algorithm under different pre-processing configurations. The results show that the no pre-processing approach yields the highest F-Score of 0.47, although accuracy remains low at 0.10, which may be attributed to class imbalance or limited appliance activity during the evaluation period. Applying Median filtering improves accuracy to 0.21 but slightly decreases the F-Score, indicating a modest trade-off between precision and recall. In contrast, Edge sharpening and its combined variants tend to lower both F-Score and accuracy, highlighting their limited effectiveness for tumble dryer disaggregation. Overall, these findings suggest that pre-processing does not consistently enhance performance for the TD appliance in this scenario, and simpler, minimally processed inputs may be more suitable for Decision Tree-based classification.

Table 4.14: TD performance for REFIT House 3 with KNN

	PR	RE	F-Score	Acc
No pre-processing	**0.34**	**0.65**	**0.44**	**0.17**
Median filtering	0.30	0.60	0.40	0.10
Edge sharpening	0.18	0.32	0.23	0.04
Median filtering + edge sharpening	0.29	0.49	0.36	0.13
Median filtering + GBF + Edge sharpening	0.27	0.47	0.34	0.08

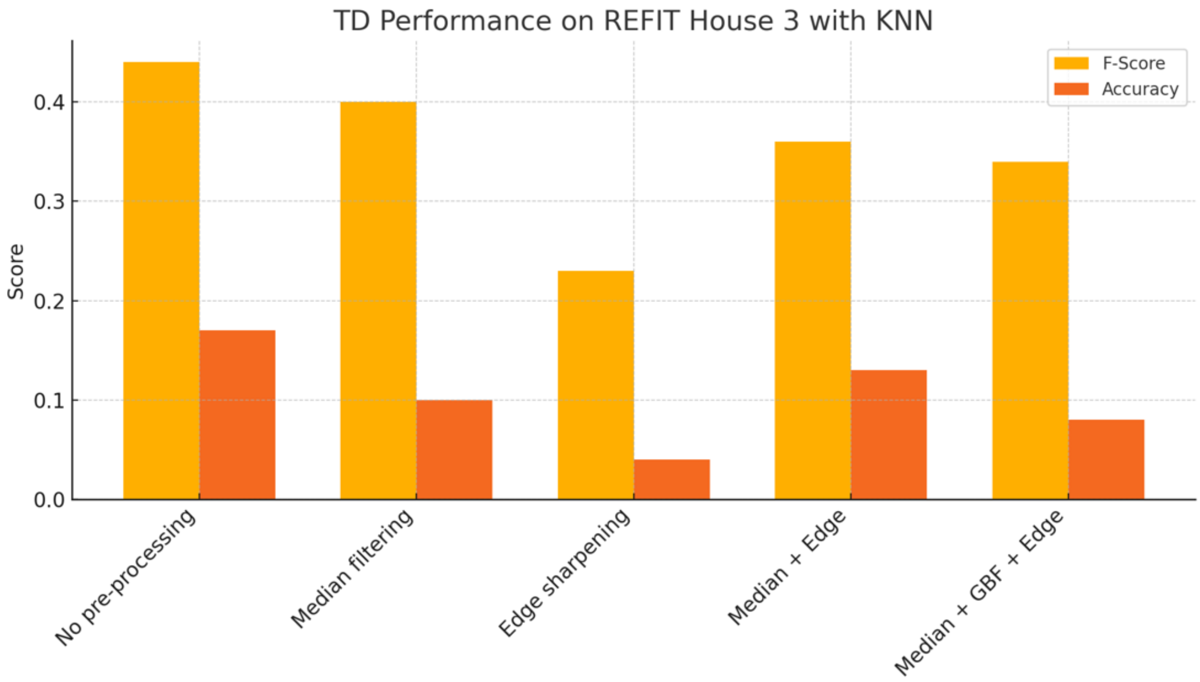


Figure 4.15: TD performance for REFIT House 3 with KNN

Figure 4.15 and Table 4.14 compare the F-Score and accuracy for tumble dryer (TD) disaggregation performance on REFIT House 3 using the KNN algorithm under different pre-processing techniques. The results, consistent with the trends summarized in Table 4.16, show that the no pre-processing configuration produces the best overall performance, achieving an F-Score of 0.44 and an accuracy of 0.17, indicating that KNN performs more effectively when applied directly to raw input data. In contrast, applying Edge sharpening alone significantly reduces classification accuracy and F-Score, with the latter dropping to 0.23, suggesting that this technique amplifies noise and obscures appliance transitions. The combined pre-processing methods, such as Median + GBF + Edge sharpening, slightly improve performance to an F-Score of 0.34, yet still fall short of the results achieved with unprocessed data, reinforcing the conclusion that KNN is most effective when operating on minimally processed signals.

Table 4.15: TD performance for REFIT House 3 with DBSCAN

	PR	RE	F-Score	Acc
No pre-processing	0.16	**0.36**	**0.22**	0.25
Median filtering	0.16	0.01	0.03	0.47
Edge sharpening	0.05	0.01	0.01	**0.48**
Median filtering + edge sharpening	0.25	0.03	0.05	0.47
Median filtering + GBF + Edge sharpening	0.12	0.01	0.02	0.47

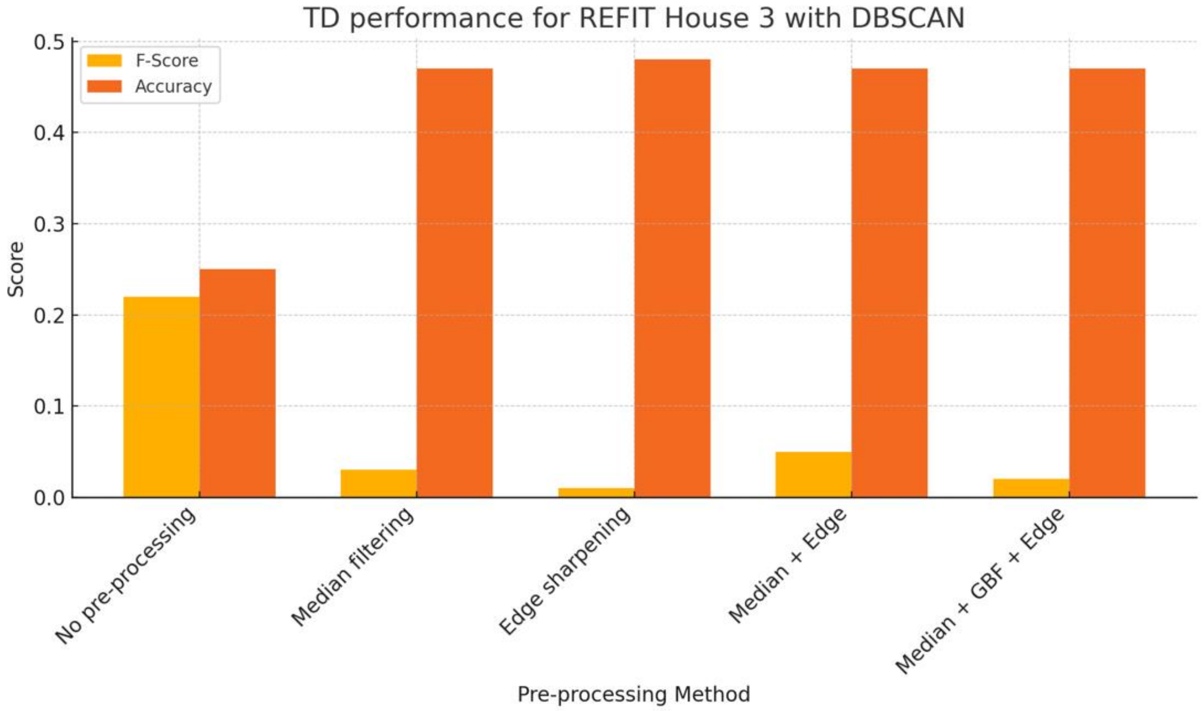


Figure 4.16: TD performance for REFIT House 3 with DBSCAN

Figure 4.16 and Table 4.15 illustrate the tumble dryer (TD) disaggregation performance on REFIT House 3 using the DBSCAN algorithm, comparing F-Score and accuracy across various pre-processing methods. As shown in Table 4.17, the best F-Score of 0.22 is obtained without any pre-processing, although the corresponding accuracy remains relatively low at 0.25. When Median filtering or Edge sharpening is applied, the F-Score drops sharply—approaching zero—indicating a substantial degradation in clustering performance. Interestingly, despite these low F-Scores, the reported accuracies for most pre-processed configurations appear relatively high (around 0.47–0.48), likely reflecting class imbalance, where the algorithm disproportionately predicts majority classes. Overall, these results emphasize the limitations of DBSCAN in accurately distinguishing tumble dryer activity, particularly when aggressive or excessive pre-processing distorts the underlying signal structure.

Tables 4.4 and 4.10 show an improvement of 16% in classification performance of the washing machine and 15% improvement in disaggregation performance of the dishwasher with the proposed two-step DT disaggregation.

Table 4.7 also shows improvement in classification accuracy of WD due to inclusion of P as additional feature compared to the benchmark [7], where DT was used without pre-processing.

As expected, supervised DT and KNN perform better than DBSCAN for all considered appliances. The benefit of pre-processing, especially for improving disaggregation accuracy, is observed clearly where performance is poor, as observed for the unsupervised DBSCAN algorithm (Tables 4.6, 4.10, 4.15) and for the challenging washing machine.

Pre-processing is not beneficial for REDD DW (Tables 4.1-4.4) and WD (Tables 4.7-4.8) at 1-min sampling resolution since the dataset is relatively less noisy than the REFIT dataset. In fact, it is detrimental because it removes some important edges. However, for the noisier (due to additional unknown appliances) REFIT houses and challenging washing machine, median filtering only is sufficient to improve classification accuracy whilst edge sharpening in addition to median filtering, helps improve the disaggregation accuracy, as observed in Tables 10-12. Tumble dryer results from

REFIT House 3 had good recall results with DT and KNN, comparable with other appliances, as we were able to pick out most instances of the appliance running but some post-processing may be needed to reduce false positives. There were no benchmarks for comparison. Summary table comparing the performance of different pre-processing filter selections shown in appendix 7.1.

4.6 Transfer Learning

For checking performance of any application output of the algorithm has to check with the actual result (ground truth), there are some different data set exist those produce by organizations and institutes supporting research like REFIT, REDD, AMPds, iAWE, GreenD, UK-DALE, COMBED, WikiEnergy, SustData, detail of each dataset is shown in table 4.16

Table 4.16 : Detail of each data set

Contribution	Dataset	Location	Duration per house	Number of houses	Appliance sample resolution	Aggregate sample resolution
[51]	REDD	USA	3-19 days	6	3 sec	1 sec & 15 kHz
[120]	BLUED	USA	8 days	1	state transition label	12 kHz
[121]	UMass Smart	USA	3 months	3	1 sec	1 sec
[122]	Tracebase	Germany	N/A	15	1-10 sec	N/A
[123]	Pecan Street Sample	USA	7 days	10	1 min	1 min
[124]	HES	UK	1 or 12 months	251	2 or 10 min	2 or 10 min
[125]	AMPds	Canada	1 year	1	1 min	1 min
[14]	iAWE	IND	73 days	1	1 or 6 sec	1 sec
[21]	UK-DALE	UK	3-17 months	4	6 sec	1-6 sec & 16 kHz
[126]	GreenD	Austria/Italy	1 year	9	1 sec	1 sec
[127]	COMBED	India	18 months	8	30 sec	30 sec
[128]	ECO	Switzerland	8 months	6	1 sec	1 sec
[129]	BERDS	USA	1 year	N/A	20 sec	20 sec
[130]	SustData	Portugal	5 years	50	50 Hz	50 Hz
[52]	REFIT	UK	2 years	20	7-8 sec	7-8 sec

Appliances with similar functions from different manufacturers often exhibit variations in energy consumption due to differences in ON/OFF durations and power consumption characteristics. If sub-meter data is unavailable for a selected house, transfer learning can be employed. Transfer learning leverages training data from other houses, and its accuracy significantly depends on the similarity between the labelled aggregate training data and the aggregate testing data. On the other hand, if sub-meter data is not available or there is no information about the appliances in the house, unsupervised NILM (Non-Intrusive Load Monitoring) methods should be utilized.

In recent years, the development of machine learning techniques, driven by the availability of large datasets, has significantly advanced energy management research. Specifically, the collection of energy data from hundreds of households across different countries has encouraged researchers to explore machine learning approaches for NILM. Two primary categories of machine learning algorithms used for NILM are supervised and unsupervised learning [131].

For example, CNNs and GRUs were used to accurately predict appliance states and energy consumption on REDD, REFIT, and UK-DALE [43]. Their results demonstrated that the trained networks could transfer effectively across datasets with minimal performance drops, compared to baseline models trained and tested on the same dataset—even when applied to unseen households within the same dataset. Both CNN- and GRU-based networks showed comparable performance; however, the GRU-based network had fewer trainable parameters and was less complex than its CNN counterpart.

Transfer learning is particularly promising for training NILM models for the following reasons:

1. Cost Reduction: Obtaining ground truth active power data for individual appliances is expensive. Transfer learning reduces the need for installing sensors on every appliance by leveraging pre-trained models.
2. Computational Savings: Pre-trained models can be reused for different appliances or domains, significantly reducing computational costs [69].

Scenarios for Transfer Learning:

Transfer learning can be categorized into three scenarios based on the training and testing datasets:

- Scenario A: Train and test on different houses within the same dataset (e.g., train on one house from the REFIT dataset and test on another house in REFIT).
- Scenario B: Train and test on different datasets within the same country (e.g., train on REFIT and test on UK-DALE).
- Scenario C: Train and test on datasets from different countries (e.g., train on REFIT and test on REDD).

Among these, Scenario C is the most challenging due to differences in appliance work cycles, which may arise from variations in temperature, lifestyle, electrical network characteristics, and manufacturers between countries.

Transfer Learning Results:

When tested on the REFIT dataset, the performance aligned with expectations. Training and testing within the same house yielded the highest accuracy. Notably, the fridge-freezer disaggregation performed better than other appliances when comparing similar power consumption patterns, such as between houses 2 and 17. This consistency underscores the importance of similarity in appliance characteristics for effective transfer learning. The detailed results are presented in Table 4.17.

Table 4.17: Result of transfer learning

			Train with %60 of refit house 2 Test with % 40 of refit house 2				Train with %60 of refit house 2 Test with % 100 of refit house 17				
Sample rate	Function	appliance	PR	RE	FM	Acc	PR	RE	FM	Acc	feature
63-64 sec	DT	kettle	0.63	0.83	0.71	0.66	0.19	0.23	0.21	0.31	ΔP, P/PMax
63-64 sec	DT	microwave	0.68	0.39	0.49	0.60	0.15	0.04	0.06	0.45	ΔP, P/PMax
63-64 sec	DT	washing machine	0.19	0.16	0.17	0.25	0.03	0.13	0.04	0.01	ΔP, P/PMax
63-64 sec	DT	Fridge-freezer	0.68	0.77	0.72	0.69	0.39	0.99	0.56	0.40	ΔP, P/PMax

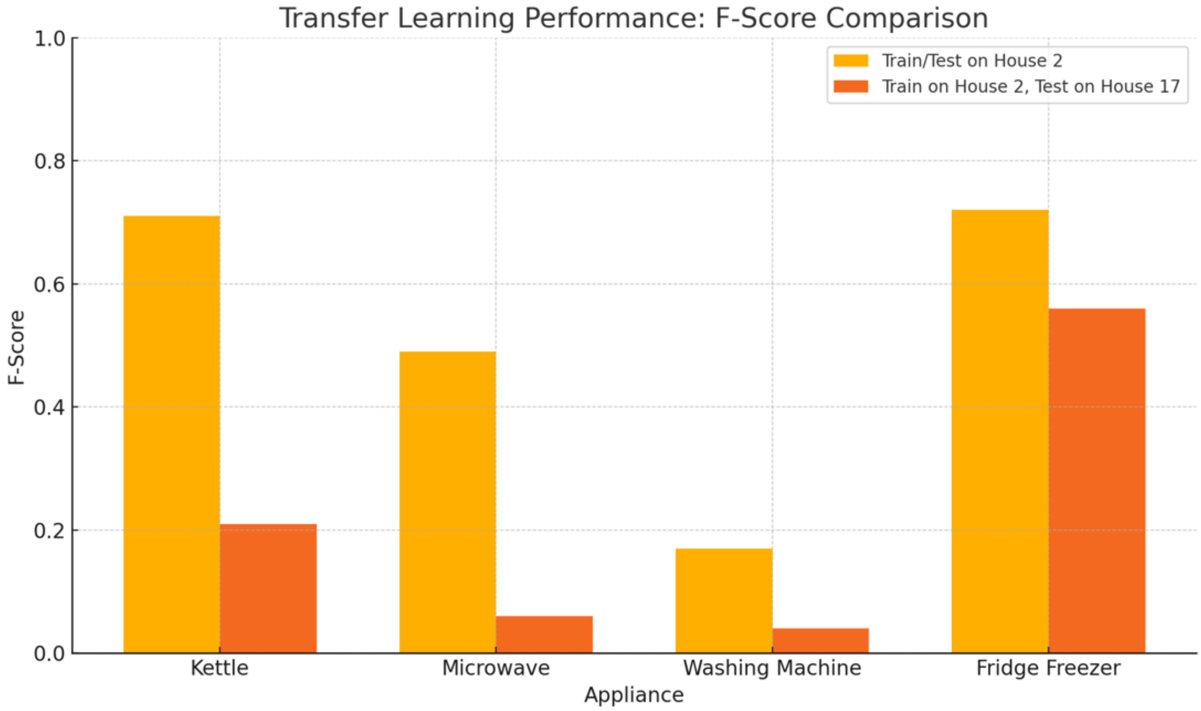


Figure 4.17: Result of transfer learning

The above figure 4.17 and table 4.17 compares the F-Score performance for transfer learning applied to four appliances. The left bars represent training and testing on REFIT House 2, while the right bars show the performance when testing on a different household (House 17) after training on House 2.

Figure 4.17 and Table 4.17 compare the F-Score performance for transfer learning applied to four different appliances. The left bars represent the results obtained when both training and testing are performed on REFIT House 2, while the right bars show the performance when the models trained on House 2 are tested on a different household (House 17). The results indicate that kettle and fridge-freezer maintain relatively higher F-Scores under transfer learning conditions, although their performance still declines notably—for instance, the kettle’s F-Score drops from 0.71 to 0.21. In contrast, the microwave and washing machine exhibit severe degradation when transferred to a new household, highlighting the strong dependence of their power signatures on individual usage patterns and operational variability. Overall, these findings underscore the inherent difficulty of generalizing NILM models trained on one household to another, particularly for appliances characterized by irregular events or diverse load profiles.

Similarity of the appliance load of the train and test data set has a high effect on performance of transfer learning

Table 4.18: Transfer learning in different situation

			Train with refit house 3 Test with refit house 2				Train with refit house 3&7 Test with refit house 2				
Sample rate	Function	appliance	PR	RE	FM	Acc	PR	RE	FM	Acc	feature
63-64 sec	ANN	Dishwasher	0.60	0.05	0.1	0.51	0.70	0.80	0.75	0.71	ΔP, P/PMax

As shown in Table 4.18, the inclusion of data from REFIT House 7 in the training set resulted in improved performance. When examining the dishwasher consumption power across these three houses—House 2 (2200W), House 3 (2100W), and House 7 (2300W)—it is evident that adding House 7 expands the range of detectable dishwasher power to between 2100W and 2300W.

This demonstrates that training a network with a broader range of data can enhance its performance by increasing its ability to generalize across variations. However, this approach also raises the risk of increased false positives. Careful consideration and strategies to mitigate false positives are essential when training models with a wider range of data.

4.7 SUMMARY

Here performance of DT, K-NN and DBSCAN algorithms in conjunction with pre-processing (median filtering, Graph Bilateral filtering and edge sharpening) for classification and estimating energy consumption of the top three appliances responsible for fires was evaluated. This helps us assess which of these simple NILM algorithms to consider for the next step of anomaly detection. Results indicate that pre-processing can improve the disaggregation performance of unsupervised DBSCAN and for appliances which are challenging to disaggregate, e.g., washing machine. DT has the best classification and disaggregation performance for all appliances of interest, comparable to state-of-the-art algorithms, and needing very little training data. A key limitation of the evaluated approaches is the low sampling rate and restricted communication bandwidth characteristic of smart meter infrastructures. These constraints reduce the availability of transient signatures and necessitate the use of simplified feature sets such as real power and ON-duration. However, an advantage of this constraint-driven design is that the proposed algorithms achieve practical deployability on low-power hardware and align closely with the capabilities of existing AMI systems. This positions the framework as both computationally efficient and directly compatible with real-world smart meter deployments. Furthermore, experiments show that the additional inclusion of aggregate power as a feature in addition to the change in power improves the performance of DT compared to previous literature. It also demonstrates improvement over the state-of-the-art with the proposed 2-step DT for improving the performance of the washing machine and dishwasher. A targeted subset of two-stage scenarios was evaluated to reflect the practical combinations of appliances that exhibit overlapping steady-state signatures. Scenarios were selected based on empirical frequency across REFIT and REDD households, similarity in power levels, and their impact on classifier ambiguity. Evaluating all possible combinations would not only be computationally prohibitive but would also include unrealistic or rarely occurring scenarios that add limited real-world value. The selected scenarios therefore represent the most challenging and practically relevant cases for improving disaggregation performance.

Only a limited number of two-stage scenarios were selected for detailed evaluation due to the practical constraints imposed by the characteristics of the dataset and the capabilities of low-power smart meter devices. The available appliances exhibited overlapping power profiles, limited variability, and inconsistent event frequencies, which restricted the number of meaningful scenario combinations that could be analysed. Additionally, to ensure that the results were interpretable and computationally feasible, the study focused on representative appliance groups with sufficient event density and clear operational patterns. This selective approach allowed the evaluation to remain rigorous while avoiding over-generalisation from scenarios that were not supported by robust data. DT may not be the best choice for transferability on unseen houses and meeting our scalability criteria, and as such further work on transfer learning with DT is needed. Building on these insights, Chapter 5 moves beyond disaggregation to focus on anomaly detection. It applies NILM-derived outputs to identify abnormal behaviours in critical appliances such as fridge-freezers, where anomalies not only indicate inefficiency but also potential safety risks. Through statistical and clustering-based methods, Chapter 5 explores how NILM can be extended from appliance identification to practical applications in fault detection, household safety, and energy efficiency.

