Advanced Filtering of Unknown Devices in Event-Based NILM

JUSTUS BREYER, Chair of Communication and Distributed Systems, RWTH Aachen University, Germany
MUHAMMAD HAMAD ALIZAI, LUMS, Pakistan
SANDEEP SAMANT, RWTH Aachen University, Germany
KLAUS WEHRLE, Chair of Communication and Distributed Systems, RWTH Aachen University, Germany

Non-Intrusive Load Monitoring (NILM), an important application of machine learning, frequently misidentifies activities of unknown devices, resulting in incorrect energy consumption estimates. This paper proposes an innovative filtering step between event detection and classification of event-based NILM to exclude events from unknown devices. This approach incorporates confidence-based classifiers, clustering, ensembling, and density-based techniques, notably Local Outlier Factor and One-Class SVM. The best techniques reduce false positives (over 93%) for unknown devices while preserving most events from known devices (less than 7% loss). This significant advancement enhances event-based NILM system accuracy, offering more reliable energy monitoring for real-world applications, and thereby contributes to broader energy conservation efforts in the context of climate change.

CCS Concepts: • **Computing methodologies → Machine learning**; *Anomaly detection*; *Ensemble methods*; *Supervised learning*.

Additional Key Words and Phrases: Energy disaggregation, Anomaly detection, Ensembling

Availability of Data and Material:
No additional material provided.

1 INTRODUCTION

Climate change necessitates reducing greenhouse gas emissions, a challenge exacerbated by the insufficient capacity of renewable energy sources to meet current demands. This calls for reducing energy consumption across various sectors, notably in residential areas. Research indicates homeowners being more inclined to save energy when they have access to detailed information about their consumption patterns [5, 10], aiding in the global sustainability effort.

In residential energy management, disaggregating household power load is crucial but challenging. Installing separate meters for each appliance, called Intrusive Load Monitoring (ILM), provides an intuitive but costly and complex solution. Non-Intrusive Load Monitoring (NILM) [14] offers a practical alternative by analyzing total consumption data from a single point, such as a smart meter, using time series analysis and machine learning. NILM, which divides into event-based and eventless methods [8], identifies appliance energy usage by detecting operational changes or using machine learning (ML) algorithms to distinguish device signatures.

NILM systems, particularly those using ML, face a critical limitation due to their dependence on specific models. This dependence limits their flexibility, as they often cannot adapt to new or changed appliances in dynamic residential environments. Event-based NILM systems, skilled at detecting device state changes, struggle with misclassification when encountering unknown devices, leading to inaccurate energy disaggregation. Depending on the share of unknown devices in the overall power consumption, these misclassifications can significantly decrease the reliability of the NILM system, as the unclassifiable activity is not simply ignored but instead assigned to one or more of the monitored devices, thereby skewing the results.

For maximum energy efficiency, the system should run on customer's premises, however, continuously updating classifiers to include new devices is burdensome for users, reducing NILM's practical use and acceptance. To overcome the manual updating of ML models for unknown device events, NILM technologies may use automated model adaptation or event separation mechanisms. These methods aim to distinguish *known* device events from *unknown* ones. Yet, they face challenges due to the unpredictable nature of residential settings, where new devices with unique energy profiles can emerge anytime. The unpredictability of distinguishing features between known and unknown devices complicates system optimization for future changes, limiting NILM's effectiveness in accurately identifying device usage in evolving residential environments.

To enhance the precision and flexibility of event-based NILM, our research makes the following novel contributions:

- Establishment of a novel filtering stage in event-based NILM, positioned between event detection and classification, specifically aimed at filtering out events from unknown devices. This includes the development of a range of methods to integrate this stage, utilizing techniques such as confidence-based, clustering, ensembling, and density-based approaches.
- Implementation and assessment of these methods using a publicly available dataset, focusing on their performance and capacity to adapt to new environments. This evaluation includes an investigation into the potential of these filters for domain transfer without necessitating retraining.
- Innovative exploration of density-based models, specifically Local Outline Factor (LOF) and One-Class Support Vector Machines (OC-SVMs), for device classification, representing a novel research direction that could significantly enhance event-based NILM system performance.

By accurately distinguishing between events from known and unknown devices, event-based NILM systems become more adaptable and reliable in dynamic residential environments. Our contributions tackle a key NILM challenge, fostering broader and more efficient adoption in the real world.

2 BACKGROUND AND LITERATURE REVIEW

This section lays the groundwork for our methodology, offering an overview of the event-based NILM pipeline and reviewing existing

Authors' addresses: Justus Breyer, breyer@comsys.rwth-aachen.de, Chair of Communication and Distributed Systems, RWTH Aachen University, Aachen, Germany; Muhammad Hamad Alizai, LUMS, Lahore, Pakistan, hamad.alizai@lums.edu.pk; Sandeep Samant, RWTH Aachen University, Aachen, Germany; Klaus Wehrle, Chair of Communication and Distributed Systems, RWTH Aachen University, Aachen, Germany, wehrle@comsys.rwth-aachen.de.

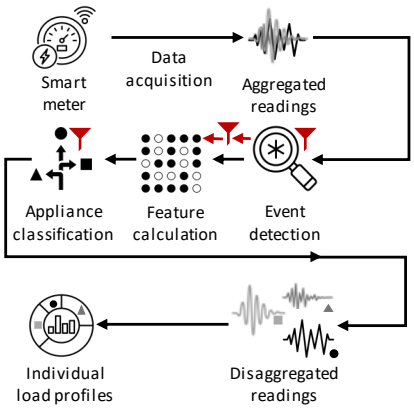


Fig. 1. Event-based NILM pipeline: Adaptations proposed in this paper are marked in red.

research on handling unknown devices and anomalies, a critical part of our study that is essential for comprehending our contributions.

2.1 Background

NILM systems can be divided into event-based and eventless categories, with each showing strengths within their specific areas. It is challenging to definitively claim one approach as superior to the other. Our study focuses on improving event-based NILM systems, and we will outline the key processes that define these systems, as depicted in black in Figure 1.

Data acquisition. The foundation of an NILM system is data collection, often through a sensor at a central point like a smart meter. Event-based NILM systems require data at high sampling rates, around 925 Hz or higher, to ensure the effectiveness of later stages, as lower rates may impair performance [15]. The sensor records a time series of aggregated power usage from all devices, providing the analysis base for subsequent NILM phases.

Event Detection. The second stage in an event-based NILM system is event detection, aimed at identifying state changes in appliances, such as turning on or off, by monitoring power consumption fluctuations. Event detection employs various methods like expert heuristics (e.g., threshold detection), probabilistic models, or matched filters [1], grounded in the Switch Continuity Principle (SCP) [14]. SCP asserts that only one appliance alters its state in a specific time frame, requiring a particular sampling rate for reliability, as it becomes less dependable over intervals of several seconds [21].

Feature Calculation. Handling time series feature calculation for detected events is variably treated in the literature, with some considering it a separate NILM component. This step is crucial for providing ML models with a manageable, reduced representation of the data. Given data complexity at several kHz, using the full dataset for classification is impractical. Feature calculation often includes initial data transformations like wavelet or Fourier transforms, with the suitability and efficiency of various features for NILM extensively studied in the literature [17, 25].

Appliance Classification. Computed features are used to link detected events to specific appliances or their operational states by training ML models on datasets. NILM model complexity varies, ranging from simpler models like RF or SVM [29] to more complex ones involving ensembling techniques [6] or neural networks [24]. For multi-label classification, models may provide confidence values for each potential device instead of a singular identification, allowing for a nuanced data interpretation.

Energy Disaggregation. In the final phase, the aggregated load is disaggregated into individual device consumption based on labeled events from prior steps. This important stage hinges on the accuracy of both event detection and classification. Mistakes in earlier stages can negatively impact this last step, causing inaccuracies in device power consumption estimates. The results of this energy disaggregation can then be relayed to the end-user, such as through a mobile application.

Creating an effective NILM system requires careful integration of its components, allowing little room for error. It is vital to recognize and mitigate factors that might reduce the system's effectiveness, such as the influence of unknown devices on the performance and accuracy of the NILM pipeline.

2.2 Related Work

Despite three decades of NILM research [2, 8], managing unknown devices has been somewhat overlooked, with significant contributions appearing only recently. Most NILM studies presuppose a static array of devices, disregarding the dynamic nature of typical households. However, this paper aims to summarize the current research on this particular challenge within NILM.

Extensible Models. To accommodate the evolving composition of household devices, developing a flexible NILM framework is crucial. Kong et al. [19] introduced a method employing a Factorial Hidden Markov Model (FHMM) that can be expanded to include new devices. However, this requires training HMMs for new devices with external data, potentially causing domain transfer problems, and lacks unknown device detection. Alternatively, Gillis and Morsie [11] proposed a decision tree-based framework that allows adding new devices by incorporating their wavelet features into the retraining process. Furthermore, Tanoni et al. [26] developed an incremental learning approach that enables deep neural networks within NILM systems to adjust to new devices, showcasing the direction toward more adaptable NILM solutions.

Predictive Maintenance. Here the aim is to detect unusual device behavior early to preempt potential malfunctions. Zangrando et al. [31] tested anomaly detection methods, finding OC-SVMs, Isolation Forests, and LOF to perform best. Azizi et al. [3] introduced a framework that evaluates the deviation of an event's features from known categories. If it exceeds a certain threshold, the system assesses anomaly frequency to identify if it signals a malfunction or a new, unmonitored device. This method combines anomaly detection with classification, boosting predictive maintenance's reliability.