Chapter 5

4 Detect Anomaly in appliances

5.1 Introduction

Anomaly detection in appliances is a critical task in energy management systems, particularly when integrated with Non-Intrusive Load Monitoring (NILM). It involves identifying unusual or faulty behaviour in electrical devices based on deviations from their typical power consumption patterns. By analysing features such as operating duration, power level, frequency of use, and time-of-day activity, anomaly detection systems can flag appliances that are malfunctioning, operating inefficiently, or consuming energy outside their expected behaviour. This is essential for enhancing household safety, enabling predictive maintenance, and reducing unnecessary energy waste. Notably, the BBC reported that malfunctioning appliances, particularly white goods, caused nearly 12,000 fires in Great Britain over three years (from January 2011 till March 2014) [10] It has been revealed that appliances such as washing machines, tumble dryers, and dishwashers are among the most common causes of household fires, based on available data, the following appliances have been identified as leading contributors to such incidents:

Table 5.1: Impact of Appliance Anomalies on Household Fire Incidents[10]

Appliance	Percentage of Appliance Fires	Approximate Number of Fires
Washing Machines	14%	1,723
Tumble Dryers	12%	1,456
Dishwashers	11%	1,324
Cookers	9%	1,080
Fridges	7%	861
Central Heating	5%	606
Toasters/Grills	4%	495
Televisions	3%	372
Washer Dryers	2%	225
Irons	1%	92

While the Scottish Fire and Rescue Service attributed 340 fires in 2019 to tumble dryers, washing machines, fridge-freezers, and dishwashers [11]. Approximately one-third of total energy consumption occurs in buildings, and anomalies or faults in appliances can lead to energy waste of up to 20% in these settings [132]. According to [133], the total electricity demand in the UK in 2020 was around 330.0 TWh. This implies that eliminating appliance-related anomalies in buildings could potentially save up to 19.8 TWh of electricity. Such a reduction not only helps lower energy costs but also significantly reduces the carbon footprint. Notably, around 40% of the UK's electricity in 2020 was generated from fossil fuels [133], meaning that addressing these inefficiencies could lead to a considerable decrease in emissions. Anomaly detection in household appliances is becoming increasingly important as we move toward smarter and more energy-efficient homes. Several studies have explored anomaly detection in household appliances using various methods. For instance, Haroon et al. and Marco et al. employed sub-metered data to identify anomalies in home appliances [132][134]. Similarly, Saeed et al. used sub-metered data to detect anomalies in standard and smart refrigerators [74]. Another study applied NILM to disaggregate appliance-level loads, then used the UNUM algorithm to detect anomalies in air conditioners and refrigerators [9].. This algorithm is based on computing the average and standard deviation of energy consumption, and it applies three heuristic rules: (1) the appliance switches between ON and OFF states too frequently; (2) the appliance remains in the ON state for an unusually extended period; and (3) the appliance remains in the ON state for a prolonged period while the OFF duration is also longer than normal. However, their findings indicated that supervised NILM, in its current form, lacks the necessary accuracy for reliable anomaly detection. In this chapter, we explore how unusual or faulty appliance behaviour can be identified without the need for installing sensors on every individual device. Instead, by using Non-Intrusive Load Monitoring (NILM), it's possible to detect when something isn't quite right—like a fridge running too long—just by analysing the overall electricity use recorded by a single smart meter. In this chapter, we refer to appliances that behave not as per manufacturer specs as "anomalous appliances," and each instance of unusual behaviour as an "anomaly." Traditionally, this kind of monitoring required separate meters for each appliance, which can be expensive and difficult to scale in modern homes that often have dozens of electrical devices. NILM offers a more practical approach by estimating the energy used by individual appliances based on patterns in the total electricity data. As smart meters become more common, researchers and energy companies are showing growing interest in using NILM to provide detailed feedback to households. This kind of insight not only helps improve safety and maintenance but also empowers people to save energy—sometimes by as much as 15%. While real-world data can be noisy and smart meters often don't record energy use in high detail, new algorithms and access to more granular data (such as measurements every few seconds inside the home) are making NILM more effective. The choice to focus on low-power smart meter devices introduces both advantages and inherent limitations that influence the design of the proposed NILM approach. On the one hand, these meters offer reduced installation costs, improved energy efficiency, and easier integration into household systems, making them suitable for large-scale deployment. However, their limited sampling rate and restricted communication bandwidth constrain the type and volume of data available for disaggregation. High-frequency features such as transient signatures or harmonic content cannot be reliably captured, necessitating the development of algorithms that operate effectively with low-resolution power data. The methodology adopted in this thesis is therefore intentionally lightweight and robust to missing or delayed measurements, reflecting realistic operational conditions for low-power, bandwidth-limited smart metering infrastructures. In this chapter, we build on these developments to show how NILM can help identify appliance faults in a practical and scalable way.

5.2 Contributions over the state of the art

Anomaly detection in energy systems has traditionally relied on high-resolution measurements, submetered data, or simple rule-based thresholds. While recent studies have explored data-driven and machine learning approaches, few have integrated anomaly detection directly with NILM outputs, particularly under smart-meter constraints. Within the research group, anomaly detection has been explored in isolation; however, hybrid frameworks combining unsupervised learning with statistical validation have not been fully developed or evaluated on real smart-meter datasets.

Existing approaches suffer from limited scalability, reliance on high-frequency data, and weak interpretability when applied to aggregate smart-meter measurements. There is a clear gap in the development of anomaly detection frameworks that are both computationally lightweight and robust enough for deployment using NILM-derived information.

This chapter addresses these gaps by introducing a novel hybrid anomaly detection framework that integrates Unsupervised Graph Signal Processing with statistical post-processing techniques. The proposed approach operates solely on aggregate smart-meter data and demonstrates effective detection of abnormal appliance behaviour under realistic conditions. By emphasising deployability, interpretability, and safety relevance, this contribution extends beyond existing methods and aligns with the third thesis contribution outlined in Chapter 1.

5.3 Appliance Characteristic

Each anomaly may be caused either by the appliance itself or by user behaviour. For instance, leaving the refrigerator door open for an extended period, running the air conditioner while the room is unoccupied, or operating it with open windows can all lead to abnormal energy consumption. Therefore, detecting anomalies—and understanding their causes—can be closely linked to occupancy patterns [136]. To capture such anomalies effectively, it is important to assess occupancy levels. While occupancy detection can provide valuable insights into energy usage, the tools required for this purpose—such as passive infrared (PIR) sensors, magnetic reed switches, or cameras—are often expensive to install and maintain [136]. In some cases, submeter data can serve as a cost-effective alternative.

An important contribution is the comprehensive review [136], which surveys AI-based methods for detecting abnormal energy-consumption patterns in buildings. The authors provide an extensive taxonomy of current techniques, categorizing them according to the type of learning algorithm (supervised, unsupervised, hybrid), detection granularity (appliance-level, building-level), feature extraction methods, and computational environments. The review identifies several pressing challenges, including the lack of standardized definitions for energy anomalies, the limited availability of labelled datasets, and the absence of consistent evaluation metrics across studies. Additionally, the authors highlight the need for reproducible research platforms and methods that preserve user privacy. These insights underline the complexity of deploying scalable and accurate anomaly detection systems and offer a foundation for guiding future developments in this area[136]

Household appliances display a wide range of operational behaviours based on their intended function, design, and user interaction. These variations affect energy consumption patterns and directly influence the performance of load disaggregation and anomaly detection techniques. Accurately characterizing these appliances is therefore crucial for developing effective pattern recognition approaches in their electrical load signature.

- Cyclic Appliances

Appliances such as refrigerators, freezers, and heat pumps exhibit cyclic operating patterns driven by internal thermostatic control systems. Their power consumption alternates between ON and OFF states in regular intervals to maintain internal conditions such as temperature. For example, fridge-freezers cycle the compressor on and off based on internal temperature thresholds. This predictable behaviour is useful for modelling and detecting anomalies such as compressor faults or abnormal cycle frequency [74][9].

- Multi-State Appliances

Devices like washing machines, dishwashers, and tumble dryers operate in multiple phases, each with different energy demands—such as filling, heating, agitating, and spinning. These appliances demonstrate discrete operational states, making their energy consumption profiles complex and time-dependent. Faults in multi-state appliances may appear as missing or abnormally extended phases, requiring time-series analysis or finite-state modelling approaches for detection [132][74].

- High-Power Instantaneous Appliances

Examples include electric kettles, toasters, and microwave ovens, which consume a high amount of power for a short duration. These appliances often have clean ON/OFF transitions and high current draw, making them relatively easy to detect in aggregate signals using NILM. However, their simplicity also limits the depth of anomaly analysis, with most anomalies being user-related (e.g., overuse or repeated cycles) [9].

- Always-On or Standby Appliances

Appliances such as modems, routers, digital clocks, and smart sensors are typically low-power and remain ON throughout the day. While they do not significantly impact peak load, their constant operation contributes to base-load energy use. Anomalies in these devices can be subtle—like a switch to an unintended active mode or slow energy creep due to firmware issues [136].

- User-Driven and Irregular Appliances

Entertainment systems, space heaters, and portable fans are often influenced by user habits and environmental factors. Their energy usage can vary widely across households and seasons. Modelling these appliances for anomaly detection is more challenging, as irregularity does not necessarily imply abnormality. Additional contextual data such as occupancy or weather may be required for meaningful insights [9][136].

In this study, we focus on fridge-freezers for appliance-level anomaly detection due to their high usage frequency and potential safety risks. The fridge-freezer is one of the few appliances that operates continuously, 24 hours a day, 7 days a week, making it an ideal candidate for monitoring long-term operational behaviour and detecting deviations that may signal mechanical inefficiency or failure. Monitoring appliance offers a dual benefit: improving energy efficiency and enhancing household safety through early fault detection.

5.3.1 Fridge-Freezers

Anomaly detection in fridge-freezers is an essential task in energy management systems, particularly when using Non-Intrusive Load Monitoring (NILM) techniques. Fridge-freezers typically exhibit predictable power consumption behaviours characterized by regular ON/OFF cycles, a stable power draw during compressor operation (typically between 100–250 W), and consistent daily energy usage. Any deviation from this pattern can be indicative of abnormal appliance behaviour. For example, unusually long compressor ON durations may suggest issues such as door seal degradation or a malfunctioning thermostat, while consistently high power levels could indicate a failing compressor [135]. Additionally, frequent short cycling of the compressor may reflect a refrigerant leak or control system fault, whereas a complete lack of power draw could point to a hardware failure or power disconnection [18] .

Although the fridge-freezer group does not have a high power demand at any given time, it operates continuously (24/7), resulting in substantial cumulative energy consumption. Moreover, the fridge-freezer follows a regular and repetitive load cycle, making it a suitable reference for modelling appliances with similar cyclic behaviour. For this reason, the fridge-freezer is selected as the focus for analysing the proposed algorithm in this study.

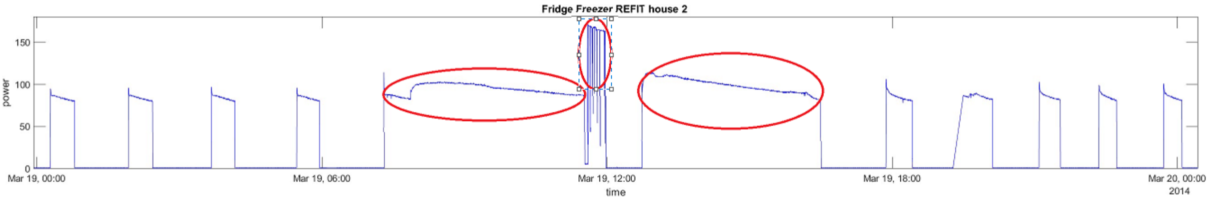


Figure 5.1: Power consumption of Fridge-Frezer for House 2 RFIT data on 19th of March 2014

In the above graph the active power consumption of fridge-freezer (REFIT data house 2) has been shown with three anomalies from 7:18 to 11:31 & 11:36 to 11:56 and 12:44 to 16:30 on 19 of March 2014 in house 2 REFIT data

Regular Cycling Pattern (Normal Behaviour):

- Clear cyclic ON/OFF compressor patterns are visible, particularly on the left and right sides of the graph (early morning and evening hours).
- These reflect typical fridge-freezer operation — turning on to cool and off when the set temperature is reached.

Anomalous Events (Circled in Red):

- **Middle Section (~12:00 PM):** There is an abnormal spike in power usage with rapid fluctuations. This may indicate:
    - A power disturbance,
    - Compressor malfunction,
    - Foresting period,
    - Or transient fault conditions.
- **Before and After the Spike:** long cycle duration :
    - These could suggest potential **frost buildup**.
    - Forgetting to close the fridge-freezer door .

Fridge-freezers are designed to operate continuously and maintain stable internal temperatures. Anomalies in their behaviour can indicate inefficient operation, developing faults, or safety hazards. These anomalies can be classified into several key types:

1. Behavioural Anomalies (Operation Pattern Deviations)

These anomalies are identified by deviations from the fridge-freezer's expected ON/OFF cycling behaviour.

- **Extended ON Duration:** Normally, the compressor operates intermittently. If it remains ON for unusually long durations, this may indicate issues such as a door left open, warm food recently placed inside, or a failing thermostat. Prolonged compressor activity increases energy consumption and accelerates component wear [137].
- **Frequent ON/OFF Cycling (Short Cycling):** When the compressor turns on and off rapidly, it can signal low refrigerant levels, clogged filters, or thermostat malfunctions. This behaviour increases energy use and can shorten compressor lifespan [138].
- **Unusual OFF Periods:** If the fridge stays OFF for longer than expected, it might suggest sensor faults or control board failure, risking food spoilage due to temperature drift [139].

2. Power Consumption Anomalies

These refer to unusual patterns in energy usage, often detectable through NILM techniques.

- **Higher-than-Normal Power Draw:** This might be due to dirty condenser coils, faulty compressor components, or overloading (e.g., blocking internal airflow). These problems force the system to work harder, drawing more power [12][18].
- **Lower-than-Normal Power Draw:** An underpowered system may reflect degraded electrical components or a compressor unable to reach optimal performance, potentially leading to ineffective cooling [140].
- **Sudden Power Spikes/Drops:** Spikes can occur due to compressor motor startup issues, relay faults, or short circuits. Conversely, sudden drops may indicate a component failure or loose connection [141].

3. Mechanical and Component-Based Anomalies

Mechanical failures often result in both functional and energy anomalies.

- **Compressor Malfunction:** A weakened or failing compressor leads to erratic temperature control and high energy use. It may run continuously or fail to start [142].
- **Defrost Heater or Timer Failure:** A broken defrost system causes frost to accumulate on the evaporator coils, blocking airflow and overloading the compressor [78].
- **Fan Motor or Vent Blockage:** Internal fans circulate cold air. If blocked or broken, temperature gradients form, causing the compressor to work harder, especially in frost-free designs [139].

4. User-Related and Environmental Anomalies

User behaviour and external factors significantly impact fridge operation:

- **Frequent Door Openings:** Each opening lets warm air enter, requiring additional cooling cycles. This is more critical during hot weather or if door seals are damaged [138].
- **Overfilling or Blocking Vents:** Overpacking blocks internal airflow and affects temperature regulation, resulting in inefficient compressor cycles [137].
- **Ambient Temperature Effects:** Placing the appliance near heat sources or in warm environments causes increased compressor activity [140].

5. Sensor and Control System Anomalies

Sensors and control boards govern how the fridge maintains temperature.

- **Faulty Thermostat or Temperature Sensor:** If these components malfunction, they may cause overcooling, undercooling, or erratic cycling behaviour [22].
- **Control Board Malfunctions:** A failing electronic control board may lead to irregular commands sent to the compressor or fan, or even full shutdown of the appliance [141].

6. Safety-Related Anomalies

These anomalies pose significant risks to property and human safety:

- **Overheating or Fire Risk:** Faulty wiring, overworked compressors, or old appliances can overheat and ignite. In the UK, fridge-freezers have been implicated in numerous domestic fires [143].

- **Electrical Smell, Smoke, or Sparks:** Indications of severe electrical faults that may not only damage the appliance but can result in household fire if not immediately addressed.
- **Unusual Noise Patterns:** Clicking, buzzing, or loud humming sounds may suggest fan issues, ice buildup, or motor degradation.[139]

Many studies, including the work by [9], primarily rely on energy consumption data for anomaly detection. However, for fridge-freezers, energy consumption is notably influenced by ambient temperature. Incorporating direct temperature measurements into the detection model would require external temperature sensors, which increases system complexity and cost. As an alternative, we propose using seasonal or monthly variations as a proxy for ambient temperature, accepting a reasonable margin of error.

Additionally, the number and duration of compressor cycles are time-dependent. During nighttime hours, both the frequency and duration of these cycles typically decrease, primarily due to reduced door openings. This reduction may also be influenced by lower ambient temperatures, reduced indoor heating, or generally cooler conditions overnight .

Tables 5.2, 5.3, 5.4, and 5.5 present the cycle duration and frequency for REFIT houses 1 and 2, demonstrating the influence of time-of-day and seasonal variations on fridge-freezer operation. These values were calculated directly from submetered data.

Table 5.2: REFIT house 1 duration in seconds and number of fridge cycles

REFIT house 1 fridge	From 24:00 till 6:00am		From 6:00 am till 12:00 pm		From 12:00 pm till 18:00		From 18 till 24:00	
Date	Number of cycles	Average ON duration of one cycle	Number of cycles	Average ON duration of one cycle	Number of cycles	Average ON duration of one cycle	Number of cycles	Average ON duration of one cycle
17/1/2014	2	1470	3	1490	2.5	1500	2.5	1530
18/1/2014	3	1480	2	1530	3	1650	2.5	1500
19/1/2014	2.5	1500	3	1620	3.5	1640	3.5	1560
20/1/2014	3	1540	2	1500	3	1660	3	1560
21/1/2014	2	1560	2.5	1590	2.5	1590	2.5	1590
22/1/2014	2.5	1560	2	1620	3	1520	2.5	1560
23/1/2014	2.5	1590	3	1729	3	1640	2.5	1560
1/7/2014	3	2060	2.5	2370	2.5	2310	3	2110
2/7/2014	2.5	2070	2.5	2100	3	2520	3	2140
3/7/2014	2	2100	3	2220	3	3120	3	2580
4/7/2014	3	2080	2.5	2190	2.5	2580	3	2260
5/7/2014	3	2360	2	2250	3	2280	2	2310
6/7/2014	3	2220	3	2360	3	2480	3	2180

Tables 5.2 present the number of compressor cycles and their average ON durations for fridge in REFIT Houses 1, segmented by time of day and season. For instance, in REFIT House 1, during the winter period (17–23 January 2014), the number of cycles between 00:00 and 06:00 was generally lower than during daytime periods, and the ON durations were shorter, consistent with reduced user interaction and lower temperature at night.

In contrast, during the summer period (1–6 July 2014), both the number of cycles and their durations increased across all time segments, with notably longer ON durations—reaching up to 3120 seconds in the afternoon (12:00–18:00) on 3 July. This clearly indicates the impact of ambient temperature on fridge-freezer operation.

These observations reinforce the need to incorporate temporal and seasonal context into anomaly detection models. Rather than relying solely on raw energy data, factoring in time-of-day and seasonal trends allows for more accurate detection of truly anomalous behaviour while reducing false positives linked to normal environmental and behavioural variation.

Table 5.3: ON cycle duration in second of Fridge for REFIT house 1

REFIT house 1	3 Months				6 Months		12 Months
	Spring 21-March-2014 22-june-2014	Summer 22-june-2014 23-September-2014	Autumn 23-September-2014 22-December-2014	Winter 22-December-2014 20-March-2015	21-March-2014 23-September-2014	23-September-2014 20-March-2015	21-March-2014 20-March-2015
T_{mean}	1509.4	1738.2	1606.7	1522.9	1618.3	1564.7	1592.6

Further analysis of seasonal variation in fridge-freezer operation was performed using data from REFIT House 1, summarized in Table 5.3. The average compressor ON durations T_{mean} were calculated over different periods: 3-month seasons, 6-month semi-annual spans, and the full year.

The results reveal a clear seasonal influence on fridge-freezer behaviour. For example, during summer (22 June to 23 September 2014), the average ON duration reached 1738.2 seconds, the highest among all seasons. This aligns with higher ambient temperatures and increased door openings, both of which raise cooling demand. In contrast, winter (22 December 2014 to 20 March 2015) recorded a lower average of 1522.9 seconds, consistent with reduced thermal load and less frequent user interaction.

When aggregating over longer periods, the trend persists. The 6-month interval from spring to autumn (21 March to 23 September 2014) showed a higher mean duration (1618.3 seconds) compared to the autumn-to-spring interval (23 September 2014 to 20 March 2015) with a lower mean of 1564.7 seconds. Over the full year, the annual average ON duration was 1592.6 seconds.

These findings support the use of seasonal and temporal features as proxies for ambient temperature and user behaviour in NILM-based anomaly detection. Instead of deploying costly external sensors, models can be seasonally calibrated using historical appliance behaviour patterns. This enhances detection accuracy while maintaining low deployment complexity.