Novelty Detection. Research on real-time new device detection has predominantly focused on eventless NILM systems. Zhang et al. [32]

used a neural network ensemble to evaluate classifier outcomes against expected performance and identify new devices through any shortfall. Welikala et al. [30] investigated device spectral signatures, employing maximum a posteriori estimates for device combinations and threshold comparison to spot new appliances. Similarly, Guo et al. [12] leveraged device signature templates for new appliance detection. Majumdar et al. [20] showed sparse coding's effectiveness in integrating new devices into dictionary coding frameworks without losing accuracy. Han et al. [13] utilized conditional generative adversarial networks for unknown device detection, marking a significant contribution to NILM's research development in identifying new devices. Lastly, de Baets et al. [9] used siamese neural networks to classify the VI-trajectories of appliances and decide based on clustering whether an appliance was known.

Gap. Addressing unknown devices is key for NILM's accurate power consumption analysis. Current NILM studies focus on eventless systems [12, 13, 20, 32], overlooking event-based NILM's benefits, such as less need for model tuning, crucial for wide-scale deployment. Our work fills this research gap by introducing a method to filter events from unknown devices in event-based systems, significantly advancing event-based NILM's practical application.

3 DESIGN

To tackle the challenges posed by unknown devices in event-based NILM, we outline our fundamental design concepts and explore various strategies for their implementation. Traditional event-based systems indiscriminately detect events from both known and unknown devices, thereby often leading to incorrect classifications and, subsequently, an elevated error rate in energy consumption estimation. To mitigate this issue, we propose modifications to the typical event-based NILM pipeline, aiming to enhance its accuracy.

3.1 Pipeline Adaptation

We explore various methods to prevent events from unknown devices from leading to inaccurate energy estimations. As shown in Figure 1, a potential intervention point is the event detection phase. By meticulously fine-tuning this stage, it may be feasible to ensure that only events from known devices proceed to the classification step. However, this approach necessitates extra threshold setting and specialization for the event detection process, potentially compromising its flexibility and adaptability to changes in the environment. We contend that a classifier designed to specifically filter out unknown events could be more effective than a custom event detector.

Alternatively, addressing the issue at the classification stage represents a more viable approach. This could involve implementing distance or confidence metrics with defined thresholds, which events must meet to be classified as belonging to a known device. This strategy has shown potential in eventless NILM systems as well, as evidenced by previous research [3, 12, 30, 32]. Essentially, setting a post-classification threshold introduces a new category (*unknown*) to the classifiers. Events classified under this category would then be excluded from the energy disaggregation process. This method could be further refined with the use of ensembling techniques [32]. A key benefit of this approach is its simplicity; it does not require adding a new component to the existing pipeline. Instead, it necessitates modifying the classifiers to include an additional confidence or distance measure alongside the device label.

Finally, introducing an independent filtering step between event detection and classification is a viable option. This intermediary filter would block events from unknown devices from proceeding to the classifier and, consequently, to the energy disaggregation phase. Possible implementations for this filtering step include unsupervised clustering algorithms or ML models tailored specifically for this purpose, such as LOF. This offers the benefit of further modularizing the NILM pipeline, thereby simplifying the process of swapping, adjusting, and optimizing various components. Additionally, a dedicated filter specifically designed for this task could potentially be more effective than an implicit filtering process integrated within the classification step. This distinct approach allows for targeted and efficient exclusion of events from unknown devices.

Filter selection. When evaluating the options for filtering out unknown events in NILM systems, it becomes clear that modifying the event detection process involves considerable implementation challenges. Although its effectiveness might be comparable to a separate filter step or a confidence-based classifier with thresholds, the overhead associated with this approach diminishes its feasibility for widespread application. Consequently, our research focuses on the latter options: implementing a separate filter step, utilizing a threshold-based confidence classifier, or a combination of both.

In our experimental setup, we only implement parts of the NILM pipeline up to the point of the respective filtering step, i.e., an event detection with a subsequent filter that may be stand-alone or integrated as confidence-threshold into a classifier. The task of optimizing classifiers and energy disaggregation algorithms that leverage the filtered events is already a well-explored area of research. Therefore, directing our efforts towards these aspects would likely yield limited additional insights. By concentrating on the filtering process, we aim to contribute novel solutions to the field of NILM, enhancing the overall system efficiency and accuracy.

3.2 Filtering Methods

To evaluate the efficacy of various filtering approaches, we commence with simple, intuitive methods using confidence-based classifiers, gradually progressing to more sophisticated solutions. Additionally, we incorporate the preliminary steps of an event-based NILM system and test all configurations using publicly accessible datasets. This section details our selection process and considerations for the different filtering strategies.

Dataset Selection. For an effective evaluation of filter performance, selecting a suitable dataset is essential. The ideal dataset should encompass a diverse range of typical household devices and capture their natural usage patterns. This diversity allows for the exclusion of specific devices during monitoring, closely replicating real-world deployment scenarios. Additionally, high-quality labeling or measurements of individual appliances, alongside aggregate household data, are necessary to validate the accuracy of the classifier and filter. While oversampling can increase the number of training samples,

the dataset should contain at least 10 events per device to capture the natural variability of device behavior. Moreover, a high sampling rate of at least 1 kHz is crucial for accurate event detection [15] and facilitates the use of harmonic-based features.

Considering these criteria, we selected the FIRED [28] dataset that monitors 66 appliances in a two-person household. The dataset offers measurement frequencies of 2 kHz and 8 kHz for isolated and aggregate measurements, respectively. Spanning 101 days, FIRED provides ample data for multiple devices, each with 10 or more events, and reliable labeling over a substantial timeframe.

Event Detection. We utilize the aggregated power data from the FIRED dataset for event detection due to its additive nature. The assumption is that changes in a device's power consumption during operational state transitions are observable in the aggregated readings, allowing for the estimation of event occurrences solely from time series analysis of power data.

Our event detection algorithm employs a probabilistic approach focusing on the identification of maxima and minima as potential events. We have refined this method based on the work of previous researchers [23, 27] that demonstrated notable enhancements compared to basic mean and median filtering techniques. The algorithm, rooted in Log Likelihood Ratio tests, consists of two primary components: detection statistics and detection activation.

Detection statistics calculate the likelihood of a change in power mean across pre- and post-event windows for a given sample x using the formula:

$$ds(x) = \frac{\mu_1 - \mu_0}{\sigma^2} \cdot \left| P(x) - \frac{\mu_0 + \mu_1}{2} \right|$$

where μ_0 and μ_1 denote the mean of the pre- and post-event windows respectively, σ^2 is the variance over the combined windows and $P(x)$ is the power value of the given sample. The detection statistics is set to 0 in a post-processing step if the absolute difference between μ_0 and μ_1 does not surpass a predefined threshold P_{thr}. The *detection activation* is responsible for actual event detection by sliding an extrema locator window over the series of detection statistics, identifying maxima and minima as potential events. These are considered as indicative of state changes of individual appliances.

Model Selection. Various methods exist for integrating an event filter into an event-based NILM pipeline. We have selected several approaches for each method to assess their effectiveness.

Initially, we consider confidence or probability-based classifiers with a threshold to exclude events from unknown devices. For this purpose, we chose SVM, LR, and RF due to their reliable confidence measures in multiclass problems. Additionally, we explored ensembling techniques known for their robustness in multiclass scenarios. Specifically, we implemented voting, stacking, and a mixture of experts (MoE) using SVM, LR, and RF classifiers. In the voting ensemble, we employed soft voting by aggregating probabilities across each class, with the highest total indicating the class assignment. Stacking involved using SVM, LR, and RF as base models, with their outputs fed into a level-1 RF that learns to combine these scores. The MoE approach utilized a binary RF for each tracked device, comparing their output probabilities. For both voting and stacking, an event is assigned to a known device if the highest confidence score exceeds a threshold. Similarly, in MoE, if any expert exceeds a threshold in its probability score, the event is classified as known.

Secondly, we combined confidence-based classifiers with unsupervised clustering techniques. This strategy involves a parallel classifier outputting a device cluster likely associated with the event, and subsequently dismisses the confidence scores of known device classes not present in the cluster. We applied k-Means for prior clustering (PC) and integrated it with SVM, RF, LR, and the voting and stacking ensembles.

Thirdly, we explored filtering algorithms separate from the classification process, typically used in anomaly detection. We selected OC-SVM and LOF, with events predicted as outliers excluded from subsequent classification. Furthermore, we investigated these filters' performance when used as classifiers themselves: Given their binary output, we created MoEs with one model per device for both OC-SVM and LOF. To our knowledge, this is the first instance of applying OC-SVM-MoE and LOF-MoE in NILM for device classification or detection of unknown devices.

Feature Selection. The operation of both the event-based classifiers and the filtering techniques, including clustering and density-based filters, necessitates a well-defined set of input features. In designing our feature vector, we drew upon existing research to balance low dimensionality with effective classification.

We selected Active Power (P) and Reactive Power (Q) as one-dimensional features, being established both as individual and combined features in prior studies [17]. Additionally, features derived from harmonics have been consistently reported to perform robustly in NILM tasks [16, 17]. To incorporate this aspect, we included Tristimulus [17, 22] in our feature set, a condensed representation of harmonics information. This addition has demonstrated success as a standalone feature [29], contributing three more dimensions.

With a total of five dimensions, the feature vector maintains a low computational complexity, facilitating efficient model processing. It has been shown [4] that this feature set achieves comparable results to a comprehensive feature selection on other real-world household datasets [17]. However, we acknowledge that the exploration of an optimal feature set remains a topic for future research and is beyond the scope of our current study. Notably, the system currently relies on amplitude changes alone and does not consider contextual information, such as the timing between amplitude rises and drops.

4 IMPLEMENTATION

Having established our preliminary choices, we now delve into the details of implementing various configurations of the modified NILM pipeline. Our discussion includes the data extracted from the dataset, the considered devices, and any data augmentations employed. Additionally, we elaborate on the training procedures for the different ML models utilized in our experimental setups. This comprehensive approach ensures a thorough understanding of the methodologies applied and the rationale behind our decisions.