Table 5.4: Monthly average ON durations (T_{mean}) in second and average ambient Temperature for the Upright Freezer in REFIT House 1(Average temperature from Met office[144])

REFIT house 1	Upright Freezer REFIT House 1											
	Jan	Feb	March	April	May	Jun	July	Aug	Sep	Oct	Nov	Dec
T_{mean}	1405	1517	1374	1426	1507	1749	1823	1729	1669	1605	1634	1536
Temp	3.5	3.8	5.4	7.9	9.7	12.8	14.5	12.2	12.4	9.1	6.4	3.1

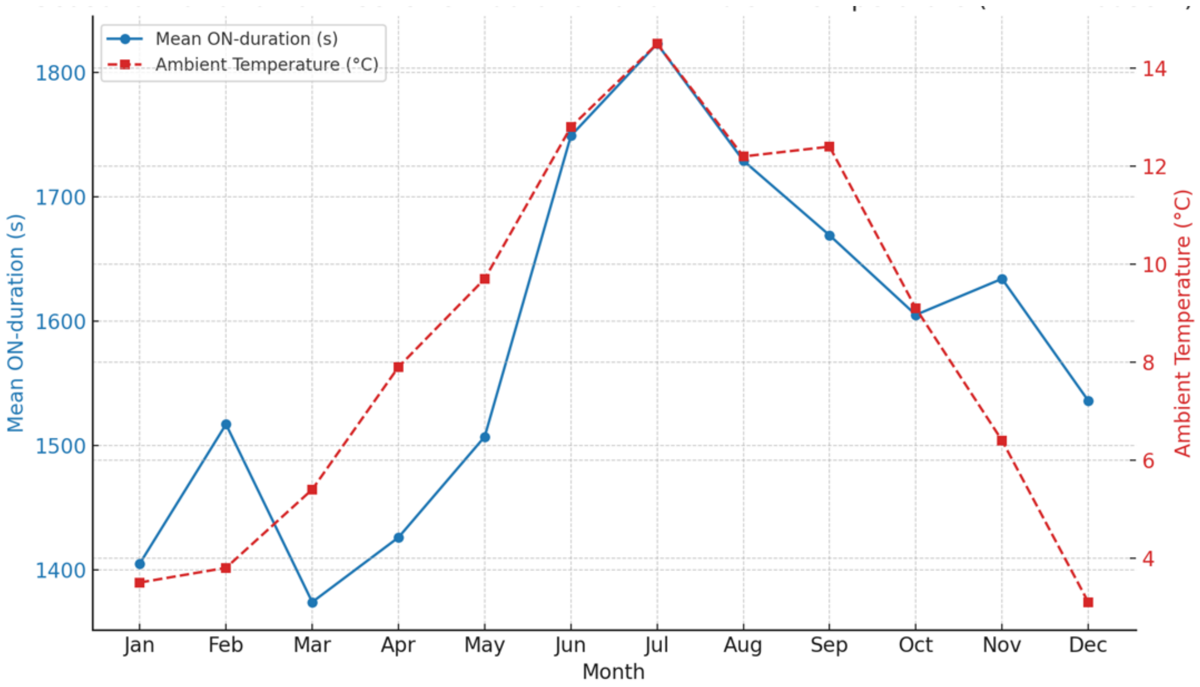


Figure 5.2: Monthly variation of Freezer ON-duration and ambient temperature (REFIT house 1)

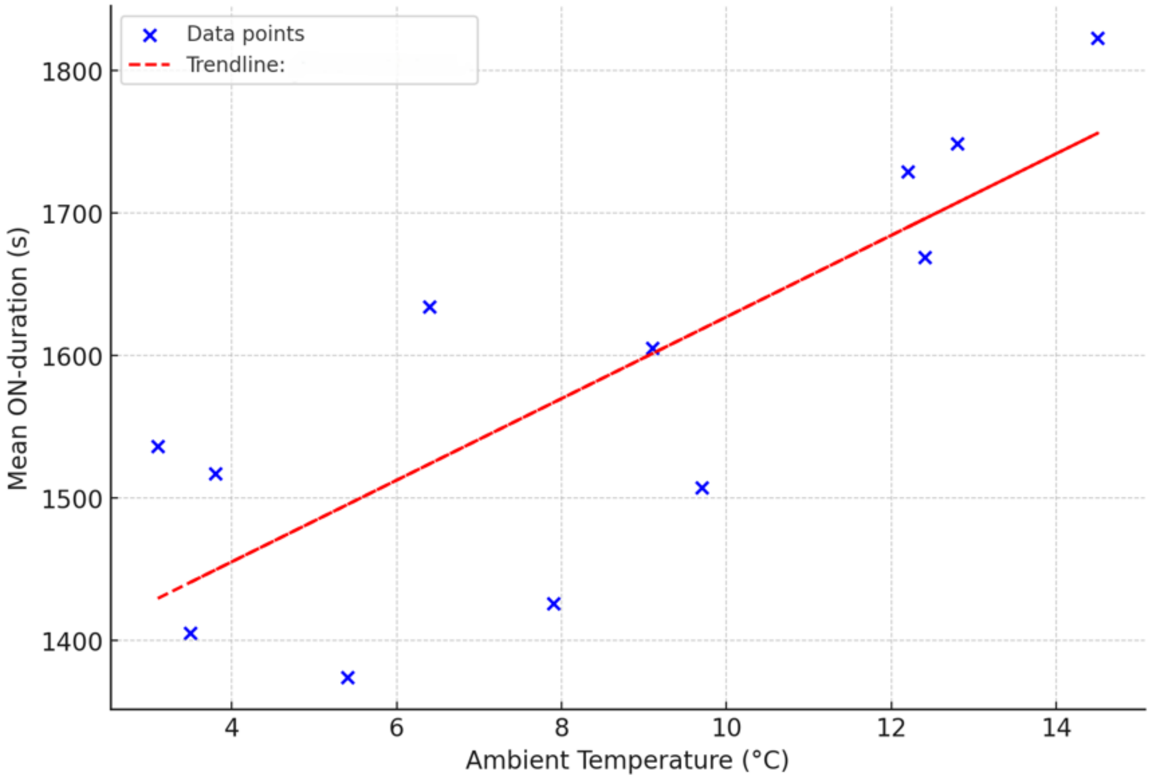


Figure 5.3: Relationship between Freezer ON-duration and ambient temperature(REFIT house 1)

In addition to the fridge data, the monthly average ON-durations for the upright freezer in REFIT House 1 also exhibit a clear seasonal pattern those shown in table 5.4. The ON-durations (T_{mean}) were lowest in March (1374 seconds) and highest in July (1823 seconds). This trend closely follows ambient temperature variations: during colder months, the freezer experiences reduced thermal load, requiring less compressor activity, whereas in warmer summer months, increased heat gain leads to longer compressor ON-durations.

Figure 5.2 presents the relationship between mean ON-duration and ambient temperature using a dual-axis graph, where the blue line represents the mean ON-duration (seconds) and the red dashed line denotes the ambient temperature (°C). The figure clearly illustrates the seasonal dependency of freezer operation, showing longer compressor cycles during the warmer months (June–August) and shorter cycles in winter. The deviations observed in January, February, and November are not anomalies but rather reflect realistic seasonal and behavioural dynamics. These variations can be attributed to factors such as changes in household activity patterns (e.g., periods of absence or altered usage during holidays) and differences between indoor and outdoor thermal conditions. While the ambient temperature data correspond to outdoor measurements, indoor kitchen temperatures may not always correlate directly, especially during winter when heating or limited ventilation can influence appliance operation. Together, these factors contribute to the apparent outliers, yet the overall trend confirms that seasonal ambient temperature remains a dominant factor influencing compressor cycling behaviour.

To further examine this relationship, Figure 5.3 presents a scatter plot of ON-duration versus ambient temperature, with a fitted regression trendline. The positive slope of the regression line confirms that higher ambient temperatures are associated with longer compressor ON-times, reinforcing the conclusion that environmental conditions—particularly seasonal temperature variations—play a critical role in determining appliance cycling performance.

Table 5.5: Monthly average ON durations (T_{mean}) in second for the Fridge-Freezer REFIT House 2 (Average temperature from Met office [144])

REFIT house 2	Fridge-Freezer REFIT House 2											
	Jan	Feb	March	April	May	Jun	July	Aug	Sep	Oct	Nov	Dec
T_{mean}	2001	1993	1679	1858	1886	2120	2240	2124	2092	1909	2011	1996
Temp	3.5	3.8	5.4	7.9	9.7	12.8	14.5	12.2	12.4	9.1	6.4	3.1

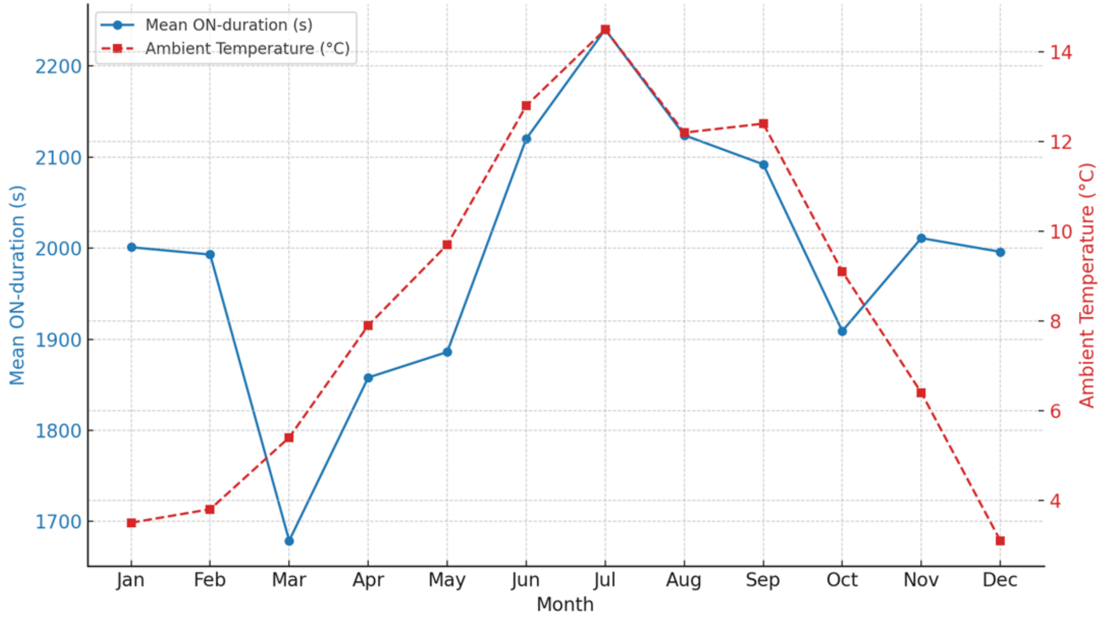


Figure 5.4: Monthly variation of Fridge-Freezer ON-duration and ambient temperature (REFIT house2)

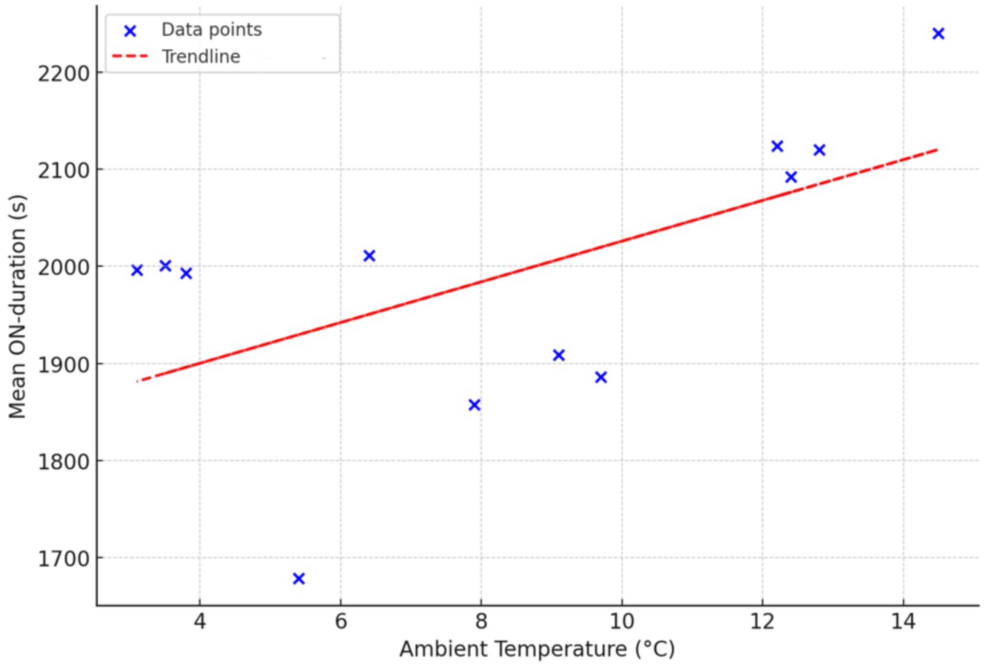


Figure 5.5:Relationship between Fridge-Freezer ON-duration and ambient temperature(REFIT house2)

The fridge-freezer in REFIT House 2 further confirms the strong seasonal dependency of compressor ON duration. Monthly averages from March 2014 to April 2015 show a consistent trend: higher ON durations in the summer months and relatively lower durations in the winter and early spring. The lowest T $_{mean}$ was recorded in March (1679 seconds), while the highest was in July (2240 seconds). This represents a substantial seasonal increase of over 33%, underscoring the impact of ambient temperature on operational load.

From May to September, ON durations remained elevated—ranging from 1886 to 2240 seconds—compared to lower values in January (2001 s), February (1993 s), and November (2011 s). Despite some variability, the data exhibits a clear summer peak and winter trough, which aligns with environmental thermal load and possibly user behaviour (e.g., more frequent access during warmer periods).

Figure 5.4 presents this relationship using a dual-axis graph, where the blue line represents the mean ON-duration (seconds) and the red dashed line shows ambient temperature (°C). The figure highlights the seasonal dependency of freezer operation, with longer compressor cycles in summer (June–August) and shorter cycles during winter.

To further examine the relationship, Figure 5.5 provides a scatter plot of ON-duration against ambient temperature with a fitted regression trendline. The positive slope confirms that higher ambient temperatures are associated with longer compressor ON-times. This reinforces the conclusion that environmental conditions, particularly seasonal ambient temperatures, play a significant role in appliance cycling behaviour.

Table 5.6: Monthly average ON durations (T mean) in second for the Fridge-Freezer REFIT House 3(Average temperature from Met office[144])

REFIT house 3	Freezer REFIT House 3											
	Jan	Feb	March	April	May	Jun	July	Aug	Sep	Oct	Nov	Dec
T $_{mean}$	630	652	789	897	919	787	936	895	770	742	694	638
Temp	3.5	3.8	5.4	7.9	9.7	12.8	14.5	12.2	12.4	9.1	6.4	3.1

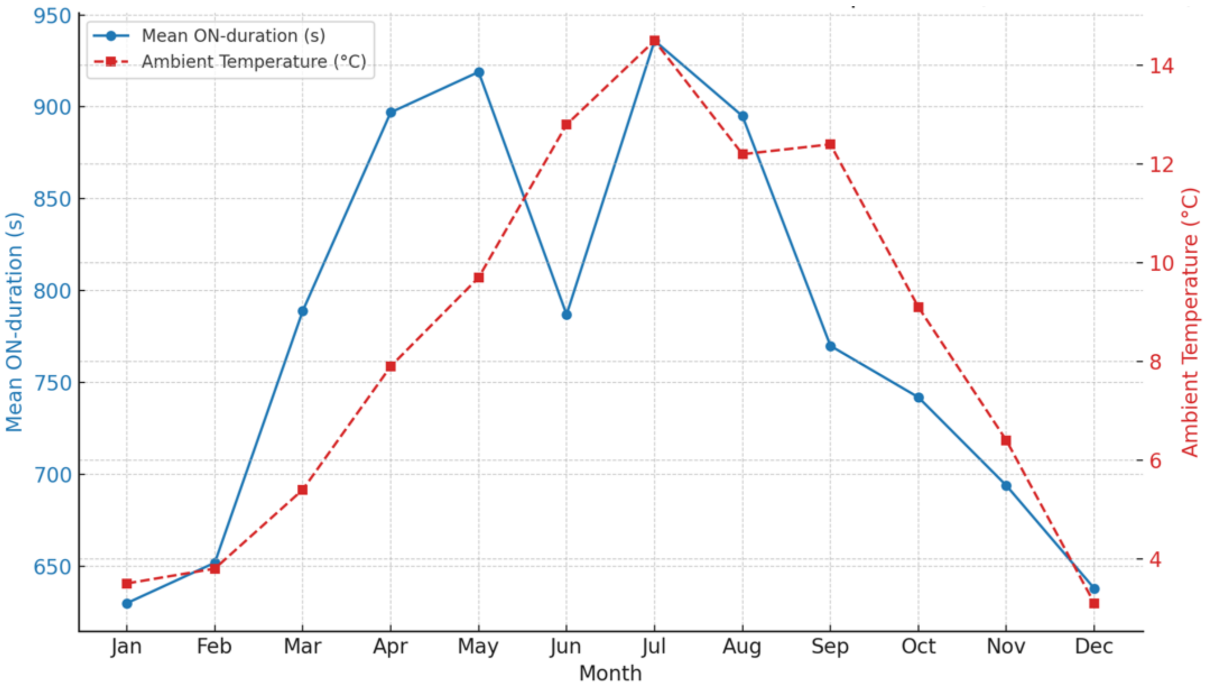


Figure 5.6: Monthly variation of Fridge-Freezer ON-duration and ambient temperature(REFIT house 3)

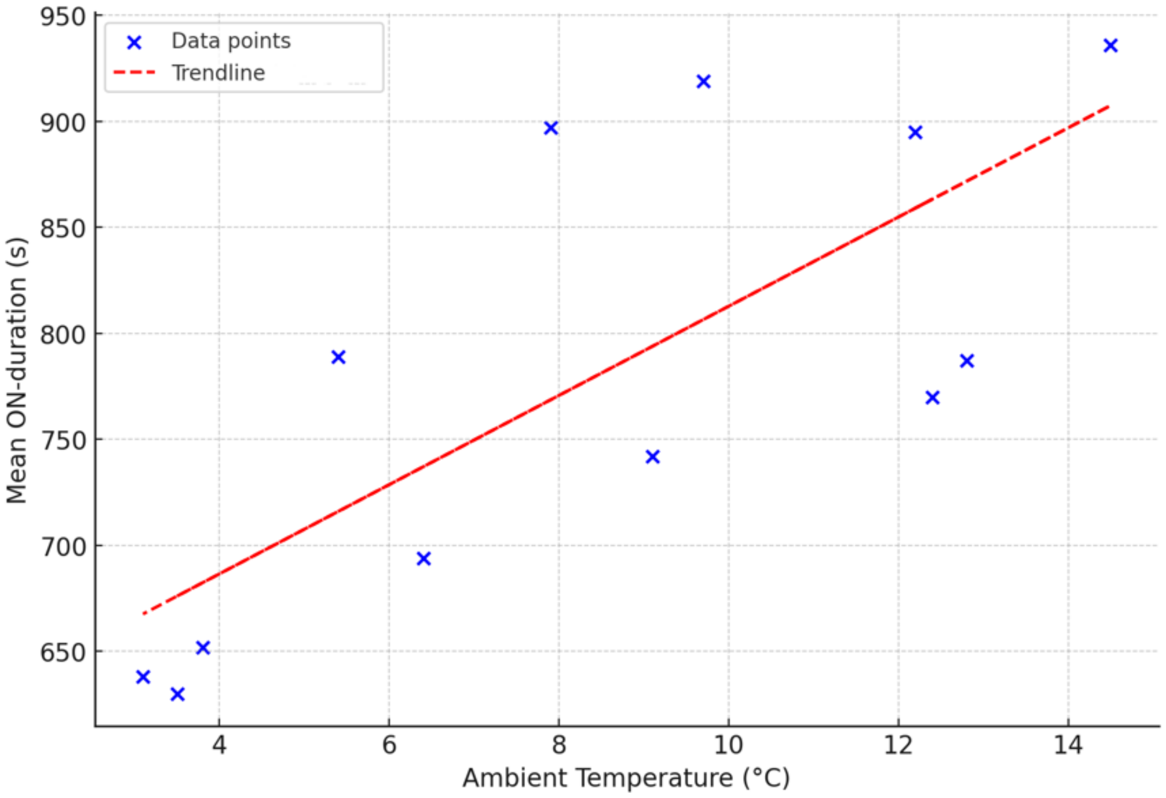


Figure 5.7:Relationship between Fridge-Freezer ON-duration and ambient temperature(REFIT house3)

The freezer in REFIT House 3 also demonstrates a strong seasonal dependency of compressor ON-duration. Monthly averages show a steady increase from winter into summer, with the lowest T mean recorded in January (630 seconds) and the highest in July (936 seconds). This represents a seasonal increase of nearly 49%, highlighting the considerable influence of ambient temperature on operational demand.

Between April and August, ON-durations remained elevated—ranging from 897 to 936 seconds—compared to much shorter cycles in the colder months (630–652 seconds in January–February). Despite some variability, the data reveals a clear summer peak and winter trough, reflecting increased thermal load in warmer conditions.

Figure 5.6 presents this relationship using a dual-axis graph, where the blue line represents mean ON-duration (seconds) and the red dashed line shows ambient temperature (°C). Longer compressor cycles are evident in the summer months (June–August), with shorter cycles during the winter period. To further investigate, Figure 5.7 provides a scatter plot of ON-duration against ambient temperature with a fitted regression trendline. The positive slope confirms that higher ambient temperatures are associated with longer ON-times. This analysis supports the conclusion that environmental conditions, particularly seasonal variations in temperature, play a major role in shaping freezer cycling behaviour in House 3.