Data Volume. As FIRED [28] partly offers precise labeling, we hence focus our analysis on this timeframe to benefit from the detailed metadata provided. The devices in FIRED are distributed across three electrical phases, $L1-L3$, and remain on the same phase throughout the labeled time period. While a device can be switched to a different phase by plugging it into another socket during deployment, the NILM system monitors all phases simultaneously, ensuring that a device is recognized regardless of its phase. To focus on devices with a high frequency of events, we analyzed the data for all three phases using our event detection algorithm over this time period. The results showed that phase $L3$ had by far the most detectable events—over seven times more than $L2$ and more than fifty times more than $L1$. Consequently, our study concentrates on data from the frequently used devices on phase $L3$.

Device Selection. As mentioned earlier, a minimum of 10 events per device is essential to ensure diversity in our training data for each class. This criterion applies to all devices, whether they are part of the training or testing set, facilitating a more flexible configuration of known and unknown devices. Additionally, this threshold aids in discerning statistically significant behavioral patterns.

Consequently, we tasked our event detector with analyzing the entire span of phase $L3$ data. Detected events were then aligned with the corresponding ground truth where feasible. Devices connected to Powermeter 11 were excluded, partly because they were not used during the labeled time period, i.e., they offered no reliable ground truth, and partly because their power consumption was below 5 W. Events with such low power ratings were not identifiable with our configuration of the event detector. However, with an average power consumption of about 250 W at the smart meter, their contribution to the overall power consumption would be below 2%, which was deemed negligible during our initial investigations. The results of this analysis are presented in Table 1, showing the final selection of devices and their respective event counts. Our selection includes both low (<200 W) and high-power consumption devices.

For these events, we computed features using the isolated measurements of each device. We selected a Region of Interest (ROI) spanning a 2 s window starting from each event. The feature calculation was conducted at a sampling rate of 2 kHz, which adequately supports the harmonics-based Tristimulus feature.

Data Augmentation. The data distribution in Table 1 reveals a significant skew in the number of events among different devices. Such an imbalance can lead to biased outcomes in machine learning models, particularly those based on clustering and density, as they might be influenced disproportionately by the majority classes.

As a countermeasure, resampling techniques are typically employed. These include undersampling the majority classes or oversampling the minority ones. For our study, we chose oversampling the minority classes, aiming to introduce greater diversity into both the training and testing datasets. Moreover, undersampling the majority classes would have reduced the number of test cases per class, thereby diminishing the statistical significance of our evaluations.

To achieve an equitable distribution of samples across all classes, we utilized the SMOTE [7] technique, which is a well-regarded method in various application domains. As a consequence, each class is adequately represented, enhancing the robustness and reliability of our model training and evaluation processes.

Table 1. Devices [28]

Device	Events	Power
Espresso Machine	1410	High
Fridge	559	High
Oven	97	High
Kettle	23	High
Coffee Grinder	95	Low
Fume Extractor	30	Low
TV	14	Low
Kitchen Spotlight	12	Low

Data Split. To effectively train our filter configurations, we divided the devices into two categories: Known and unknown. The data from unknown devices is withheld during training and exclusively used for testing. In contrast, for known devices, we allocate 67% of the samples to training, using a 2:1 split ratio.

The supervised multi-class classifiers, namely RF, LR, and SVM, along with k-Means for prior clustering and the components of the voting and stacking ensembles, utilize the entire set of training data. However, for the MoE ensembles, each binary classifier is trained differently: Every expert receives the complete set of training samples for its corresponding device as positive examples. For negative samples, the training data of all known devices are evenly distributed across the other experts. Specifically, for N known devices, an expert is trained with all the training samples of one known device as positive instances and a combined $(N-1)^{th}$ portion of the training samples of each other known device as negative instances. This approach aligns with scenarios where training is limited to a specific subset of devices. It also ensures that the negative samples for each expert are distinct and that there is an equal balance of positive and negative samples for training each expert.

Training. To train the filters, we optimized them using 5-fold cross-validation combined with RandomizedSearchCV for hyperparameter tuning. Details regarding the hyperparameter tuning, including the search spaces and results, are presented in Table 2. In addition, the thresholds for the confidence values were fine-tuned using a binary search method. We employed a knee-criterion based on the recall for the class of known devices, setting the threshold at a point where identification of almost all known device instances was still possible while maximizing the exclusion of unknown devices. This approach stems from our main objective for the filters to primarily eliminate events from unknown devices.

For the prior clustering approach, we fixed the number of clusters at 2, given the relatively small total number of known devices. The number of necessary clusters correlates strongly with the amount of known devices and their respective number of distinguishable states, with two clusters achieving best results in our settings.

Table 2. Hyperparameters

Model	Parameter
SVM	$C \in \{10^{-2}, 10^{-1}, \mathbf{10^0}, ..., 10^3\}$
	$\gamma \in \{10^4, 10^3, ..., \mathbf{10^{-2}}\}$
	$kernel \in \{\mathbf{RBF}\}$
LR	$C \in \{10^{-3}, 10^{-2}, ..., \mathbf{10^3}\}$
	$penalty \in \{l1, \mathbf{l2}\}$
RF	$n\_estimator \in \{10, 50, 100, \mathbf{1000}\}$
	$max\_depth \in \{10, 20, ..., \mathbf{90}, ..., 120\}$
	$min\_samples\_split \in \{2, \mathbf{6}, 10\}$

5 EVALUATION

After detailing our filter implementations' methodology, we outline scenarios to test their effectiveness, featuring mixes of known and unknown devices for specific assessment aspects. We first cover the evaluation metrics used, then describe the evaluation scenarios, and, finally, share and discuss our findings.

5.1 Evaluation Metrics

To thoroughly evaluate our filters' effectiveness in accurately excluding events from unknown devices, we have implemented several constraints within the NILM pipeline. Each filter configuration employs the same set of identified events to ensure consistency, thereby minimizing variability that might arise from the event detection stage and making it easier to directly associate the outcomes with the filter setup in use. Additionally, we assess the performance of each filter immediately after its application, bypassing any further steps such as classification or energy disaggregation. This decision is made to avoid the complexity of evaluating the combined effectiveness of multiple components in the pipeline, which would complicate attributing performance changes exclusively to the filter. Given the significant impact that filtered events can have on subsequent stages of the pipeline, our evaluation focuses on the precision with which events are processed by the filters.

To this end, we employ several standard metrics: Precision, Recall, F1-Score, and Accuracy, whereby the discussion of results is mostly based on the former two as they provide more detailed insight into the behavior of the models. The latter (F1-Score and Accuracy) are instead drawn upon for comparison to other approaches in the literature. In our analysis, the class of unknown devices is treated as the positive class, while known devices are the negative class, aligning with our task of filtering out events from the positive class.

5.2 Scenarios

We consider four scenarios to evaluate our filters, each designed to test different aspects of their performance under various conditions.

- **Scenario 1 - Balanced Device Split:** Eight devices are divided into equal groups of known and unknown, including both high and low power consumers and multi-state appliances. This setup, with a relatively high number of unknowns, aims to mimic real-world conditions more closely than typical novelty detection setups in NILM, which often test only a couple of new devices [9, 12, 20, 30].

Table 3. Composition of devices in each scenario

Category	Scen. 1	Scen. 2	Scen. 3	Scen. 4
Known	E. Machine K. Spotlight Oven Fume Extr.	Fridge Kettle TV C. Grinder	E. Machine K. Spotlight Oven Fume Extr. C. Grinder	Fridge Kettle TV
Unknown	Fridge Kettle TV C. Grinder	E. Machine K. Spotlight Oven Fume Extr.	Fridge[1] Kettle[1] TV[1]	Fridge[1] Kettle[1] TV[1]

[1] = Device from UK-DALE [18]

- **Scenario 2 - Reversed Split Evaluation:** This scenario swaps the known and unknown device groups to test the robustness of our findings and confirm they are not due to chance. It acts as an initial test for filter effectiveness, with the potential for more complex combinations to be explored.
- **Scenario 3 - Cross-Domain Challenge with New Devices:** Known devices from one dataset are tested against unknown devices from another, assessing filter performance in a new domain and simulating deployment in a different household. This specific use case does not appear to have been addressed in previous work on detecting new devices.
- **Scenario 4 - Matching Device Types Across Domains:** This tests filter adaptability to domain shifts while keeping device types constant, examining if filters can recognize known devices despite changes in domain or if cross-domain signature differences hinder detection. The cross-validation on novelty detection referenced in [9] was performed on houses within the same electrical grid. In contrast, we chose to use a dataset from a completely different domain.

Table 3 details the known and unknown device sets per scenario. As clearly shown by the power distribution of the devices (Figure 2), the unknown devices can easily be mistaken for known devices based solely on power consumption—e.g., in Scenario 1, the fridge could be confused with an espresso machine, the kettle with an oven, and the TV and coffee grinder with a fume extractor. Our analysis evaluates filters' effectiveness within the same domain (Scenarios 1 and 2) and in cross-domain contexts (Scenarios 3 and 4), expanding the typical evaluation scope of novelty detection in NILM.

5.3 Scenarios 1 & 2: Familiar Domain

Our analysis begins within the filters' trained domain, focusing on confidence-based classifiers. Despite distinct training datasets for Scenarios 1 and 2, we maintained a consistent decision boundary, leading to high precision in Scenario 1 across all models (SVM, LR, RF, and ensembles) but a notable precision drop for LR in Scenario 2, as shown in Figure 3a. This drop, potentially due to an unfit threshold or classification bias, prompts a deeper investigation into model-specific feature prioritization, revealing LR's preference of Tristimulus over Active (P) and Reactive (Q) Power, in contrast to RF's prioritization, as a likely cause for low performance.