Table 5.7: Monthly average ON durations (T mean) in second for the Fridge-Freezer REFIT House 5(Average temperature from Met office[144])

REFIT house 5	Fridge-Freezer REFIT House 5											
	Jan	Feb	March	April	May	Jun	July	Aug	Sep	Oct	Nov	Dec
T mean	682	664	629	665	765	773	824	787	779	729	764	656
Temp	3.5	3.8	5.4	7.9	9.7	12.8	14.5	12.2	12.4	9.1	6.4	3.1

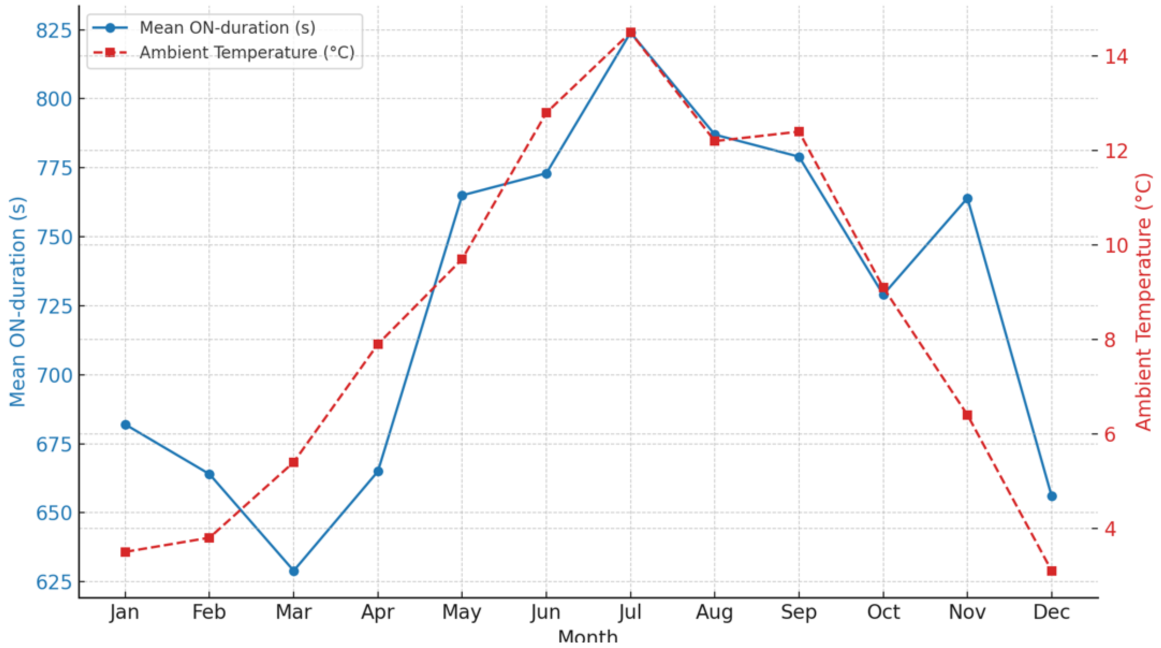


Figure 5.8: Monthly variation of Fridge-Freezer ON-duration and ambient temperature(REFIT house5)

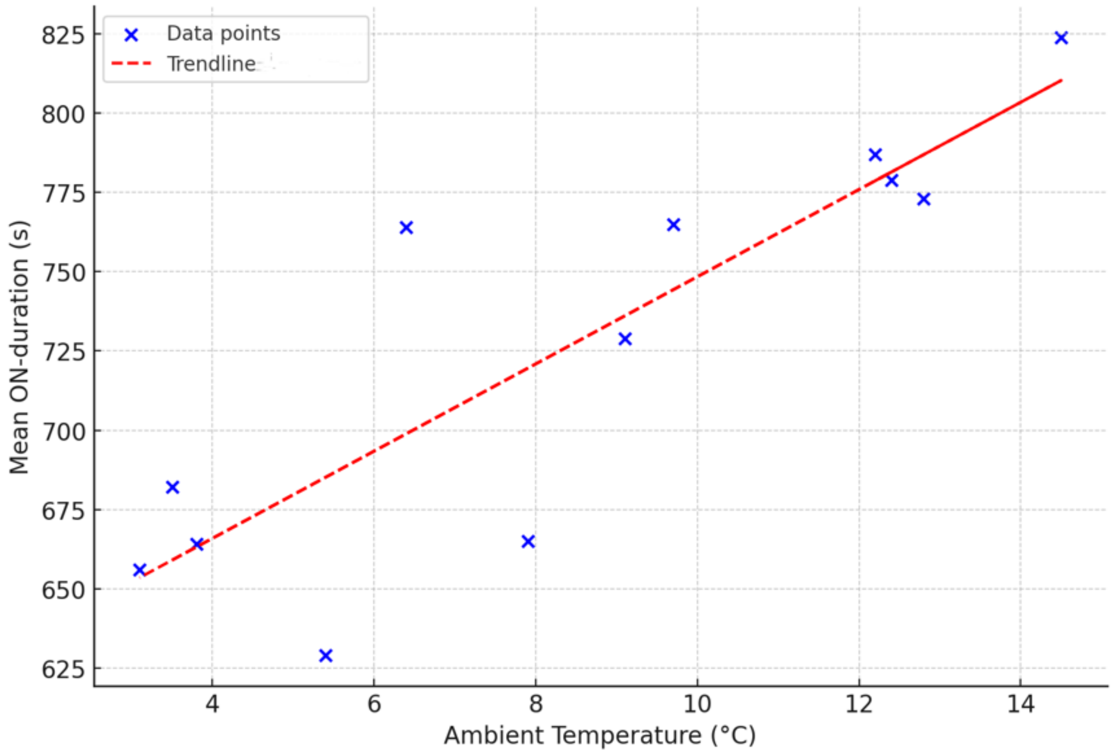


Figure 5.9: Relationship between Fridge-Freezer ON-duration and ambient temperature(REFIT house5)

The fridge-freezer in REFIT House 5 shows an even stronger seasonal dependency. The lowest T $_{mean}$ was recorded in March (629 seconds), while the highest occurred in July (824 seconds), representing a seasonal increase of about 31%. This indicates that the appliance's operational profile is highly sensitive to ambient temperature fluctuations.

From May through September, ON-durations consistently exceeded 765 seconds, peaking in midsummer, while the colder months of January–April exhibited much lower values (629–665 seconds). The resulting pattern reveals a distinct summer maximum and winter minimum, consistent with higher thermal demand in warm conditions and reduced compressor load in cooler months.

Figure 5.8 illustrates this seasonal relationship with a dual-axis graph, where the blue line shows mean ON-duration and the red dashed line indicates ambient temperature. Compressor activity is clearly elevated in the summer period, while shorter cycles dominate in winter. Figure 5.9 further examines this dependency through a scatter plot with regression trendline. The strong positive slope demonstrates that higher ambient temperatures are directly associated with longer compressor ON-durations, confirming the strong temperature sensitivity of the fridge-freezer in House 5

When compared to the upright freezer in REFIT House 1, the fridge-freezer in House 2 exhibits both higher absolute ON-durations and a greater degree of seasonal variation. This difference may be attributed to factors such as larger internal volume, older appliance age, or distinct household usage patterns. Notably, the consistent rise in ON-duration during warmer months underscores the importance of incorporating seasonal baseline calibration within anomaly detection frameworks to distinguish normal seasonal fluctuations from genuine faults.

A similar pattern is observed in the freezer in House 3 and the fridge-freezer in House 5, both of which display pronounced seasonal sensitivity. In House 3, ON-durations increased by approximately 49%

between winter and summer, whereas in House 5, compressor cycles extended by around 31% across the same period. These findings demonstrate that while freezers and fridge-freezers differ in their degree of temperature responsiveness, both consistently exhibit the characteristic summer peak and winter trough, confirming a strong environmental dependency in compressor cycling.

The deviations observed in certain months are not anomalies but reflect realistic seasonal and behavioural dynamics. Such variations can arise from changes in household activity patterns—for example, reduced occupancy or altered food storage practices during holidays—as well as from differences between indoor and outdoor thermal conditions. Although the recorded ambient temperature data correspond to outdoor measurements, indoor kitchen environments often deviate due to heating, ventilation, and cooking activity, particularly in colder months. These factors collectively explain the apparent outliers, yet the overarching trend affirms that seasonal ambient temperature remains the dominant driver of compressor behaviour.

Failure to explicitly account for seasonal variability may lead NILM-based models to misclassify normal summer operation as anomalous, especially for thermally sensitive appliances such as fridge-freezers. In the absence of direct ambient temperature measurements, monthly or seasonal average ON-durations (T_mean) can serve as an effective, low-cost proxy for environmental influence. Integrating such baseline seasonal adjustments significantly improves the accuracy and robustness of NILM-based anomaly detection, ensuring that detected deviations more accurately represent genuine appliance malfunctions rather than expected seasonal effects.

5.4 Detect anomaly with unsupervised NILM and rule-based approaches

Rule-based methods for anomaly detection have been widely adopted in NILM research due to their transparency, interpretability, and reliance on prior knowledge of appliance behaviour. [9] proposed a rule-based approach for detecting appliance anomalies based on disaggregated power signals derived from smart meter data. The technique defines a set of logical rules corresponding to the expected operational patterns of household appliances, including duration of use, power magnitude, time-of-day activity, and on/off cycling frequency.

For instance, a typical rule might state that a refrigerator should cycle every 2–4 hours within a certain power threshold. If this pattern is disrupted—such as prolonged inactivity or extended continuous operation—the system flags the appliance as anomalous. These rules are grounded in domain knowledge and can effectively identify behavioural deviations in appliances such as fridges, washing machines, kettles, and microwaves.

However, in Haroon et al.'s study, these rules were applied to appliance-level signals obtained via supervised NILM techniques. While supervised NILM algorithms benefit from labelled training data and known appliance signatures, their performance is highly contingent on the availability and quality of such data. Supervised models often suffer from poor generalization when applied to new households or unseen appliance types. Moreover, minor errors in disaggregation—such as load misclassification or missed events—can lead to significant inaccuracies in anomaly detection, particularly false positives or false negatives in rule violation checks.

To address these shortcomings, this work proposes the integration of a rule-based framework [8] with an unsupervised NILM technique based on Graph Signal Processing (UGSP). UGSP offers a data-driven, label-free disaggregation method that is inherently well-suited to event detection and sparse signal reconstruction. It models the aggregate power signal as a graph, where each node corresponds to a data point or event and edges capture similarities, and then employs edge-preserving filters to extract appliance-specific switching events and their associated power levels without requiring any prior training.

Previous research has demonstrated the effectiveness of UGSP for load disaggregation. In particular, [17] proposed a training-less NILM solution that exploits GSP principles to disaggregate aggregate power measurements without relying on submetered training data. By representing appliance events as graph signals and applying graph Laplacian filtering, they showed that GSP can successfully separate loads at low sampling rates, making it a promising candidate for practical smart-meter-based NILM. Such work establishes UGSP as a robust approach for load decomposition and event identification in electrical measurements.

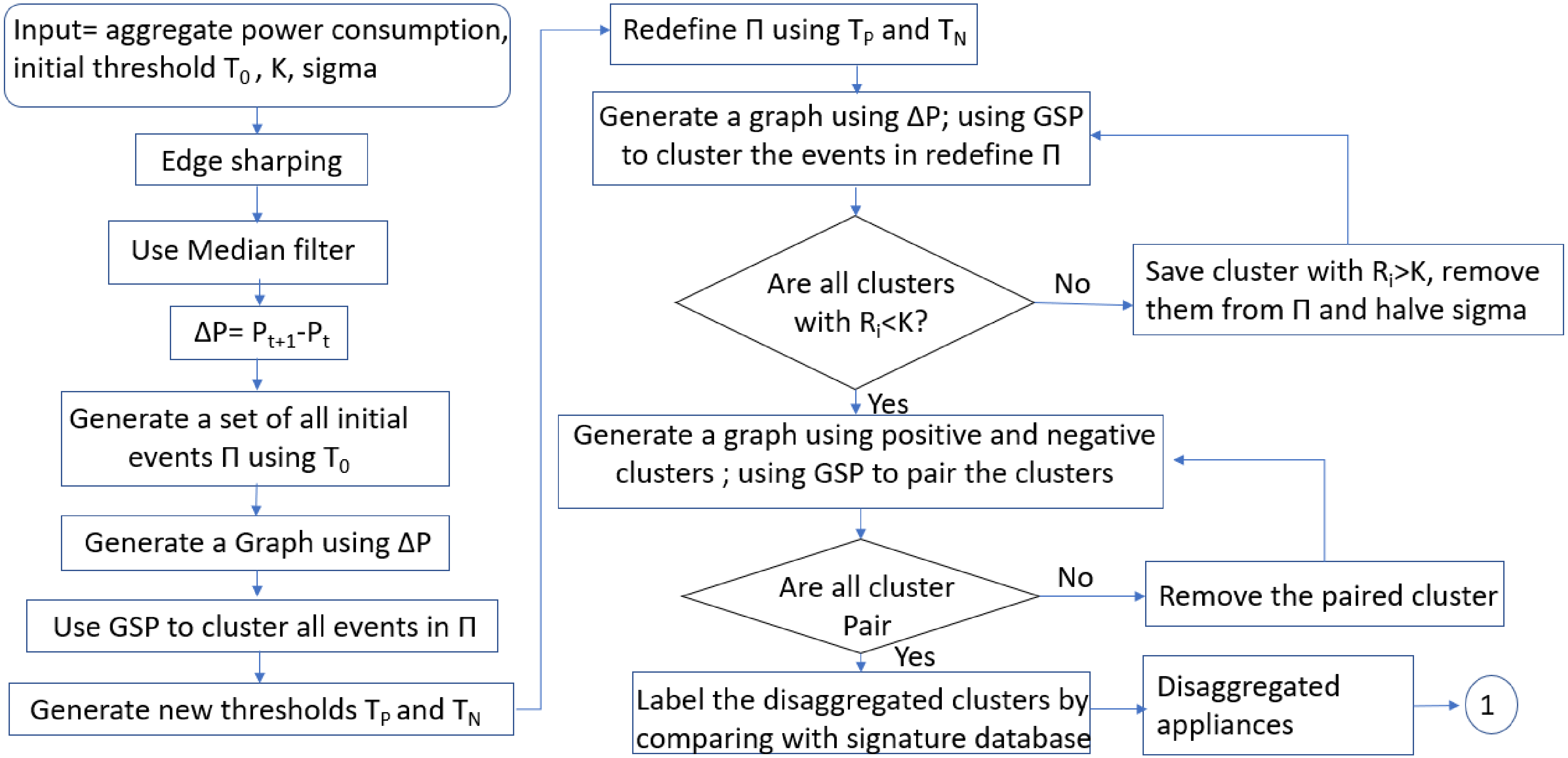
The novelty of this thesis lies in extending UGSP from disaggregation into the domain of anomaly detection. While previous studies have focused on recovering appliance-level signals, this work leverages the disaggregated outputs to detect abnormal behaviours, such as extended ON-durations in fridge-freezers. By combining UGSP with rule-based and clustering-based post-processing, the proposed framework preserves the temporal fidelity of appliance signatures while enabling diagnostic insights in an unsupervised manner. To the best of our knowledge, this is the first attempt to employ UGSP for appliance anomaly detection, thereby bridging the gap between NILM-based load monitoring and condition-based diagnostics.

By combining UGSP with a rule-based anomaly detection system, the proposed approach enhances robustness and applicability in real-world scenarios. The disaggregated appliance signals obtained through UGSP are used as input to the rule engine, which applies predefined conditions to detect anomalies. For example, if UGSP reveals that a refrigerator has not cycled for over 12 hours, despite an ambient temperature requiring cooling, the rule engine can infer a potential compressor malfunction.

This hybrid methodology leverages the generalization strength of unsupervised learning and the interpretability of rule-based logic, enabling effective anomaly detection in households where labelled appliance data is unavailable or impractical to obtain.

In Figures 5.10 and 5.11 flowchart for detect anomaly with unsupervised GSP has been shown, in graph 5.2 , unsupervised GSP has been used for disaggregate the load(as used in chapter 4) and then used rule based approach as shown in [9] used for detect anomaly.

Also in graph 5.3 unsupervised GSP has been used for disaggregate the load(as used in chapter 4) and then used GSP for classification of ON duration of fridge-freezer and Clusters with lower number of data consider as anomaly, this method can use for detect anomaly off line or use larger windows for detect anomaly that increase the availability of normal data in the selected widows.


Input= aggregate power consumption, initial threshold T_0 , K, sigma
Edge sharping
Use Median filter
$\Delta P= P_{t+1}-P_t$
Generate a set of all initial events Π using T_0
Generate a Graph using ΔP
Use GSP to cluster all events in Π
Generate new thresholds T_P and T_N
Redefine Π using T_P and T_N
Generate a graph using ΔP; using GSP to cluster the events in redefine Π
Are all clusters with R_i<K?
No
Save cluster with R_i>K, remove them from Π and halve sigma
Yes
Generate a graph using positive and negative clusters ; using GSP to pair the clusters
Are all cluster Pair
No
Remove the paired cluster
Yes
Label the disaggregated clusters by comparing with signature database
Disaggregated appliances
1

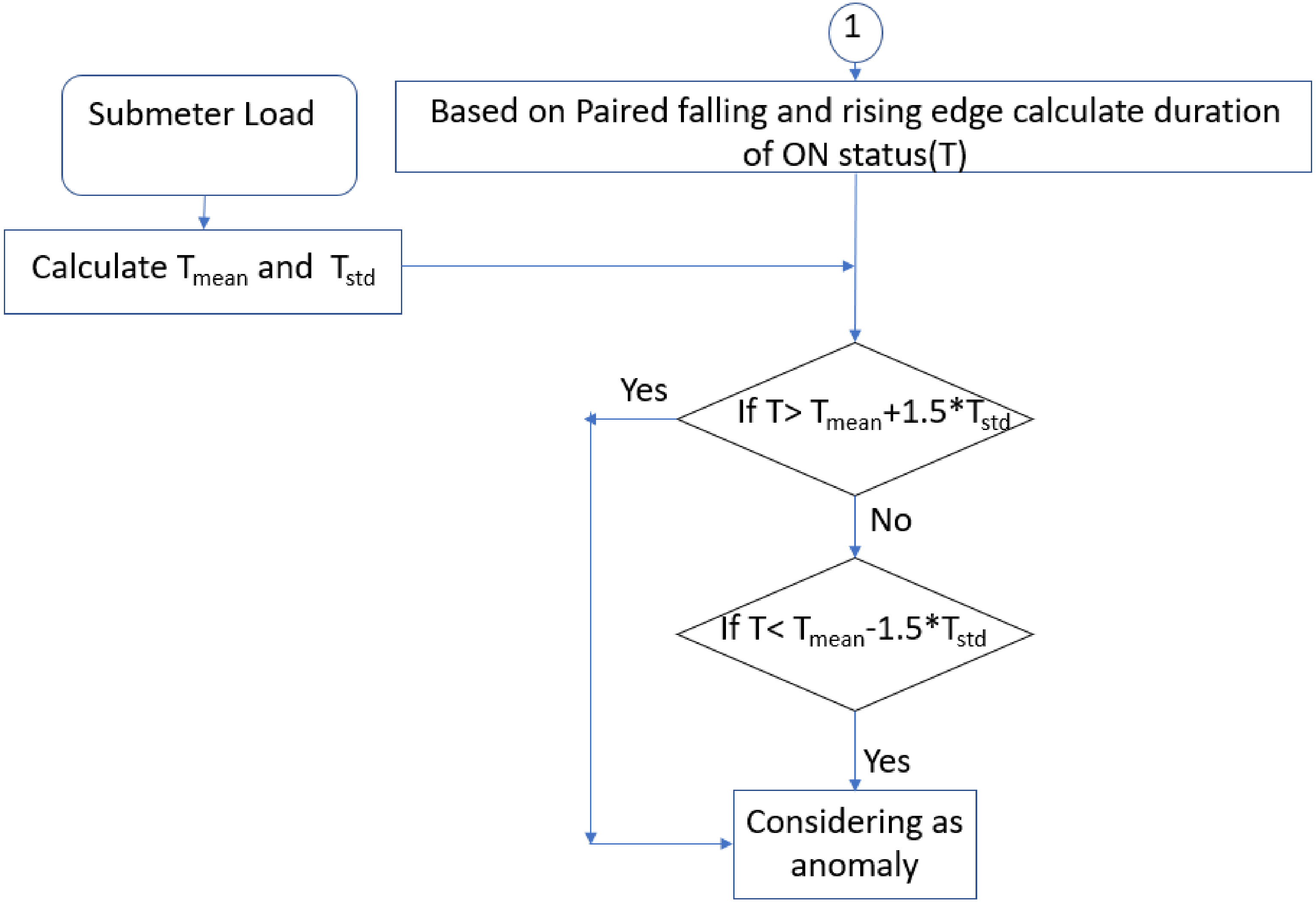
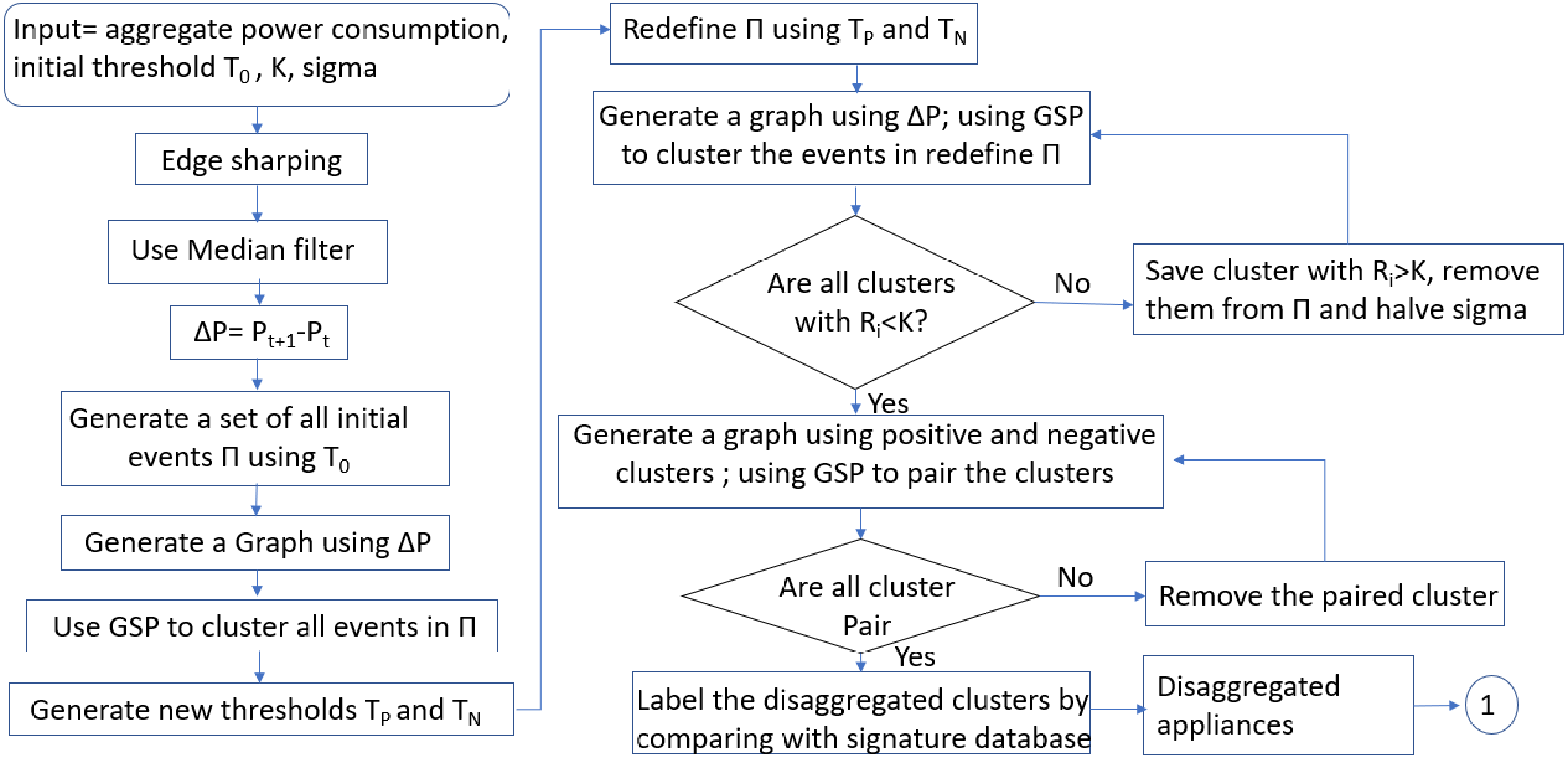

1
Submeter Load
Based on Paired falling and rising edge calculate duration of ON status(T)
Calculate T_{mean} and T_{std}
Yes
If T> T_{mean}+1.5*T_{std}
No
If T< T_{mean}-1.5*T_{std}
Yes
Considering as anomaly

Figure 5.10: Flowchart for detecting anomaly by UGSP and rule base


Input= aggregate power consumption, initial threshold T_0 , K, sigma
Edge sharping
Use Median filter
$\Delta P= P_{t+1}-P_t$
Generate a set of all initial events Π using T_0
Generate a Graph using ΔP
Use GSP to cluster all events in Π
Generate new thresholds T_P and T_N
Redefine Π using T_P and T_N
Generate a graph using ΔP; using GSP to cluster the events in redefine Π
Are all clusters with R_i<K?
No
Save cluster with R_i>K, remove them from Π and halve sigma
Yes
Generate a graph using positive and negative clusters ; using GSP to pair the clusters
Are all cluster Pair
No
Remove the paired cluster
Yes
Label the disaggregated clusters by comparing with signature database
Disaggregated appliances
1