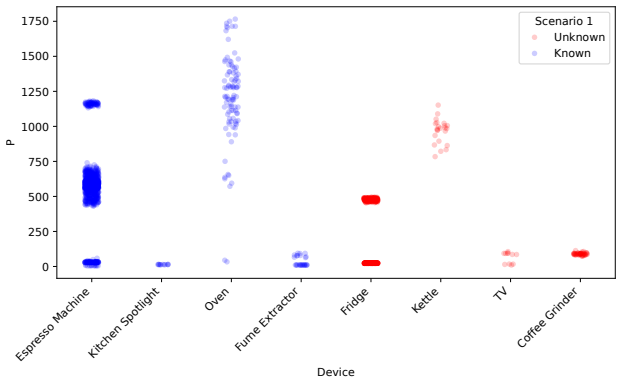
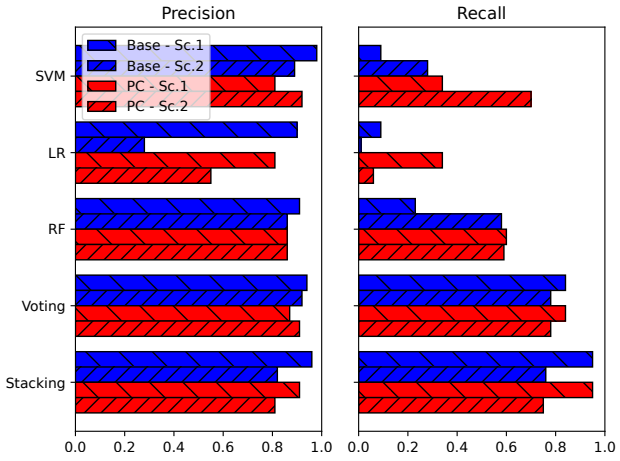
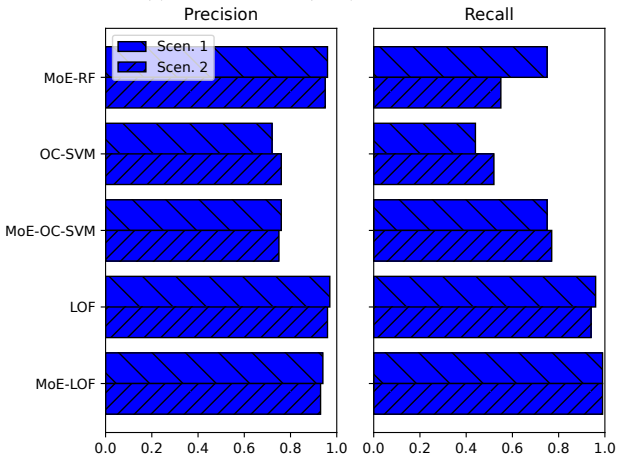


Fig. 2. The power distribution of chosen devices.



(a) Filters without (*base*) and with PC.



(b) MoE- and density-based Filters.

Fig. 3. Precision and recall of different filters

Single classifiers showed generally low recall, identifying less than 35% of events from unknown devices, except for RF exceeding 50% recall in one instance. Ensembling techniques improved robustness, detecting over 70% of unknown events without compromising precision, underscoring their potential in NILM systems.

PC slightly reduced precision but notably improved recall for single classifiers, enhancing their ability to identify unknown devices. However, ensemble methods did not benefit from PC, showing decreased performance in recognizing some known device states. Ensemble methods emerge as superior in maintaining precision-recall balance, showing their effectiveness for diverse scenarios.

Specialized filters like MoE ensembles and density-based models (OC-SVM and LOF) were explored; surprisingly, MoE did not perform as well as other ensembles. OC-SVM struggled with low precision and recall, failing to correctly identify several devices. Specifically, it failed to identify the fridge, kettle, and coffee grinder as unknown devices, while mistakenly classifying the kitchen spotlight and fume extractor as known devices. In Scenario 2, OC-SVM missed identifying the oven and espresso machine as unknown devices and incorrectly classified the TV, coffee grinder, and fridge as known. These results closely align with the division between high and low power consuming devices, as outlined in Table 1. The OC-SVM seems to struggle with creating a sufficiently nuanced decision boundary for effective binary classification across multiple devices. Even when used in an MoE configuration, while there was an increase in recall, the precision remained disappointingly low, indicating that OC-SVM's limitations are not effectively addressed by reducing the number of devices in the known device group. In contrast, LOF exhibited exceptional performance, with precision and recall rates exceeding 93%, misclassifying less than 7% of known device events as unknown. These findings hint at its significant potential for NILM applications, especially when integrated into MoE ensembles for robust filtering and classification across scenarios.