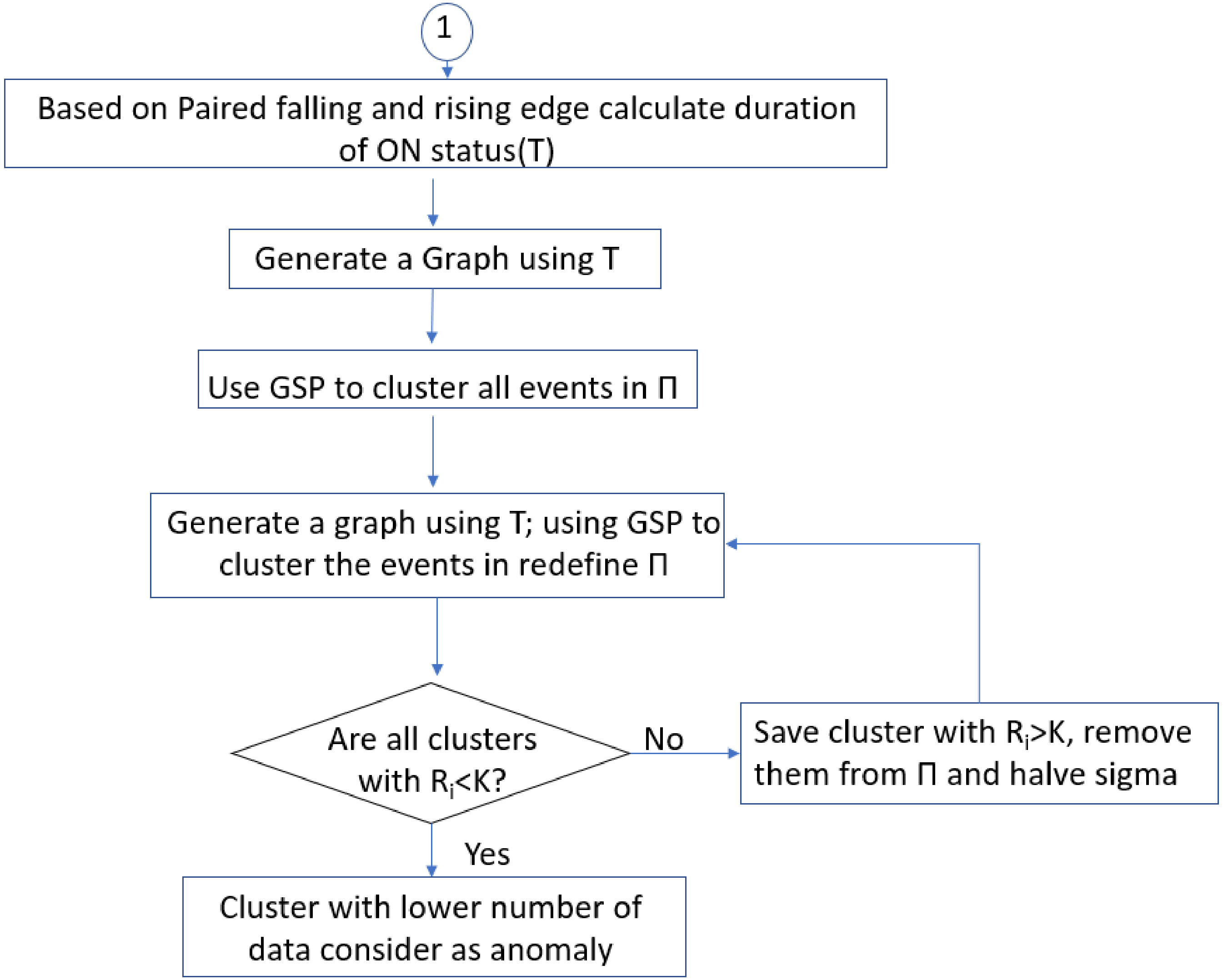

1
Based on Paired falling and rising edge calculate duration of ON status(T)
Generate a Graph using T
Use GSP to cluster all events in Π
Generate a graph using T; using GSP to cluster the events in redefine Π
Are all clusters with R_i<K?
No
Save cluster with R_i>K, remove them from Π and halve sigma
Yes
Cluster with lower number of data consider as anomaly

Figure 5.11. Flowchart for detecting anomaly by UGSP and clustering

The above algorithms were evaluated using active power readings from three houses in an open-access dataset, both down sampled to a 1-minute resolution to align with standard smart meter data intervals: Houses 2, 15 & 21 of the REFIT dataset [52]. These houses were selected due to their comprehensive submetered appliance-level data, clear labelling, and inclusion of diverse high-energy devices, making them well-suited for performance assessment. Additionally, the availability of anomaly data for these houses provides a valuable basis for validating the results.

The algorithms in figure 5.10 consists of two sequential stages: Unsupervised Graph Signal Processing (UGSP)-based Non-Intrusive Load Monitoring (NILM) for appliance disaggregation and anomaly detection.

Stage 1 — Disaggregation via Graph Signal Processing (GSP)

Pre-processing

1. Edge sharpening enhances switching transitions in P_t.

2. A median filter suppresses impulsive noise while preserving edges.

Event detection

3. Compute ΔP_t and form the initial event set $\Pi = \{ t : |\Delta P_t| \geq T_0 \}$.

Graph construction and clustering

4. Build a weighted graph $G=(V,E,W)$ with nodes $V \equiv \Pi$. Event features .

5. Apply a GSP/spectral clustering method to group events.

Threshold refinement

6. From cluster polarities, estimate T_P and T_N; redefine Π using these thresholds.

Iterative purification

7. Rebuild G and re-cluster with updated Π.

8. For each cluster compute quality R_i.

9. If any $R_i > K$: save those clusters, remove their nodes from Π, set $\sigma \leftarrow \sigma/2$, and repeat.

10. Otherwise, proceed to pairing.

Positive–negative pairing and labelling

11. Construct a bipartite match between positive and negative clusters based on magnitude balance, typical dwell, and temporal co-occurrence; remove paired clusters.

12. Compare final cluster signatures with a reference database to label appliances..

Stage 2 Anomaly Detection (3σ Rule)

ON-duration extraction

13. Pair each rising edge with its subsequent falling edge to obtain durations.

Baseline estimation

14. From a healthy window(Windows with no anomalies):

Calculate T_{mean} ,T_{std}

Rolling estimates may be used to track seasonal drift.

Decision rule

15. For each new duration T, flag an anomaly if $T > T_{mean} + 1.5 \cdot T_{std}$ or $T < T_{mean} - 1.5 \cdot T_{std}$.(formula Based on [8])

Figure 5.11 presents an algorithm whose first stage mirrors the preceding method. In Stage 2, Graph Signal Processing (GSP) is applied to cluster the ON-duration sequences of the selected appliances, and the cluster with the smallest cardinality is treated as the anomalous class.

For clarity, key parameters are defined as follows: the ON-duration threshold represents the minimum continuous active interval required to classify an appliance cycle; ΔP denotes the change in power used to detect switching events; the initial threshold T0 is used to filter out low-magnitude edges and TP and TN thresholds used for positive and negative edges calculation respectively and UGSP operates by constructing a graph Laplacian from event sequences, followed by spectral clustering to group events before post-processing. The rule-based method applies appliance-specific duration and power bounds derived from statistical profiles in REFIT.

The underlying assumption in clustering-based anomaly detection is that normal operating behaviours occur more frequently and therefore form larger, denser clusters, while abnormal or faulty behaviours are comparatively rare. Consequently, the cluster with the smallest cardinality is treated as the anomalous class, since it represents patterns that occur infrequently in the dataset. These rare instances are likely to correspond to unusual operational modes, sensor errors, or appliance faults. By identifying the least populated cluster, the method leverages the natural imbalance between normal and abnormal behaviours, allowing anomalies to be detected without requiring labelled training data. This approach is particularly useful in NILM applications, where labelled anomalies are scarce, and clustering offers a scalable way to isolate rare but potentially critical events.

The above algorithms utilize active power readings from the open-access REFIT dataset, specifically Houses 2 and 21 [52]. These houses were selected due to their comprehensive submetered appliance-level data, clear labelling, and inclusion of diverse high-energy devices—making them well-suited for evaluating unsupervised NILM algorithms under varied operational conditions. They also provide submetered data that allows performance evaluation against ground truth. Furthermore, each house contains only a single fridge-freezer; this is important, as the presence of two identical appliances with similar features would negatively impact anomaly detection, with one device's events potentially obscuring those of the other in the NILM process.

For all results presented, testing was carried out over continuous one-month periods: House 2 (1–30 June 2014 and 1–30 March 2014) and House 21 (1–31 August 2014 and 1–30 June 2014). Evaluation metrics included Precision (PR), Recall (RE), and F-Score. In each results table, the anomalies detected by the proposed algorithm are reported. Performance was also benchmarked against the reference approach described in [8], in order to highlight improvements in detection accuracy and robustness.

To improve algorithmic performance, various pre-processing methods were applied prior to NILM execution. In households with multiple fridge-freezers, disaggregation can in principle still be attempted; however, this typically requires additional post-processing to separate the highly similar load signatures. Approaches such as clustering, statistical grouping, or probabilistic assignment may be employed to distinguish overlapping ON/OFF events. While these methods can enhance disaggregation accuracy, they inevitably alter or reconstruct the original time-series behaviour of each appliance. This presents a critical limitation for anomaly detection, which depends on preserving authentic operational patterns such as cycle duration, compressor duty cycle, and power draw fluctuations. Post-processing may obscure genuine anomalies, such as prolonged ON states or short cycling, or generate artificial irregularities that appear as faults. Consequently, although the presence of multiple fridge-freezers can be addressed through post-processed disaggregation, subsequent anomaly detection becomes unreliable because the diagnostic signals of individual appliances are no longer faithfully retained. Therefore, in this study, only REFIT houses containing a single fridge-freezer were selected for analysis. This ensured that the disaggregation process could be carried out without extensive post-processing, thereby preserving the natural load signatures required for reliable anomaly detection.

Table 5.8:Anomaly detection by UGSP result for fridge-freezer REFIT house 2

Duration of investigate	Anomaly detection by UGSP and rule based	Anomaly detection by UGSP and clustering	Anomaly check by submitter data manually	Status Algorithm UGSP and rule based	Status Algorithm UGSP and clustering
1 Jun 2014 to 30 Jun 2014	02/06/2024 14:31-16:06	02/06/2024 14:31	02/06/2024 14:31-16:58	TP	TP
	02/06/2024 17:47-20:25	02/06/2024 17:47	02/06/2024 17:47-20:25	TP	TP
	03/06/2024 12:50-13:46	03/06/2024 12:50	03/06/2024 12:50-13:46	TP	TP
	04/06/2024 10:54-12:48	04/06/2024 10:54	04/06/2024 11:22-12:48	TP	TP
	06/06/2024 12:19-13:51	06/06/2024 12:19	06/06/2024 12:19-13:51	TP	TP
	06/06/2024 14:43-15:37	-	06/06/2024 14:43-15:37	TP	FN
	06/06/2024 16:25-17:32	06/06/2024 16:25	06/06/2024 16:25-17:32	TP	TP
	06/06/2024 18:12-19:06	-	06/06/2024 18:12-19:06	TP	FN
	06/06/2024 23:49-00:46	06/06/2024 23:49	06/06/2024 23:49-00:46	TP	TP
	07/06/2024 04:53-05:57	07/06/2024 04:53	07/06/2024 04:53-05:57	TP	TP
	08/06/2024 10:31-11:35	08/06/2024 10:31	08/06/2024 10:31-11:35	TP	TP
	09/06/2024 11:55-12:58	09/06/2024 11:55	09/06/2024 11:55-12:58	TP	TP
	10/06/2024 19:59-21:23	10/06/2024 19:59	10/06/2024 19:59-21:23	TP	TP
	11/06/2024 11:17-12:13	-	11/06/2024 11:17-12:13	TP	FN
	12/06/2024 09:58-10:55	12/06/2024 09:58	12/06/2024 09:58-10:55	TP	TP
	14/06/2024 10:18-11:27	14/06/2024 10:18-11:27	14/06/2024 10:18	TP	TP
	15/06/2024 08:58-09:49	-	15/06/2024 08:58-09:49	TP	FN
	16/06/2024 11:26-12:26	16/06/2024 11:26	16/06/2024 11:26-12:26	TP	TP
	16/06/2024 15:56-17:01	16/06/2024 15:56	16/06/2024 15:56-17:01	TP	TP
	17/06/2024 14:29-15:34	17/06/2024 14:29	17/06/2024 14:29-15:34	TP	TP
	17/06/2024 16:47-17:47	17/06/2024 16:47	17/06/2024 16:47-17:47	TP	TP

	18/06/2024 08:30-09:23	-	17/06/2024 08:30-09:23	TP	FN
	19/06/2024 13:39-14:48	19/06/2024 13:39	19/06/2024 13:39-14:48	TP	TP
	20/06/2024 14:31-15:33	20/06/2024 14:31	20/06/2024 14:31-15:33	TP	TP
	21/06/2024 14:15-15:09	-	21/06/2024 14:15-15:09	TP	FN
	21/06/2024 15:41-17:06	21/06/2024 15:41	21/06/2024 15:41-17:06	TP	TP
	22/06/2024 18:12-19:18	22/06/2024 18:12	22/06/2024 18:12-19:18	TP	TP
	23/06/2024 08:17-09:15	23/06/2024 08:17	23/06/2024 08:17-09:15	TP	TP
	23/06/2024 13:48-14:42		23/06/2024 13:48-14:42	TP	FN
	23/06/2024 16:59-18:00	23/06/2024 16:59	23/06/2024 16:59-18:00	TP	TP
	24/06/2024 02:26-18:00	24/06/2024 02:26	24/06/2024 02:26-18:00	TP	TP
	24/06/2024 18:57-19:51	-	24/06/2024 18:57-19:51	TP	FN
	24/06/2024 22:46-00:10	24/06/2024 22:46	24/06/2024 22:46-00:10	TP	TP
	25/06/2024 01:15-02:08	-	25/06/2024 01:15-02:08	TP	FN
	25/06/2024 12:20-18:20	25/06/2024 12:20	25/06/2024 12:20-18:20	TP	TP
	26/06/2024 06:35-07:48	25/06/2024 06:35	25/06/2024 06:35-07:48	TP	TP
	26/06/2014 16:33-17:59	26/06/2014 16:33	26/06/2014 16:33-17:59	TP	TP
	27/06/2014 16:14-17:16	27/06/2014 16:14	27/06/2014 16:14-17:16	TP	TP
	30/06/2014 19:55-20:46	-	30/06/2014 19:55-20:46	TP	FN

Based on submetered data, the number of anomalies with ON duration greater than 2957 s is 78. In comparison, reference [8] detected only 2 anomalies during the same period on submetered data.

Table 5.9:Anomaly detection performance , fridge-freezer, house 2 REFIT data, June 2024

	TP	FN	FP	Precision	Recall	F-Score
Anomaly detection by UGSP and rule based	39	39	0	1	0.5	0.66
Anomaly detection by UGSP and clustering	29	49	0	1	0.37	0.54

Table 5.8 presents the anomaly detection performance for the fridge-freezer in REFIT House 2 over the period 1–30 June 2014, using UGSP-based disaggregation combined with two different post-processing approaches: a rule-based method and a clustering method. An anomaly threshold was set at an ON duration greater than 2957 s (calculated as > T_{mean} + $1.5 \cdot T_{std}$ =2120+1.5×558T). Based on the submetered ground-truth data, a total of 78 anomalies were identified during this period. By contrast, reference [8] reported only two anomalies for the same dataset, suggesting that their method applied a more restrictive definition of anomalies or used different thresholding criteria.

The UGSP with rule-based anomaly detection identified 39 anomalies (true positives) and missed 39 (false negatives), achieving a Precision of 1.0, Recall of 0.5, and an F-Score of 0.66. In comparison, the UGSP with clustering detected 29 anomalies and missed 49, with a Precision of 1.0, Recall of 0.37, and an F-Score of 0.54. Both approaches produced no false positives, indicating that all detected events were valid anomalies, but differed in sensitivity. The rule-based approach demonstrated higher recall and overall performance compared to the clustering approach. Clustering-based anomaly detection has the advantage that it does not require additional prior knowledge or labelled data from the appliance load. Instead, it groups cycles based on similarity in duration, power, or shape, and identifies outliers as anomalies. This makes clustering particularly suitable for offline analysis, where large batches of data can be processed retrospectively to detect unusual events. However, the need for batch processing and the computational cost of clustering limit its suitability for real-time or embedded applications. In contrast, rule-based methods operate on simple thresholds and can be applied directly during NILM disaggregation. They are lightweight, interpretable, and capable of detecting anomalies in real-time, which makes them well suited for online anomaly detection in practical household monitoring systems. The trade-off, however, is that rule-based methods may be less adaptive to variations in appliance behaviour compared to clustering, and can miss anomalies that do not fit the predefined rule set.

Table 5.10:Anomaly detection by UGSP result for fridge-freezer REFIT house 2

Duration of investigate	Anomaly detection by UGSP and rule based	Anomaly detection by UGSP and clustering	Anomaly check by submitter data manually	Status Algorithm UGSP and rule based	Status Algorithm UGSP and clustering
1 March 2014 to 30 March 2014	2/03/2014 4:29-5:13	2/03/2014 4:29	2/03/2014 4:29-5:13	TP	TP
	3/03/2014 10:21-11:36	3/03/2014 10:21	3/03/2014 10:52-11:36	TP	TP
	8/03/2014 00:48-1:29	-	8/03/2014 00:48-1:29	TP	FN
	8/03/2014 23:14-00:00	8/03/2014 23:14	8/03/2014 23:14-00:01	TP	FN
	9/03/2014 21:20-22-31	-	9/03/2014 21:54-23-38	TP	FN
	10/03/2014 1:03-1:59	10/03/2014 1:03-1:59	10/03/2014 1:03-1:59	TP	TP
	10/03/2014 5:50-6:37	10/03/2014 5:50	10/03/2014 5:50-6:37	TP	TP
	10/03/2014 9:36-10:27	10/03/2014 9:36	10/03/2014 9:36-10:27	TP	TP
	10/03/2014 11:29-12:16	10/03/2014 11:29	10/03/2014 11:29-12:16	TP	TP
	13/03/2014 5:10-5:52	-	13/03/2014 5:10-5:52	TP	FN
	14/03/2014 12:20-13:01	-	14/03/2014 12:20-13:01	TP	FN
	15/03/2014 2:49-3:41	15/03/2014 2:49	15/03/2014 2:49-3:41	TP	TP
	17/03/2014 22:23-2317	17/03/2014 22:23	17/03/2014 22:23-2317	TP	TP
	18/03/2014 9:35-10:16		18/03/2014 9:35-10:16	TP	FN
	18/03/2014 13:55-15:33	18/03/2014 13:55-15:33	18/03/2014 13:55-15:33	TP	TP
	18/03/2014 19:39-20:20	-	18/03/2014 19:39-20:20	TP	FN
	19/03/2014 12:44-15:30	19/03/2014 12:44-15:30	19/03/2014 12:44-16:30	TP	TP

	20/03/2014 10:18-10:59	-	20/03/2014 10:18-10:59	TP	FN
	20/03/2014 11:18-12:22	20/03/2014 11:18-12:22	20/03/2014 11:18-12:22	TP	TP
	21/03/2014 4:32-5:20	21/03/2014 4:32	21/03/2014 4:32-5:20	TP	TP
	23/03/2014 4:22-5:26	-	23/03/2014 4:43-5:26	TP	FN
	24/03/2014 9:25-10:6	-	24/03/2014 9:25-10:6	TP	FN
	29/03/2014 2:14-3:02	29/03/2014 2:14	29/03/2014 2:14-3:02	TP	TP

Based on submetered data, the number of anomalies with ON duration greater than 2387 s is 67. In comparison, reference [8] detected only 2 anomalies during the same period on submeter data. The parameters used for disaggregation with UGSP were set as Sigma = 250 and Ri = 0.04. For the fridge-freezer, the maximum ON-duration threshold was determined based on the statistical values T_{mean} = 1679 s and T_{std} = 472 s, yielding a cutoff of 2387 s (i.e., $T_{mean} + 1.5 \times T_{std}$). Any ON-duration exceeding this threshold was considered anomalous.

Table 5.11:Anomaly detection performance , fridge-freezer, house 2 REFIT data, June 2024

	TP	FN	FP	Precision	Recall	F-Score
Anomaly detection by UGSP and rule based	23	44	0	1	0.34	0.51
Anomaly detection by UGSP and clustering	13	54	0	1	0.19	0.27

Table 5.10 summarises the anomaly detection results for the fridge-freezer in REFIT House 2 over the period 1–30 March 2014. An anomaly threshold was set at an ON duration greater than 2387. Based on the submetered ground-truth data, a total of 67 anomalies were identified during this period, whereas reference [8] reported only two anomalies for the same dataset, again highlighting the sensitivity of anomaly detection outcomes to the chosen thresholding approach.

As shown in Table 5.11, the UGSP with rule-based post-processing detected 23 anomalies (true positives) and missed 44 (false negatives), yielding a Precision of 1.0, Recall of 0.34, and F-Score of 0.51. In contrast, the UGSP with clustering detected only 13 anomalies and missed 54, with a Precision of 1.0, Recall of 0.19, and F-Score of 0.27. Both methods produced no false positives, confirming that the anomalies flagged were valid, but their sensitivity varied considerably. The rule-based method again outperformed clustering, although recall remained relatively low compared to the submetered baseline.

Table 5.12: Anomaly detection by UGSP result for fridge-freezer REFIT house 21

Duration of investigate	Anomaly detection by UGSP and rule based	Anomaly detection by UGSP and clustering	Anomaly check by submitter data manually	Status Algorithm UGSP and rule based	Status Algorithm UGSP and clustering
1 Aug 2014 to 31 Aug 2014	05/08/2024 17:54-18:56	-	05/08/2024 17:54-18:56	TP	FN
	06/08/2024 18:02-19:15	06/08/2024 18:02	06/08/2024 18:02-19:15	TP	TP

	07/08/2024 17:59-19:04	-	07/08/2024 17:59-19:04	TP	FN
	09/08/2024 18:47-19:42	-	09/08/2024 18:47-19:42	TP	FN
	09/08/2024 20:36-21:32	-	09/08/2024 20:36-21:32	TP	FN
	12/08/2024 17:35-19:12	12/08/2024 17:35	12/08/2024 17:35-19:12	TP	TP
	14/08/2024 17:59-18:57	-	14/08/2024 17:59-18:57	TP	FN
	15/08/2024 18:01-19:08	-	15/08/2024 18:01-19:08	TP	FN
	16/08/2024 02:24-03:30	-	16/08/2024 02:24-03:30	TP	FN
	17/08/2024 01:44-03:37	17/08/2024 01:44	17/08/2024 01:44-03:37	TP	TP
	30/08/2024 11:58-01/09/2024 8:35	30/08/2024 11:58	30/08/2024 11:58-01/09/2024 8:35	TP	TP

Based on submetered data, the number of anomalies with ON duration greater than 3243 s is 48. In comparison, reference [8] detected only 3 anomalies during the same period. The parameters used for disaggregation with UGSP were set as Sigma = 250 and Ri = 0.04. For the fridge-freezer, the maximum ON-duration threshold was determined based on the statistical values T_{mean} = 2571 s and T_{std} = 448 s, yielding a cutoff of 3243 s (i.e., T_{mean} + 1.5 × T_{std}). Any ON-duration exceeding this threshold was considered anomalous.

Table 5.13:Anomaly detection performance , fridge-freezer, house 21 REFIT data, Aug 2024

	TP	FN	FP	Precision	Recall	F-Score
Anomaly detection by UGSP and rule based	11	37	0	1	0.23	0.37
Anomaly detection by UGSP and clustering	4	44	0	1	.08	0.15

Table 5.12 summarises the anomaly detection results for the fridge-freezer in REFIT House 21 during the period 1–31 August 2014. The anomaly threshold was defined as an ON duration greater than 3243s.

Submetered ground-truth data indicated 48 anomalies over this period, while reference [8] reported only three anomalies for the same dataset, again illustrating the strong dependence of anomaly counts on thresholding criteria and methodology.

As shown in Table 5.13, the UGSP with rule-based post-processing detected 11 anomalies and missed 37, resulting in a Precision of 1.0, Recall of 0.23, and F-Score of 0.37. By comparison, the UGSP with clustering identified only 4 anomalies and missed 44, with a Precision of 1.0, Recall of 0.08, and F-Score of 0.15. As in previous cases, both methods produced no false positives, but recall was notably lower in House 21 than in House 2, reflecting greater difficulty in detecting long-duration anomalies in this dataset.

These findings reinforce the earlier trend: the rule-based approach consistently outperforms clustering in terms of recall and F-Score, making it more reliable for online anomaly detection. However, overall performance was weaker in House 21 compared to House 2, which may be attributed to variations in fridge-freezer usage patterns, anomaly distribution, and signal-to-noise conditions in the aggregate load.

Table 5.14: Anomaly detection by UGSP result for fridge-freezer REFIT house 21

Duration of investigate	Anomaly detection by UGSP and rule based	Anomaly detection by UGSP and clustering	Anomaly check by submitter data manually	Status Algorithm UGSP and rule based	Status Algorithm UGSP and clustering
1 Jun 2014 to 31 Jun 2014	01/06/2024 07:23-09:20	01/06/2024 07:23	01/06/2024 07:23-09:20	TP	TP
	01/06/2024 20:48-21:43	-	01/06/2024 20:48-21:43	TP	FN
	01/06/2024 22:49-23:45	-	01/06/2024 22:49-23:45	TP	FN
	02/06/2024 04:44-06:52	02/06/2024 04:44-	02/06/2024 04:44-06:52	TP	TP
	02/06/2024 18:19-20:35	02/06/2024 18:19	02/06/2024 18:19-20:35	TP	TP
	03/06/2024 15:20-16:45	03/06/2024 15:20	03/06/2024 15:20-16:45	TP	TP
	03/06/2024 22:17-00:00	03/06/2024 22:17	03/06/2024 22:17-00:00	TP	TP
	04/06/2024 10:59-11:56	-	04/06/2024 10:59-11:56	TP	FN
	04/06/2024 23:22-00:30	-	04/06/2024 23:22-00:30	TP	FN
	06/06/2024 03:01-03:57	-	06/06/2024 03:01-03:57	TP	FN
	06/06/2024 16:10-17:31	06/06/2024 16:10	06/06/2024 16:10-17:31	TP	TP
	07/06/2024 18:43-19:58	-	07/06/2024 18:43-19:58	TP	FN
	08/06/2014 21:07-22:03	-	08/06/2014 21:07-22:03	TP	FN
	09/06/2014 20:50-21:46	-	09/06/2014 20:50-21:46	TP	FN
	13/06/2014 05:54-06:59	-	13/06/2014 05:54-06:59	TP	FN
	14/06/2014 05:14-07:20	14/06/2014 05:14	14/06/2014 05:14-07:20	TP	TP
	14/06/2014 19:46-20:41	-	14/06/2014 19:46-20:41	TP	FN
	15/06/2014 02:04-03:40	15/06/2014 02:04	15/06/2014 02:04-03:40	TP	FN
	16/06/2014 00:13-01:25	-	16/06/2014 00:13-01:25	TP	FN
	16/06/2014 03:39-04:42	-	16/06/2014 03:39-04:42	TP	FN
	16/06/2014 03:39-04:42	-	16/06/2014 03:39-04:42	TP	FN
	17/06/2014 21:19-22:21	-	17/06/2014 21:19-22:21	TP	FN
	17/06/2014 23:44-00:44	-	17/06/2014 23:44-00:44	TP	FN
	18/06/2014 17:40-18:59	18/06/2014 17:40	18/06/2014 17:40-18:59	TP	TP
	18/06/2014 22:01-22:59	-	18/06/2014 22:01-22:59	TP	FN
	18/06/2014 23:58-00:55	-	18/06/2014 23:58-00:55	TP	FN
	19/06/2014 02:00-02:55	-	19/06/2014 02:00-02:55	TP	FN
	19/06/2014 04:04-04:59	-	19/06/2014 04:04-04:59	TP	FN

	19/06/2014 17:06-18:53	19/06/2014 17:06	19/06/2014 17:06-18:53	TP	TP
	19/06/2014 16:03-17:42	19/06/2014 16:03	19/06/2014 16:03-17:42	TP	TP
	22/06/2014 17:16-19:03	22/06/2014 17:16	22/06/2014 17:16-19:03	TP	TP
	23/06/2014 14:59-17:54	23/06/2014 14:59	23/06/2014 14:59-17:54	TP	TP
	23/06/2014 18:08-20:02	23/06/2014 18:08	23/06/2014 18:08-20:02	TP	TP
	25/06/2014 00:33-01:28	-	25/06/2014 00:33-01:28	TP	FN
	28/06/2014 06:40-09:13	28/06/2014 06:40	28/06/2014 06:40-09:13	TP	TP

Based on submetered data, the number of anomalies with ON duration greater than 3508 s is 75. In comparison, reference [8] detected only 2 anomalies during the same period. The parameters used for disaggregation with UGSP were set as Sigma = 250 and Ri = 0.04. For the fridge-freezer, the maximum ON-duration threshold was determined based on the statistical values T_{mean} = 2668 s and T_{std} = 560 s, yielding a cutoff of 3508 s (i.e., T_{mean} + 1.5 × T_{std}). Any ON-duration exceeding this threshold was considered anomalous.

Table 5.15:Anomaly detection performance , fridge-freezer, house 21 REFIT data, Jun 2024

	TP	FN	FP	Precision	Recall	F-Score
Anomaly detection by UGSP and rule based	36	39	0	1	0.48	0.65
Anomaly detection by UGSP and clustering	15	60	0	1	0.2	0.33

Table 5.14 presents the anomaly detection results for the fridge-freezer in REFIT House 21 during the period 1–30 June 2014. The anomaly threshold was defined as an ON duration greater than 3508

Submetered ground-truth data indicated 75 anomalies in this period, whereas reference [8] identified only two anomalies for the same dataset. This discrepancy again underlines how strongly anomaly counts are influenced by the choice of thresholding and methodological definitions.

As reported in Table 5.15, the UGSP with rule-based anomaly detection achieved 36 true positives and missed 39 anomalies, yielding a Precision of 1.0, Recall of 0.48, and F-Score of 0.65. In contrast, the UGSP with clustering detected 15 anomalies and missed 60, giving a Precision of 1.0, Recall of 0.20, and F-Score of 0.33. Both approaches produced no false positives, but the rule-based method again outperformed clustering in terms of recall and F-Score.

Table 5.16: Anomaly detection by UGSP result for fridge-freezer REFIT house 15

Duration of investigate	Anomaly detection by UGSP and rule based	Anomaly detection by UGSP and clustering	Anomaly check by submitter data manually	Anomaly based on Haroon Paper	Status Algorithm UGSP and rule based	Status Algorithm UGSP and clustering
1 March 2014 to 31 March 2014	01/03/2014 6:12-7:52	01/03/2014 6:12	01/03/2014 6:12-8:28	-	TP	TP
	01/03/2014 10:56-11:34	-	-	-	FP	TN
	03/03/2014 06:03-06:39	-	03/03/2014 06:00--06:39	-	TP	FN
	05/03/2014 06:23-07:30	05/03/2014 06:23	05/03/2014 06:23-07:30	-	TP	TP
	06/03/2014 09:26-10:13	-	-	-	FP	TN
	14/03/2014 20:19-20:55	-	14/03/2014 20:19-20:55	-	TP	FN
	18/03/2014 14:48-15:25	-	18/03/2014 14:48-15:25	-	TP	FN
	18/03/2014 21:33-22:11	-	18/03/2014 21:33-22:11	-	TP	FN
	19/03/2014 20:23-20:59	-	19/03/2014 20:23-20:59	-	TP	FN
	21/03/2014 19:48-20:49	21/03/2014 19:48	21/03/2014 19:48-20:49	-	TP	TP
	22/03/2014 19:05-21:20	22/03/2014 19:05	22/03/2014 19:05-22:43	-	TP	TP
	22/03/2014 20:29-21:20	-	22/03/2014 20:29-21:20	-	TP	FN
	26/03/2014 18:42-19:58	26/03/2014 18:42	26/03/2014 18:42-21:04	-	TP	TP
	27/03/2014 06:46-07:37	-	-	-	FP	TN
	30/03/2014 19:45-20:31	-	30/03/2014 19:45-20:31	-	TP	FN
	30/03/2014 20:40-21:19	-	30/03/2014 20:40-21:19	-	TP	FN
	30/03/2014 19:28-20:08	-	30/03/2014 19:28-20:08	-	TP	FN

Based on submetered data, the number of anomalies with ON duration greater than 2037 s is 62. In comparison, reference [8] detected no anomalies during the same period. The parameters used for disaggregation with UGSP were set as Sigma = 250 and Ri = 0.04. For the fridge-freezer, the maximum ON-duration threshold was determined based on the statistical values T_{mean} = 1503 s and T_{std} = 380 s, yielding a cutoff of 2037 s (i.e., $T_{mean} + 1.5 \times T_{std}$). Any ON-duration exceeding this threshold was considered anomalous.

Table 5.17:Anomaly detection performance , fridge-freezer, house 15 REFIT data, March 2024

	TP	FN	FP	Precision	Recall	F-Score
Anomaly detection by UGSP and rule based	14	48	3	0.82	0.23	0.36
Anomaly detection by UGSP and clustering	5	57	0	0.23	0.08	0.12

Table 5.16 shows the anomaly detection results for the fridge-freezer in REFIT House 15 over the period 1–31 March 2014. The anomaly threshold was defined as an ON duration greater than 2037 s. Submetered ground-truth data indicated 62 anomalies during this period, whereas reference [8] reported none, reflecting significant methodological differences in anomaly definition.

As presented in Table 5.17, the UGSP with rule-based anomaly detection achieved 14 true positives and missed 48 anomalies, with 3 false positives. This yielded a Precision of 0.82, Recall of 0.23, and an F-Score of 0.36. In contrast, the UGSP with clustering detected only 5 anomalies and missed 57, with no false positives but a much lower Precision of 0.23, Recall of 0.08, and F-Score of 0.12.

Table 5.18: Anomaly detection by UGSP result for fridge-freezer REFIT house 15

Duration of investigate	Anomaly detection by UGSP and rule based	Anomaly detection by UGSP and clustering	Anomaly check by submitter data manually	Anomaly based on Haroon Paper	Status Algorithm UGSP and rule based	Status Algorithm UGSP and clustering
1 June 2014 to 31 June 2014	05/06/2014 12:17-13:22	05/06/2014 12:17	-	-	FP	FP
	05/06/2014 20:24-21:00	-	-	-	FP	TN
	06/06/2014 03:28-03:56	-	06/06/2014 03:28-03:56	-	TP	FP
	06/06/2014 05:23-05:50	-	06/06/2014 05:23-05:50	-	TP	FP
	06/06/2014 21:15-21:53	-	-	-	FP	TN
	08/06/2014 20:15-21:10	08/06/2014 20:15	08/06/2014 20:15-21:10	-	TP	TP
	08/06/2014 22:49-23:45	08/06/2014 22:49	08/06/2014 22:49-23:45	-	TP	TP
	10/06/2014 06:12-06:40	-	-	-	FP	TN
	10/06/2014 17:55-18:25	-	-	-	FP	TN
	10/06/2014 05:38-06:39	10/06/2014 05:38	10/06/2014 05:17-07:11	-	TP	TP
	15/06/2014 17:21-18:03	-	-	-	FP	TN
	15/06/2014 21:07-21:50	-	-	-	FP	TN
	16/06/2014 05:04-06:00	16/06/2014 05:04	16/06/2014 05:04-06:00	-	TP	TP
	17/06/2014 06:03-06:38		17/06/2014 05:39-06:07	-	TP	FN
	18/06/2014 04:00-04:27	-	18/06/2014 04:00-04:27	-	TP	FN
	19/06/2014 20:58-22:01	19/06/2014 20:58	19/06/2014 20:58-22:01	-	TP	TP
	21/06/2014 17:35-18:02	-	21/06/2014 17:35-18:02	-	TP	FN
	22/06/2014 19:43-20:11	-	22/06/2014 19:43-20:11	-	TP	FN
	23/06/2014 06:11-06:43	-	23/06/2014 06:11-06:43	-	TP	FN
	24/06/2014 06:12-06:42	-	24/06/2014 06:12-06:42	-	TP	FN
	24/06/2014 17:39-19:02	-	-	-	FP	FP
	25/06/2014 05:47-06:36	-	-	-	FP	FP
	26/06/2014 17:54-18:23	-	26/06/2014 17:54-18:23	-	TP	FN
	30/06/2014 07:39-08:06	-	30/06/2014 07:39-08:06	-	TP	FN

Based on submetered data, the number of anomalies with ON duration greater than 2037 s is 36. In comparison, reference [8] detected no anomalies during the same period. The parameters used for disaggregation with UGSP were set as Sigma = 250 and Ri = 0.04. For the fridge-freezer, the maximum ON-duration threshold was determined based on the statistical values T_{mean} = 1378 s and T_{std} = 144 s, yielding a cutoff of 1594 s (i.e., T_{mean} + 1.5 × T_{std}). Any ON-duration exceeding this threshold was considered anomalous.

Table 5.19: Anomaly detection performance , fridge-freezer, house 15 REFIT data, June 2024

	TP	FN	FP	Precision	Recall	F-Score
Anomaly detection by UGSP and rule based	15	21	9	0.63	0.41	0.5
Anomaly detection by UGSP and clustering	5	31	5	0.5	0.14	0.21

Table 5.18 presents the anomaly detection results for the fridge-freezer in REFIT House 15 during the period 1–30 June 2014. The anomaly threshold was defined as an ON duration greater than 1594 s . Ground-truth submetered data indicated 36 anomalies for this period, whereas reference [8] reported none, again demonstrating that anomaly definition is highly sensitive to the chosen threshold and methodology.

As summarised in Table 5.19, the UGSP with rule-based post-processing detected 15 anomalies and missed 21, with 9 false positives. This resulted in a Precision of 0.63, Recall of 0.41, and an F-Score of 0.50. In comparison, the UGSP with clustering detected only 5 anomalies, missed 31, and produced 5 false positives, yielding a Precision of 0.50, Recall of 0.14, and an F-Score of 0.21.

Table 5.20: Consolidated anomaly detection performance across REFIT Houses 2, 21, and 15 (different months)

House	Period	Method	TP	FN	FP	Precision	Recall	F-Score
2	Mar-14	UGSP + Rule-based	23	44	0	1	0.34	0.51
2	Jun-14	UGSP + Rule-based	39	39	0	1	0.5	0.66
21	Jun-14	UGSP + Rule-based	36	39	0	1	0.48	0.65
21	Aug-14	UGSP + Rule-based	11	37	0	1	0.23	0.37
15	Mar-14	UGSP + Rule-based	14	48	3	0.82	0.23	0.36
15	Jun-14	UGSP + Rule-based	15	21	9	0.63	0.41	0.5
2	Mar-14	UGSP + Clustering	13	54	0	1	0.19	0.27
2	Jun-14	UGSP + Clustering	29	49	0	1	0.37	0.54
21	Jun-14	UGSP + Clustering	15	60	0	1	0.2	0.33
21	Aug-14	UGSP + Clustering	4	44	0	1	0.08	0.15
15	Mar-14	UGSP + Clustering	5	57	0	0.23	0.08	0.12
15	Jun-14	UGSP + Clustering	5	31	5	0.5	0.14	0.21

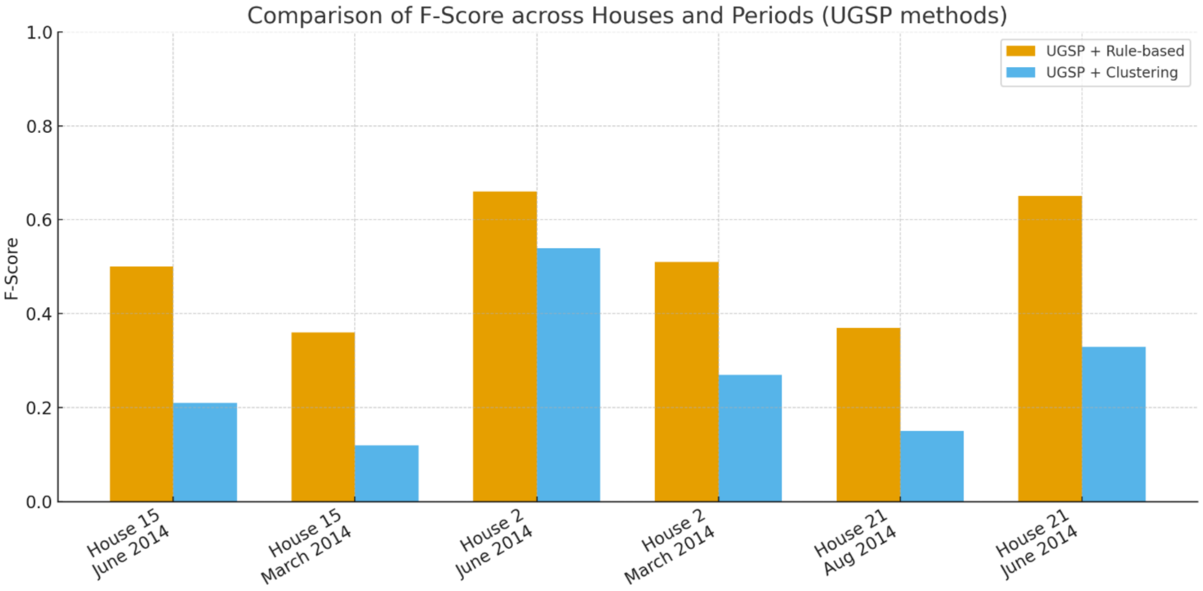


Figure 5.12: summary bar chart showing the F-Score performance of UGSP + Rule-based vs. UGSP + Clustering across all test cases (Houses 2, 21, 15 over March, June, and August).

Table 5.20 summarises the anomaly detection performance across all case studies, covering REFIT Houses 2, 21, and 15 over different months. The results show clear variation in detection accuracy depending on both the chosen post-processing method and the dataset period. UGSP combined with rule-based thresholds consistently achieved higher recall and F-Score compared to UGSP with clustering, although some cases (e.g., House 15, June 2014) revealed non-negligible false positives. In contrast, clustering produced very few false positives but suffered from low recall, resulting in poor overall F-Scores across all households.

Figure 5.12 provides a visual summary of F-Score performance across the test cases. It highlights that the rule-based approach outperformed clustering in every dataset, with House 2 (June 2014) showing the strongest performance (F-Score = 0.66), and House 21 (August 2014) and House 15 (March 2014) showing the weakest (F-Scores of 0.37 and 0.36, respectively).

5.5 Summary

This chapter proposes anomaly detection for fridge-freezers using unsupervised Graph Signal Processing (UGSP) applied to the REFIT dataset. The study employed two post-processing strategies—rule-based thresholding and clustering—to identify anomalies defined as unusually long ON-durations. The results across multiple households and time periods revealed several important insights.

First, the choice of post-processing method significantly affects detection performance. The rule-based approach consistently achieved higher recall and F-Score compared to clustering, making it more suitable for online, real-time monitoring applications. However, in certain cases (e.g., House 15), rule-based detection introduced false positives, highlighting the sensitivity of fixed thresholds to variations in appliance usage. Clustering, on the other hand, produced fewer false positives but suffered from very low recall, limiting its applicability where timely fault detection is required.

Second, performance was strongly dataset-dependent. Detection accuracy varied across households and time periods, with House 2 (June 2014) yielding the highest F-Score (0.66) and House 21 (August 2014) and House 15 (March 2014) showing the weakest performance. These variations illustrate the impact of household-specific appliance behaviour, usage patterns, and signal-to-noise conditions on NILM-based anomaly detection.

Finally, the anomaly counts obtained in this study were substantially higher than those reported in reference [8], where far fewer anomalies were identified under the same conditions. This discrepancy underscores the importance of carefully defining anomaly thresholds and evaluation metrics in order to ensure fair comparison across different methods.

Chapter 6

5 Conclusions and Future Work

6.1 Conclusions

This thesis has investigated the critical role of feature selection, classification, and anomaly detection in advancing the performance and applicability of Non-Intrusive Load Monitoring (NILM). Through a combination of supervised, unsupervised, and rule-based approaches, the research systematically explored the impact of signal pre-processing, model selection, and post-processing techniques on NILM accuracy, generalizability, and operational robustness. The findings collectively address four central research questions, each contributing to a comprehensive understanding of how NILM can be optimized for real-world energy monitoring and appliance-level analytics.

The first research question examined how effective feature selection can enhance NILM performance while reducing computational cost. As demonstrated in Chapter 3, feature selection emerged as a decisive factor in both photonic sensing applications and smart meter–based NILM. Using Artificial Neural Networks (ANN) as a benchmark model, the study showed that selecting a minimal yet discriminative set of features — such as ΔP, Δt, steady-state power levels, and transition signatures — leads to significant improvements in classification accuracy and generalization capability. Moreover, feature dimensionality reduction directly decreases model complexity and training time, enabling faster inference and making NILM more suitable for real-time and embedded applications, where computational efficiency is essential. The outcomes affirm that a well-engineered feature space not only facilitates model interpretability but also enhances resilience against noise and non-stationary load conditions.

The second research question focused on identifying NILM algorithms most suitable for near real-time disaggregation. As presented in Chapter 4, this work compared a range of supervised and unsupervised learning techniques — including Decision Trees (DT), K-Nearest Neighbours (KNN), K-Means, DBSCAN, and Graph Signal Processing (GSP/UGSP). The comparative analysis revealed that supervised algorithms generally deliver higher disaggregation accuracy due to their ability to exploit labelled training data and learn distinct appliance signatures. However, unsupervised techniques demonstrated greater scalability and independence from labelled datasets, making them attractive for large-scale or newly instrumented households. Furthermore, the introduction of a two-stage grouping and disaggregation framework significantly improved performance for appliances with overlapping power signatures, such as dishwashers and washing machines. This hierarchical approach allowed the system to first cluster appliances with similar characteristics and then perform a refined classification within each subgroup, improving accuracy while maintaining computational efficiency.

The third research question investigated the transferability of NILM models across datasets and households. As discussed in Chapter 4, transfer learning experiments using data from REFIT Houses 2 and 17 revealed that while models trained on one household can retain partial generalization capabilities, performance often degrades when exposed to unseen environments. This degradation arises from variations in appliance types, user behaviour, load characteristics, and background noise. For example, while relatively stable appliances like kettles and fridge-freezers showed moderate transferability, complex or behaviourally driven appliances such as washing machines and microwaves exhibited pronounced declines in classification accuracy. These findings underscore the limitations of direct model transfer in NILM and highlight the need for domain adaptation, calibration strategies, and context-aware modelling to improve cross-household generalization. Incorporating additional metadata — such as occupancy patterns, time-of-use information, and appliance condition indicators — may further enhance model adaptability and robustness in heterogeneous environments.

The fourth research question addressed appliance-level anomaly detection from aggregate smart meter data, representing one of the thesis's most applied and safety-oriented contributions. Chapter

5 demonstrated that integrating Unsupervised Graph Signal Processing (UGSP) with rule-based anomaly detection frameworks enables effective identification of abnormal appliance behaviours — such as extended ON durations, incomplete cycles, or missing operational events — without the need for labelled fault data. Across multiple REFIT households (Houses 2, 21, and 15), the combined UGSP + rule-based approach consistently achieved higher recall and F-Scores than UGSP + clustering, particularly for long ON-duration anomalies in fridge-freezers and washing machines. While clustering methods were useful for exploratory analysis and offline fault pattern discovery, the rule-based approach proved more reliable for online monitoring and real-time decision support. The results also indicated that while anomaly detection precision was consistently high, recall remained moderate, reflecting the intrinsic challenge of detecting rare and intermittent appliance faults. Nonetheless, the demonstrated framework confirms that NILM-based anomaly detection can serve as a low-cost, scalable, and non-intrusive diagnostic tool for household energy systems, supporting preventive maintenance and energy safety.

Taken together, this research provides a validated methodological framework for NILM that integrates feature engineering, algorithm selection, and anomaly detection into a unified process. The findings highlight the complementary roles of supervised and unsupervised learning: the former excelling in precision and interpretability, and the latter offering flexibility and adaptability to unlabelled data environments. Moreover, the study reveals the trade-offs between pre-processing complexity and classification performance, emphasizing that excessive signal manipulation can distort event boundaries and degrade model accuracy. The cross-household transfer learning experiments further expose the need for hybrid NILM architectures capable of learning transferable representations, possibly through deep feature embeddings or semi-supervised adaptation frameworks.

In conclusion, this thesis advances both the theoretical and practical understanding of NILM by bridging the gap between algorithmic optimization and real-world deployment. The integration of UGSP-based disaggregation, rule-based anomaly detection, and careful feature selection establishes a strong foundation for scalable, interpretable, and cost-effective load monitoring systems. Future extensions of this work could explore adaptive NILM models that learn from continuous data streams, context-driven energy analytics for behaviour-aware prediction, and integration with IoT-enabled smart grids for holistic household energy intelligence. Ultimately, this research contributes to the broader goal of enabling efficient, safe, and intelligent energy management through data-driven NILM frameworks.

6.2 Future Work

While this thesis has provided valuable insights into feature selection, classification, and anomaly detection for Non-Intrusive Load Monitoring (NILM), several promising directions remain for further research. Building on the methodologies and findings presented in the preceding chapters, future work should focus on enhancing the generalizability, adaptability, and real-world applicability of NILM frameworks, particularly in the context of smart meter analytics and intelligent energy management.

A primary avenue for future investigation involves improving transferability and generalization across different households and datasets. As discussed in Chapter 4, transfer learning experiments revealed significant degradation in model performance when NILM algorithms trained on one household were tested on another. This issue primarily arises from variations in appliance characteristics, user behaviour, and background noise profiles. Future studies should therefore explore domain adaptation and federated learning frameworks to enable cross-household model sharing without the need for extensive retraining or centralized data storage. Such approaches could enhance data privacy while promoting scalability across larger populations. Moreover, integrating contextual and metadata features—including appliance age, occupancy patterns, environmental conditions, and building characteristics—could substantially improve model robustness and generalization by allowing NILM systems to adapt to situational differences between households.

Another promising direction concerns the development of advanced NILM algorithms that leverage recent progress in deep learning. Although traditional machine-learning algorithms such as Decision Trees, K-Nearest Neighbours, and DBSCAN have demonstrated practical utility, emerging architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer models could offer improved feature extraction and temporal learning capabilities. Future research could investigate these architectures at low sampling rates, representative of smart meter data, and examine hybrid GSP–deep learning models that combine the structural interpretability of Graph Signal Processing with the representational strength of deep networks. In addition, incorporating temporal and contextual signals—such as outdoor temperature, energy tariffs, and user schedules—may further enhance the accuracy of appliance classification and enable NILM systems to operate as intelligent, context-aware energy management tools.

The next area of opportunity involves real-time and adaptive anomaly detection. In Chapter 5, a rule-based anomaly detection framework was proposed and validated using UGSP disaggregation outputs. While effective, this framework relied on fixed thresholds that were uniformly applied across all datasets. In practice, static thresholds can result in false positives in some contexts and missed detections in others, especially when appliances age or environmental conditions change. Future work should therefore aim to develop adaptive and self-learning anomaly detection systems capable of dynamically adjusting their parameters based on historical trends, seasonal variations, and appliance degradation. Approaches such as rolling statistical windows, Bayesian updating, and reinforcement learning could be used to continuously refine detection thresholds in real time. Furthermore, deploying these adaptive systems on embedded smart meter platforms or edge computing devices would enable online, low-latency anomaly detection, improving safety and reliability without requiring cloud-level computation.

Future research should also focus on validation using larger and more diverse datasets to ensure scalability and representativeness. While this thesis primarily used the REFIT and REDD datasets, additional evaluation on other public datasets such as UK-DALE, ECO, and AMPds, as well as region-specific datasets, would provide broader insights into algorithm robustness. Conducting cross-regional analyses would help account for differences in appliance types, voltage standards, and behavioural practices. Such studies would strengthen the external validity of NILM algorithms and inform the development of globally applicable NILM frameworks.

An additional line of investigation involves the integration of NILM with user feedback mechanisms and Internet of Things (IoT) technologies. Coupling NILM-based analytics with user-facing mobile or web interfaces could allow residents to receive real-time notifications of abnormal appliance behaviour, facilitating prompt corrective actions and improved energy awareness. Combining NILM outputs with auxiliary IoT sensor data—such as temperature, vibration, or acoustic measurements—could further enhance reliability through multimodal anomaly detection. This fusion of electrical and environmental sensing would provide richer contextual information, reducing uncertainty and enabling more precise fault diagnostics.

A related enhancement involves the continued development of dynamic thresholding mechanisms for NILM anomaly detection. The static thresholds used in this work, though effective in initial validation, do not account for evolving household behaviour or environmental variability. Future approaches should aim to construct adaptive threshold models that evolve alongside appliance usage patterns and operating conditions. Techniques from adaptive control theory, probabilistic modelling, and time-varying statistical estimation could provide the mathematical foundation for such models, enabling NILM systems to sustain accuracy and reliability over long-term deployments.

A further and closely related direction for future research is the extension of the proposed two-stage disaggregation framework through tighter integration with transfer learning techniques. In this thesis, the two-stage strategy demonstrated improved robustness by first grouping appliances with similar load characteristics before performing fine-grained disaggregation. Future work could explore transferring knowledge at each stage separately: for example, learning generic appliance group representations (Stage 1) across multiple households, while adapting appliance-specific classifiers

(Stage 2) using limited local data. Such a hierarchical transfer learning approach could reduce sensitivity to household-specific variations while preserving classification accuracy.

Finally, future research should seek to quantify the broader safety, energy, and policy impacts of NILM deployment. NILM-based anomaly detection offers significant potential to identify electrical faults and hazardous conditions before they escalate, thereby reducing the risk of appliance-related fires or energy losses. Quantifying these safety benefits would provide compelling evidence for policymakers and utilities to support large-scale implementation. Furthermore, by offering granular insights into end-use consumption, NILM can play a key role in carbon reduction and demand response programs, helping consumers modify their behaviour in line with sustainability goals. Integrating NILM frameworks into national smart meter infrastructures could thus enable cost-effective, scalable anomaly detection and energy management solutions that contribute directly to net-zero and energy-efficiency policy targets.

In summary, future research should aim to develop NILM systems that are not only technically robust and computationally efficient but also contextually intelligent, adaptive, and user-centric. Advancements in algorithmic generalization, data diversity, and real-time adaptability will enable NILM to evolve from a primarily research-focused field into a practical, widely deployed technology supporting smart, sustainable, and safe energy systems. The integration of NILM with emerging technologies such as IoT, edge computing, and federated learning will further reinforce its role as a cornerstone of the next generation of intelligent energy analytics and household monitoring infrastructures.

6.3 Final Remarks

This thesis has advanced understanding of NILM through systematic exploration of feature selection, disaggregation, and anomaly detection. While significant challenges remain—particularly in transferability and online detection—the results show that NILM, when combined with lightweight rule-based frameworks, can move beyond energy feedback into safety-critical applications. The integration of NILM with anomaly detection represents a practical pathway toward safer, smarter, and more sustainable residential energy systems.

Chanter 7

Appendix

7.1. Summary table comparing the performance of different pre-processing filter selections

Table	Case	Algorithm	No pre-processing	Median filtering	Edge sharpening	Median filtering + edge sharpening	Median filtering + GBF + Edge sharpening
4.1	DW performance for REDD House 1 with DT	DT	PR=0.65, RE=0.84, F=0.73, Acc=0.63	PR=0.59, RE=0.83, F=0.69, Acc=0.61	PR=0.67, RE=0.72, F=0.69, Acc=0.63	PR=0.67, RE=0.63, F=0.65, Acc=0.61	PR=0.64, RE=0.66, F=0.64, Acc=0.60
4.2	DW performance for REDD House 1 with KNN	KNN	PR=0.65, RE=0.83, F=0.73, Acc=0.63	PR=0.65, RE=0.81, F=0.72, Acc=0.63	PR=0.70, RE=0.61, F=0.66, Acc=0.62	PR=0.67, RE=0.65, F=0.66, Acc=0.62	PR=0.48, RE=0.72, F=0.57, Acc=0.50
4.3	DW performance for REDD House 1 with DBSCAN	DBSCAN	PR=0.46, RE=0.65, F=0.54, Acc=0.52	PR=0.47, RE=0.11, F=0.18, Acc=0.51	PR=0.59, RE=0.14, F=0.23, Acc=0.52	PR=0.58, RE=0.23, F=0.33, Acc=0.54	PR=0.47, RE=0.65, F=0.34, Acc=0.51
4.4	DW performance for REFIT House 2 with DT	DT	PR=0.63, RE=0.90, F=0.74, Acc=0.67	PR=0.73, RE=0.79, F=0.76, Acc=0.73	PR=0.53, RE=0.58, F=0.56, Acc=0.53	PR=0.63, RE=0.65, F=0.64, Acc=0.62	PR=0.53, RE=0.72, F=0.61, Acc=0.54
4.5	DW performance for REFIT House 2 with KNN	KNN	PR=0.62, RE=0.87, F=0.72, Acc=0.65	PR=0.62, RE=0.83, F=0.70, Acc=0.64	PR=0.44, RE=0.72, F=0.55, Acc=0.42	PR=0.48, RE=0.77, F=0.59, Acc=0.46	PR=0.62, RE=0.72, F=0.67, Acc=0.63
4.6	DW performance for REFIT House 2 with DBSCAN	DBSCAN	PR=0.44, RE=0.63, F=0.52, Acc=0.42	PR=0.48, RE=0.92, F=0.63, Acc=0.45	PR=0.34, RE=0.29, F=0.31, Acc=0.37	PR=0.51, RE=0.47, F=0.49, Acc=0.51	PR=0.47, RE=0.39, F=0.43, Acc=0.48
4.7	WD performance for REDD House 1 with DT	DT	PR=0.94, RE=0.98, F=0.96, Acc=0.96	PR=0.81, RE=0.84, F=0.83, Acc=0.82	PR=0.72, RE=0.67, F=0.69, Acc=0.70	PR=0.80, RE=0.88, F=0.84, Acc=0.83	PR=0.68, RE=0.94, F=0.79, Acc=0.74

Table	Case	Algorithm	No pre-processing	Median filtering	Edge sharpening	Median filtering + edge sharpening	Median filtering + GBF + Edge sharpening
4.8	WD performance for REDD House 1 with KNN	KNN	PR=0.91, RE=0.98, F=0.94, Acc=0.94	PR=0.23, RE=0.61, F=0.33, Acc=0.22	PR=0.23, RE=0.61, F=0.33, Acc=0.22	PR=0.20, RE=0.51, F=0.29, Acc=0.25	PR=0.47, RE=0.49, F=0.48, Acc=0.47
4.9	WD performance for REDD House 1 with DBSCAN	DBSCAN	PR=0.63, RE=0.71, F=0.67, Acc=0.64	PR=0.06, RE=0.02, F=0.03, Acc=0.32	PR=0.54, RE=0.22, F=0.31, Acc=0.51	PR=0.44, RE=0.08, F=0.14, Acc=0.49	PR=0.54, RE=0.10, F=0.16, Acc=0.51
4.10	WM performance for REFIT House 2 with DT	DT	PR=0.18, RE=0.44, F=0.26, Acc=0.18	PR=0.27, RE=0.54, F=0.36, Acc=0.07	PR=0.08, RE=0.11, F=0.09, Acc=0.02	PR=0.32, RE=0.39, F=0.35, Acc=0.31	PR=0.29, RE=0.30, F=0.29, Acc=0.18
4.11	WM performance for REFIT House 2 with KNN	KNN	PR=0.15, RE=0.16, F=0.16, Acc=0.20	PR=0.20, RE=0.35, F=0.25, Acc=0.07	PR=0.09, RE=0.17, F=0.12, Acc=0.11	PR=0.26, RE=0.40, F=0.32, Acc=0.20	PR=0.27, RE=0.33, F=0.30, Acc=0.27
4.12	WM performance for REFIT House 2, DBSCAN	DBSCAN	PR=0.12, RE=0.76, F=0.21, Acc=0.04	PR=0.11, RE=0.56, F=0.19, Acc=0.07	PR=0.07, RE=0.42, F=0.13, Acc=0.04	PR=0.10, RE=0.57, F=0.18, Acc=0.24	PR=0.12, RE=0.52, F=0.16, Acc=0.24
4.13	TD performance for REFIT House 3 with DT	DT	PR=0.34, RE=0.78, F=0.47, Acc=0.10	PR=0.32, RE=0.52, F=0.40, Acc=0.21	PR=0.23, RE=0.30, F=0.26, Acc=0.14	PR=0.23, RE=0.43, F=0.30, Acc=0.01	PR=0.25, RE=0.44, F=0.32, Acc=0.06
4.14	TD performance for REFIT House 3 with KNN	KNN	PR=0.34, RE=0.65, F=0.44, Acc=0.17	PR=0.30, RE=0.60, F=0.40, Acc=0.10	PR=0.18, RE=0.32, F=0.23, Acc=0.04	PR=0.29, RE=0.49, F=0.36, Acc=0.13	PR=0.27, RE=0.47, F=0.34, Acc=0.08
4.15	TD performance for REFIT House 3 with DBSCAN	DBSCAN	PR=0.09, RE=0.88, F=0.16, Acc=0.02	PR=0.08, RE=0.80, F=0.15, Acc=0.09	PR=0.05, RE=0.45, F=0.09, Acc=0.04	PR=0.09, RE=0.68, F=0.16, Acc=0.08	PR=0.09, RE=0.61, F=0.16, Acc=0.24

Bibliography:

[1] Smart Meter Statistics in Great Britain: Quarterly Report to end September 2024 published on 28 November 2024

[2] G.Poursharif, 2018. Investigating the Ability of Smart Electricity Meters to Provide Accurate Low Voltage Network Information to the UK Distribution Network Operators (Doctoral dissertation, University of Sheffield)

[3] Department for Energy Security and Net Zero, Smart Meter Targets Framework: Government response to a consultation on minimum installation requirements for Year 3 (2024) and Year 4 (2025), July 2023

[4] K.He, D.Jakovetic, , B. Zhao, V. Stankovic, L. Stankovic, and S. Cheng, 2019. A Generic Optimisation-based Approach for Improving Non-intrusive Load Monitoring. IEEE Transactions on Smart Grid.

[5] A.Ridi, C. Gisler, and J.Hennebert, 2014, October. Appliance and state recognition using hidden markov models. In 2014 International Conference on Data Science and Advanced Analytics (DSAA) (pp. 270-276). IEEE.

[6] Mohammad Khazaei, Lina Stankovic, and Vladimir Stankovic 2019. "Trends and challenges in smart metering analytics." In 2019 MTMI International Conference on Emerging Issues in Business, Technology and Applied Sciences, pp. 111-117.

[7] Kanghang He, Lina Stankovic, Jing Liao, and Vladimir Stankovic. 2016, “Non-intrusive load disaggregation using graph signal processing,” IEEE Trans. Smart Grids, vol. 9, pp. 1739-1747.

[8] Haroon Rashid, Vladimir Stankovic, Lina Stankovic, and Pushpendra Singh 2019. "Evaluation of non-intrusive load monitoring algorithms for appliance-level anomaly detection." In Proc. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8325-8329. IEEE

[9] Haroon Rashid, Pushpendra Singh, Vladimir Stankovic, and Lina Stankovic 2019. "Can non-intrusive load monitoring be used for identifying an appliance’s anomalous behaviour?" Elsevier Applied Energy 238: 796-805.

[10] https://www.bbc.co.uk/news/uk-33124925

[11] https://www.firescotland.gov.uk/news-campaigns/news/2020/01/electricity-safety-(7).aspx

[12] George William Hart 1992, “Nonintrusive appliance load monitoring,” Proceedings of the IEEE, vol. 80, no. 12, pp. 1870–1891.

[13] J Zico Kolter and Tommi S Jaakkola, 2012 “Approximate inference in additive factorial hmms with application to energy disaggregation.,” in AISTATS, vol. 22, pp. 1472–1482

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

[15] Mingjun Zhong, Nigel Goddard, and Charles Sutton 2015. "Latent Bayesian melding for integrating individual and population models." In Advances in Neural Information Processing Systems, pp. 3618-3626.

[16] Stephen Makonin, Fred Popowich, Ivan V. Bajić, Bob Gill, and Lyn Bartram 2015. "Exploiting HMM sparsity to perform online real-time nonintrusive load monitoring." IEEE Transactions on Smart Grid 7, no. 6: 2575-2585.

[17] Bochao Zhao, Lina Stankovic, and Vladimir Stankovic 2016. "On a training-less solution for non-intrusive appliance load monitoring using graph signal processing." IEEE Access 4: 1784-1799.

[18] Zoha, A., Gluhak, A., Imran, M. A., & Rajasegarar, S. (2012). Non-intrusive load monitoring approaches for disaggregated energy sensing: A survey. Sensors, 12(12), 16838-16866. https://doi.org/10.3390/s121216838

[19] He, K., Stankovic, V., & Stankovic, L. (2016). Feature selection for Non-Intrusive Load Monitoring. Smart Grid and Renewable Energy, 7(2), 90–100.

[20] Parson, O., Ghosh, S., Weal, M., & Rogers, A. (2012). Non-intrusive load monitoring using prior models of general appliance types. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 26, No. 1).

[21] Kelly, J., & Knottenbelt, W. (2016). Neural NILM: Deep neural networks applied to energy disaggregation. Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments.

[22] Zhao, B., He, K., Stankovic, L., & Stankovic, V. (2018). Improving event-based non-intrusive load monitoring using graph signal processing. IEEE Access, 1-15. https://doi.org/10.1109/ACCESS.2018.2871343

[23] Krystalakos, O., Nalmpantis, C., & Vrakas, D. (2018). Sliding window approach for online energy disaggregation using deep neural networks. Energy and Buildings, 158, 1523–1532.

[24] Cominola, A., Giuliani, M., Piga, D., Castelletti, A., & Rizzoli, A. E. (2021). A hybrid signature-based and model-based approach for automatic appliance fault detection and diagnosis. Applied Energy, 283, 116259.

[25] Al-Waisi, Z., & Agyeman, M. O. (2018, September). On the Challenges and Opportunities of Smart Meters in Smart Homes and Smart Grids. In Proc. 2nd Int. Symposium on Computer Science and Intelligent Control (p. 16). ACM.

[26]	Uribe-Pérez, N., Hernández, L., De la Vega, D., & Angulo, I. (2016). State of the art and trends review of smart metering in electricity grids. Applied Sciences, 6(3), 68.
[27]	ICCS-NTUA &AF Mercados EMI, 2015, Study on cost benefit analysis of Smart Metering Systems, FINAL REPORT in EU Member states
[28]	Emmanuel, M., & Rayudu, R. (2016). Communication technologies for smart grid applications: A survey. Journal of Network and Computer Applications, 74, 133-148.
[29]	Bian, D., Kuzlu, M., Pipattanasomporn, M., Rahman, S., & Shi, D. (2019). Performance evaluation of communication technologies and network structure for smart grid applications. IET Commun., 13(8), 1025-1033.
[30]	Khazaei, M., Stankovic, L., & Stankovic, V. (2020, November). Evaluation of low-complexity supervised and unsupervised NILM methods and pre-processing for detection of multistate white goods. In 5th International Workshop on Non Intrusive Load Monitoring.
[31]	Alahakoon, D., & Yu, X. (2015). Smart electricity meter data intelligence for future energy systems: A survey. IEEE Transactions on Industrial Informatics, 12(1), 425-436.
[32]	Prado, J. G., González, A., & Riaño, S. (2017). Adopting smart meter events as key data for low-voltage network operation. CIRED-Open Access Proceedings Journal, 2017(1), 924-928
[33]	Moghaddass, R., & Wang, J. (2017). A hierarchical framework for smart grid anomaly detection using large-scale smart meter data. IEEE Transactions on Smart Grid, 9(6), 5820-5830
[34]	Wang, Y., Chen, Q., Hong, T., & Kang, C. (2018). Review of smart meter data analytics: Applications, methodologies, and challenges. IEEE Transactions on Smart Grid, 10(3), 3125-3148.
[35]	Jokar, P., Arianpoo, N., & Leung, V. C. (2015). Electricity theft detection in AMI using customers' consumption patterns. IEEE Transactions on Smart Grid, 7(1), 216-226
[36]	Jindal, A., Dua, A., Kaur, K., Singh, M., Kumar, N., & Mishra, S. (2016). Decision tree and SVM-based data analytics for theft detection in smart grid. IEEE Transactions on Industrial Informatics, 12(3), 1005-1016.
[37]	Depuru, S. S. S. R., Wang, L., Devabhaktuni, V., & Green, R. C. (2013). High performance computing for detection of electricity theft. International Journal of Electrical Power & Energy Systems, 47, 21-30. Eco-Bot: http://eco-bot.eu/
[38]	Smart Meters and Losses: Best Practice Review. UK Power Networks (Operations) Limited, 2016.

[39]	Buzau, M. M., Tejedor-Aguilera, J., Cruz-Romero, P., & Gómez-Expósito, A. (2018). Detection of non-technical losses using smart meter data and supervised learning. IEEE Transactions on Smart Grid, 10(3), 2661-2670
[40]	Hong, T., Pinson, P., & Fan, S. (2014). Global energy forecasting competition 2012.
[41]	Espinoza, M., Joye, C., Belmans, R., & De Moor, B. (2005). Short-term load forecasting, profile identification, and customer segmentation: A methodology based on periodic time series. IEEE Trans. Power Systems, 20, 1622-1630.
[42]	Murray, D., Stankovic, L., Stankovic, V., Lulic, S., Sladojevic, S. (2019). Transferability of neural network approaches for low-rate energy disaggregation. ICASSP IEEE Int. Conf. Acs., Speech & Sig. Proc. pp. 8330-8334
[43]	Figueiredo, M., Ribeiro, B., & de Almeida, A. T. (2014). Electrical signal feature extraction for non-intrusive load monitoring based on the S-transform. Signal Processing, 96, 18-29.
[44]	Zander, S., Nguyen, T., & Armitage, G. (2005). Automated traffic classification and application identification using machine learning. In Proceedings of the IEEE Conference on Local Computer Networks (pp. 250–257). IEEE. https://doi.org/10.1109/LCN.2005.44
[45]	Gao, J., Giri, S., Kara, E. C., & Berges, M. (2019). Plaid: A public dataset of high-resolution electrical appliance measurements for load identification research. Energy and Buildings, 128, 198-211.
[46]	Stankovic, L., Stankovic, V., Lulic, I., & Clauß, S. (2016). Detecting small appliances in smart meter data using prior models and change detection. 2016 IEEE 21st International Conference on Digital Signal Processing (DSP), 274–278.
[47]	Stankovic, L., Lulic, I., & Stankovic, V. (2015). A robust disaggregation method for low-frequency smart meter data. Energy and Buildings, 86, 710–720.
[48]	Murray, D., Stankovic, L., & Stankovic, V. (2017). An explainable fall detection system using raw smart meter data. Energies, 10(11), 1680.
[49]	A. Ridi, C. Gisler, and J. Hennebert. ACS-F2 - A New Database of Appliance Consumption Analysis. In Proceedings of the International Conference on Soft Computing and Pattern Recognition (SocPar 2014), 2014
[50]	de Paiva Penha, D., & Castro, A. R. G. (2017, December). Convolutional neural network applied to the identification of residential equipment in non-intrusive load monitoring systems. In 3rd International Conference on Artificial Intelligence and Applications (pp. 11-21).

[51] KOLTER, J. Zico; JOHNSON, Matthew J. REDD: A public data set for energy disaggregation research. In: Workshop on Data Mining Applications in Sustainability (SIGKDD), San Diego, CA. 2011. p. 59-62.

[52] David Murray, Lina Stankovic, and Vladimir Stankovic 2017. "An electrical load measurements dataset of United Kingdom households from a two-year longitudinal study." Scientific Data 4, no. 1: 1-12.

[53] Altrabalsi, H., Stankovic, V., Liao, J., & Stankovic, L. (2016). Low-complexity energy disaggregation using appliance load modelling. AIMS Energy, 4(1), 884-905.

[54] Elafoudi, G., Stankovic, L., & Stankovic, V. (2014). Power disaggregation of domestic smart meter readings using dynamic time warping. In 6th Int. Symp. Communications, Control and Signal Processing (ISCCSP), 2014 IEEE

[55] Kamat, S. P. (2004, November). Fuzzy logic based pattern recognition technique for non-intrusive load monitoring. In 2004 IEEE Region 10 Conference TENCON 2004. (Vol. 100, pp. 528-530). IEEE.

[56] Srinivasan, D., Ng, W. S., & Liew, A. C. (2005). Neural-network-based signature recognition for harmonic source identification. IEEE Transactions on Power Delivery, 21(1), 398-405.

[57] M. Khazaei and S. Jadid, "Contingency ranking using neural networks by Radial Basis Function method," 2008 IEEE/PES Transmission and Distribution Conference and Exposition, Chicago, IL, USA, 2008, pp. 1-4, doi: 10.1109/TDC.2008.4517045.

[58] T.M. Mitchell, Machine Learning, McGraw-Hill Companies, USA, 1997

[59] M. Garofalakis, D. Hyun, R. Rastogi and K. Shim, “Building Decision Trees with Constraints”, Data Mining and Knowledge Discovery, vol. 7, no. 2, 2003, pp. 187 – 214

[60] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1, pp. 281–297

[61] Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD), pp. 226–231

[62] Shuman, D. I., Narang, S. K., Frossard, P., Ortega, A., & Vandergheynst, P. (2013). The Emerging Field of Signal Processing on Graphs: Extending High-Dimensional Data Analysis to Networks and Other Irregular Domains. IEEE Signal Processing Magazine, 30(3), 83–98

[63] Zhao, L., & Stankovic, V. (2015). Blind signal separation for non-intrusive load monitoring. Energy Informatics Conference.

[64] Zhong, M., Wang, Z., & Goddard, N. (2018). Deep learning for energy disaggregation: A comparative study. Energy Informatics Journal.

[65] Makonin, S., & Popowich, F. (2016). Non-intrusive load monitoring with off-the-shelf hardware and software. Springer Energy Efficiency Journal.

[66] Anderson, K., Berges, M., & Gupta, S. (2020). Advances in non-intrusive load monitoring. IEEE Smart Grid Conference.

[67] Wang, Y., Chen, Z., & Zhang, C. (2021). Federated learning for privacy-preserving NILM. IEEE Transactions on Smart Grid

[68] Makonin, S., Bajic, I. V., & Popowich, F. (2014). Real-Time Energy Disaggregation on a Budget. Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments (BuildSys).

[69] Zhang, C., Zhong, M., Wang, Z., et al. (2018). Sequence-to-Point Learning with Neural Networks for Energy Disaggregation. AAAI Conference on Artificial Intelligence.

[70] Faustine, A., Mvungi, N. H., Kaijage, S., & Michael, K. (2017). A survey on non-intrusive load monitoring methods and techniques for energy disaggregation problems. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7(6), e1265. https://doi.org/10.1002/widm.1265

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

[72] Chicco, G., Napoli, R., & Piglione, F. (2021). Load disaggregation approaches for non-intrusive load monitoring: A review. IEEE Transactions on Industrial Informatics, 17(6), 3893–3904

[73] Kita, D. M., Miranda, B., Favela, D., Bono, D., Michon, J., Lin, H., .. & Hu, J. (2018). High- performance and scalable on-chip digital Fourier transform spectroscopy. Nature communications, 9(1), 4405.

[74] Hosseini, S. S., Agbossou, K., Kelouwani, S., Cardenas, A., & Henao, N. (2020). A Practical Approach to Residential Appliances on-Line Anomaly Detection: A Case Study of Standard and Smart Refrigerators. IEEE Access, 8, 57905-57922.

[75] Ganu T, Rahayu D, Seetharam DP, Kunnath R, Kumar AP, Arya V, et al. Socketwatch: an autonomous appliance monitoring system. Pervasive computing and communications (PerCom), 2014 IEEE international conference on. IEEE; 2014. p. 38–43.

[76] Pereira W, Ferscha A, Weigl K. Unsupervised detection of unusual behaviours from smart home energy data. International conference on artificial intelligence and soft computing. Springer; 2016. p. 523–34.

[77] Ahmadi-Karvigh S, Ghahramani A, Becerik-Gerber B, Soibelman L. Real-time activity recognition for energy efficiency in buildings. Appl Energy 2018;211:146–60.

[78] Himeur, Y., Alsalemi, A., Bensaali, F., & Amira, A. (2021). Smart power consumption abnormality detection in buildings using micromoments and improved K-nearest neighbours. International Journal of Intelligent Systems, 36(6), 2865

[79] D. Pyle, Data preparation for data mining, 1st Vol., Morgan Kaufmann publisher, San Francisco, 1999

[80] I. Guyon, N. Matic and V. Vapnik, "Discovering informative patterns and data cleaning", In: Fayyad UM, Piatetsky-Shapiro G, Smyth P and Uthurusamy R. (ed) Advances in knowledge discovery and data mining, AAAI/MIT Press, California, 1996, pp. 181-203

[81] E. Simoudis, B. Livezey B and R. Kerber R , "Integrating inductive and deductive reasoning for data mining", In: Fayyad UM, Piatetsky-Shapiro G, Smyth P and Uthurusamy R. (Eds.) Advances in knowledge discovery and data mining, AAAI/MIT Press, California, 1996, pp. 353-373

[82] Beniwal, S., & Arora, J. (2012). Classification and feature selection techniques in data mining. International journal of engineering research & technology (ijert), 1(6), 1-6.

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

[84] Herrero-Bermello, A., Li, J., Khazaei, M., Grinberg, Y., Velasco, A. V., Vachon, M., .. & Schmid, J. H. (2019). On-chip Fourier-transform spectrometers and machine learning: a new route to smart photonic sensors. Optics letters, 44(23), 5840-5843

[85] Ferrari, M., & Quaresima, V. (2012). A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. NeuroImage, 63(2), 921–935. https://doi.org/10.1016/j.neuroimage.2012.03.049

[86] Akca, B. I., & Doerr, C. R. (2019). Polarization-independent silicon photonic grating couplers with subwavelength-engineered waveguide ports. IEEE Photonics Technology Letters, 31(2), 90–93. https://doi.org/10.1109/LPT.2018.2884880

[87] Ma, K., Chen, K., Zhu, N., Liu, L., & He, S. (2018). Non-blocking 4× 4 silicon electro-optic switches based on broadband variable optical attenuators. IEEE Photonics Journal, 11(4), 4900107. https://doi.org/10.1109/JPHOT.2018.2847446

[88] Huang, J., Yang, J., Zhang, H., Zhang, J., Wu, W., & Chang, S. (2016). Low-loss and broadband three-mode polarization rotator based on a silicon-on-insulator platform. IEEE Photonics Technology Letters, 28(18), 2677–2680. https://doi.org/10.1109/LPT.2016.2599071

[89] Velasco, A. V., Cheben, P., Bock, P. J., Delâge, A., Schmid, J. H., Lapointe, J., .. & Vachon, M. (2013). High-resolution Fourier-transform spectrometer chip with microphotonic silicon spiral waveguides. Optics Letters, 38(5), 706–708. https://doi.org/10.1364/OL.38.000706

[90] Soref, R. A., Leonardis, F. D., & Passaro, V. M. N. (2019). Multiband photonic crystal devices for waveguiding and sensing. Journal of Lightwave Technology, 37(12), 3192–3201. https://doi.org/10.1109/JLT.2019.2902482

[91] Akca, B. I. (2017). Polarization-independent photonic crystal filters for integrated optics. Optics Express, 25(7), 1487–1496. https://doi.org/10.1364/OE.25.001487

[92] Souza, M. C. M. M., Grieco, A., Frateschi, N. C., & Fainman, Y. (2018). Integrated photonic Fourier-transform spectrometers with micro-ring resonators. Nature Communications, 9(1), 665. https://doi.org/10.1038/s41467-018-02984-7

[93] Herrero-Bermello, A., Velasco, A. V., Podmore, H., Cheben, P., Schmid, J. H., Janz, S., .. & Corredera, P. (2017). Advances in subwavelength grating waveguides. Optics Letters, 42(11), 2239–2242. https://doi.org/10.1364/OL.42.002239

[94] Sharpe, S. W., Johnson, T. J., Sams, R. L., Chu, P. M., Rhoderick, G. C., & Johnson, P. A. (2004). Gas-phase databases for infrared spectral analysis. Applied Spectroscopy, 58(12), 1452–1461. https://doi.org/10.1366/0003702042641281

[95] Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., & Movellan, J. (2005). Recognizing facial expression: Machine learning and application to spontaneous behaviour. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (pp. 568–573). IEEE. https://doi.org/10.1109/CVPR.2005.297

[96] Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics, 16(6), 321–332.

[97] Zeifman, M., & Roth, K. (2011). Nonintrusive appliance load monitoring: Review and outlook. IEEE Transactions on Consumer Electronics, 57(1), 76-84.

[98] Canny, J. (1986). A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 8(6), 679–698.

[99] Wei, H., Zhang, J., Wang, H., & Yang, Q. (2019). Bilateral graph filtering for non-intrusive load monitoring. Energy and Buildings, 202, 109373. https://doi.org/10.1016/j.enbuild.2019.109373

[100] Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and practice (2nd ed.). OTexts. https://otexts.com/fpp2/

[101] Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

[102] Leeb, S. B., Kirtley, J. L., & Shaw, S. R. (1995). Transient event detection in spectral envelope estimates for nonintrusive load monitoring. In Proceedings of the IEEE

Instrumentation and Measurement Technology Conference (Vol. 2, pp. 1169–1174). https://doi.org/10.1109/IMTC.1995.515111

[103] Singh, K., & Santoso, S. (2012). Event detection and classification in NILM. In IEEE Power and Energy Society General Meeting (pp. 1–8).

[104] Chen, S., & Jacobsen, H. A. (2017). Toward accurate non-intrusive load monitoring: Benchmarking state-of-the-art algorithms. Energy and Buildings, 139, 368–377.

[105] Kaselimi, M., Doulamis, N., Voulodimos, A., Doulamis, A., & Protopapadakis, E. (2019). Bayesian-optimized bidirectional LSTM regression model for non-intrusive load monitoring. IEEE Transactions on Smart Grid, 10(5), 5768–5777.

[106] Schafer, J. L. (1997). Analysis of incomplete multivariate data. Chapman and Hall/CRC.

[107] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data mining: Practical machine learning tools and techniques (4th ed.). Morgan Kaufmann.

[108] Du, X., Gao, Y., & Wang, J. (2020). Power normalization in NILM systems. Energy and AI, 1, 100004.

[109] Laverty, D. M., Morrow, D. J., & Crossley, P. A. (2018). Phase alignment in three-phase systems: A survey. IEEE Transactions on Power Delivery, 33(3), 1187–1195.

[110] Bousbiat, H., Kessal, M., Bouallegue, A., Boudour, M., & Hadjami, A. (2018). Load disaggregation using Hidden Markov Models and event detection techniques. IEEE Transactions on Smart Grid, 9(6), 6074–6081.

[111] Liu, Y., Makonin, S., & Bajic, I. V. (2019). A novel recurrent neural network for non-intrusive load monitoring based on gated recurrent units. IEEE Transactions on Smart Grid, 10(1), 447–456. https://doi.org/10.1109/TSG.2017.2763021

[112] Bonfigli, R., Principi, E., Fagiani, M., Severini, M., Squartini, S., & Piazza, F. (2018). Non-intrusive load monitoring by using active and reactive power in additive Factorial Hidden Markov Models. Applied Energy, 208, 1595–1607.

[113] Mauch, L., & Yang, B. (2016). A new approach for supervised power disaggregation by using a deep recurrent LSTM network. IEEE Transactions on Smart Grid, 8(3), 1274–1282.

[114] Rafiq, M., Luo, H., Hussain, A., & Luo, H. (2020). An event-driven appliance identification approach using graph-based semi-supervised learning. IEEE Transactions on Smart Grid, 11(5), 4223–4233.

[115] Koutitas, G., Hall, S., & Demestichas, P. (2019). Deep learning-based multi-modal appliance classification. IEEE Access, 7, 96345–96355.

[116] Shen, X., Zhang, Y., He, Z., & Jiang, C. (2020). Non-intrusive load monitoring with domain adaptation and active learning. IEEE Transactions on Industrial Informatics, 16(7), 4612–4621.

[117] Fitra Hidiyanto and Abdul Halim. 2020. KNN Methods with Varied K, Distance and Training Data to Disaggregate NILM with Similar Load Characteristic. In Proceedings of the 3rd Asia Pacific Conference on Research in Industrial and Systems Engineering 2020 (APCORISE 2020). Association for Computing Machinery, New York, NY, USA, 93–99.

[118] Jing Liao, Georgia Elafoudi, Lina Stankovic, and Vladimir Stankovic, 2014. "Non-intrusive appliance load monitoring using low-resolution smart meter data." In Proc. 2014 IEEE International Conference on Smart Grid Communications (SmartGridComm), pp. 535-540.

[119] Zhixiang Xu, Ningxuan Guo, Yinan Wang, and Gangfeng Yan 2019. "Identifying Fridge Consumption Non-intrusively based on temporal characteristic." In 2019 IEEE 3rd Conference on Energy Internet and Energy System Integration (EI2), pp. 2770-2775. IEEE.

[120] Anderson, K., Berges, M., & Ocneanu, A. (2012). BLUED: A fully labeled public dataset for event-based non-intrusive load monitoring research. Proceedings of the 2nd KDD Workshop on Data Mining Applications in Sustainability (SustKDD).

[121] Barker, S., Mishra, A., Irwin, D., Shenoy, P., & Albrecht, J. (2012). Smart: An open data set and tools for enabling research in sustainable homes*. Proceedings of the 2nd KDD Workshop on Data Mining Applications in Sustainability (SustKDD).

[122] Reinhardt, A., Köhler, S., & Santini, S. (2012). On the accuracy of appliance identification based on smart meter data. Proceedings of the 2nd KDD Workshop on Knowledge Discovery from Sensor Data (SensorKDD).

[123] Pecan Street Inc. Dataport. (2013). Pecan Street Dataport. Retrieved from https://dataport.cloud

[124] Palmer, J., & Terry, N. (2011). Housing energy fact file 2011. Department of Energy & Climate Change.

[125] Makonin, S., Popowich, F., Bajic, I. V., Gill, B., & Bartram, L. (2013). AMPds: The Almanac of Minutely Power dataset (Version 1). Proceedings of the IEEE International Conference on Smart Grid Communications (SmartGridComm), 168–173.

[126] Monacchi, A., Egarter, D., Elmenreich, W., D'Alessandro, S., & Tonello, A. M. (2014). GREEND: An energy consumption dataset of households in Italy and Austria. IEEE International Conference on Smart Grid Communications, 511–516.

[127] Chakraborty, S., Dutta, K., & Sharma, A. (2015). COMBED: A combined energy dataset for smart energy research. Proceedings of the 2015 IEEE PES Innovative Smart Grid Technologies Conference (ISGT).

[128] Beckel, C., Sadamori, L., Santini, S., & Staake, T. (2014). ECO: A dataset of household electricity, water, and gas consumption. Proceedings of the 5th ACM Workshop on Embedded Systems for Energy-Efficient Buildings.

[129] Melfi, R., Rosenblum, B., Nordman, B., & Christensen, K. (2015). BERDS: Building energy research data set. Energy and Buildings, 94, 165–173.

[130] Ferreira, P., Ruano, A., Silva, S., & Conceição, P. (2017). SustData: A public dataset for building energy research. Energy and Buildings, 142, 249–257.

[131] D’Incecco, M., Squartini, S., & Zhong, M. (2019). Transfer learning for non-intrusive load monitoring. IEEE Transactions on Smart Grid, 11(2), 1419-1429.

[132] Rashid, H., Batra, N., & Singh, P. (2018). Rimor: Towards identifying anomalous appliances in buildings. Proceedings of the 5th Conference on Systems for Built Environments (BuildSys '18), 33–42. Association for Computing Machinery. https://doi.org/10.1145/3276774.3276797

[133] Department for Business, Energy & Industrial Strategy. (2021). Digest of UK energy statistics (DUKES) 2021: Chapter 5 – Electricity.

[134] Hosseini, S. S., Agbossou, K., Kelouwani, S., Cardenas, A., & Henao, N. F. (2021). Detection of anomalies in household appliances from disaggregated load consumption. Energy Informatics, 4(1), 1–18.

[135] Piga, D., Pierucci, S., & Ferraris, M. (2016). Can non-intrusive load monitoring be used for identifying an appliance’s anomalous behaviour? Energy Efficiency, 9(5), 1325–1343.

[136] Himeur, Y., Ghanem, K., Alsalemi, A., Bensaali, F., & Amira, A. (2021). Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives. Applied Energy, 287, 116601.

[137] Alsalemi, A., et al. (2020). Anomaly detection and diagnosis of appliances using power consumption time series. Energy Reports, 6, 401–406. https://doi.org/10.1016/j.egyr.2020.01.008

[138] De Almeida, A., Fonseca, P., Schlomann, B., & Feilberg, N. (2012). Characterization of the household electricity consumption in the EU, potential energy savings and specific policy recommendations. Energy and Buildings, 43(8), 1884–1894.

[139] Rashid, H., Kim, J., & Hassan, R. (2021). Appliance fault detection using NILM data. Sustainable Energy, Grids and Networks, 25, 100425.

[140]	Jin, X., McArthur, S. D. J., & McDonald, J. R. (2011). An approach to generation of synthetic power load profiles for domestic appliances. IEEE Transactions on Power Systems, 26(3), 1314–1323.
[141]	Tabone, M. D., Callaway, D. S., & Madsen, R. (2017). Fault detection for household refrigerators: Field testing results. Applied Energy, 189, 81–90.
[142]	Chicco, G., Napoli, R., & Piglione, F. (2003). Load pattern clustering for short-term load forecasting of anomalous days. Electric Power Systems Research, 69(3), 283–293.
[143]	UK Fire Services. (2023). Home appliance fire safety statistics. Retrieved from https://www.fireservice.co.uk
[144]	https://www.metoffice.gov.uk/