Validation. Our analysis identified MoE-LOF as the most reliable performer, with LOF as a strong contender. To further validate MoE-LOF, we conducted an additional evaluation across all detectable events for all devices and phases within the labeled timeframe of FIRED, employing a realistic setting without oversampling events. Using the same two sets of known devices from Scenarios 1 and 2, but introducing a wider variety of unknown devices, we achieved F1-scores of at least 93% in both cases. However, for the known models in Scenario 1, recall decreased to 87% due to the increased diversity of unknown devices. Importantly, the models were not tuned but were trained using consistent parameters.

5.4 Scenarios 3 & 4: Domain Transfer

Exploring domain transfer in NILM systems, our analysis targets understanding filter efficacy in unfamiliar settings, aiming to simplify NILM setup expansion.

Scenario 3: Unknown Devices. Preliminary tests on domain transfer highlight challenges in recognizing unknown devices from the UK-DALE dataset. Filters showed variable success, as seen in Figure 4a; notably, the LR classifier exhibited a 0% recall, indicating a complete inability to correctly identify events from the unknown devices. Other classifiers also struggled, particularly with the kettle and, to some extent, the fridge. The introduction of PC significantly

improved the recall for the kettle, achieving 100% across all classifiers, and ameliorated some difficulties with the fridge.

The MoE ensemble using RF and the stacking ensemble performed comparably, indicating potential benefits of applying PC to MoE configurations. OC-SVM underperformed, even when extended to an MoE setup. In contrast, LOF excelled, maintaining 100% recall, illustrating its robustness in cross-domain applications.

Scenario 4: Known Devices. Investigating the impact of domain shifts on recognizing known device classes revealed a stark performance variation (Figure 4b). Training filters on similar device events from the FIRED dataset altered recognition rates, with some filters achieving 100% recall. Notably, the addition of PC does not negatively impact the filters' performance with the kettle, an exception to the general trend of reversed performance.

Ensembling methods, however, seem to reduce the ability of filters to recognize devices as known in a new domain. Most strikingly, the LOF filter, while exceptionally sensitive and effective in its original domain, fails entirely in this domain transfer scenario, demonstrating a 0% success rate across all tests.

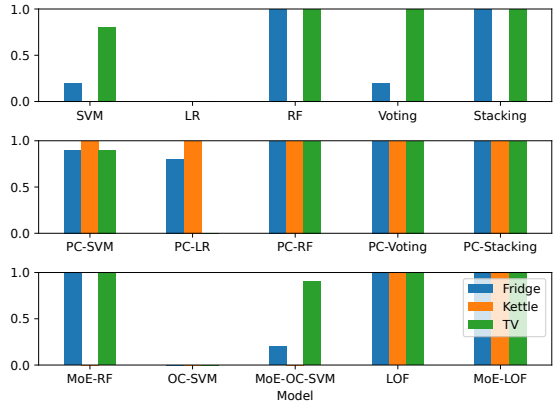
This analysis underscores the complexities associated with domain transfer in NILM systems. It reveals that while some filters adapt well to new environments, others, particularly those highly sensitive to specific device signatures like LOF, may require retraining or recalibration to maintain effectiveness.

5.5 Discussion

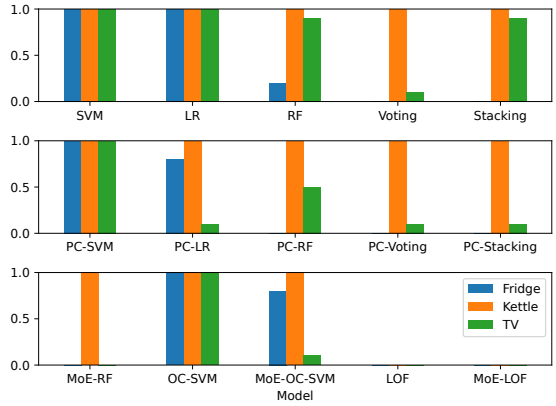
Our evaluation revealed key insights into the proposed filtering methods for events from unknown devices. Simple thresholding is found inadequate, regardless of the extent of hyperparameter tuning, suggesting its effectiveness is less reliant on these settings. Instead, the device mix significantly impacts performance. Device-specific issues, like distinguishing between different power levels and multi-state appliances, suggest the need for specialized models or classes for distinct device states.

Integrating a secondary clustering algorithm substantially enhances threshold-based filters, though improvements are device-dependent. Ensemble methods emerge robust across scenarios, yet adding PC analysis in this case shows minimal benefits. For density-based approaches, OC-SVM lacks versatility for diverse devices and does not improve with a MoE ensemble. Instead, it mainly discriminates between high- and low-power consuming devices. On the flipside, LOF displays promising accuracies, despite sensitivity to device signature variations, which might improve with more diverse data. With consistent F1-scores and accuracies above 93% in detecting multiple unknown devices, it shows potential to be the most promising prospect within a known domain when compared to the literature [9, 12, 13, 20, 30, 32], calling for a study with comparable data between the approaches.

LOF's efficacy in a MoE setup suggests potential for future research in device composition correlation, training sample requirements, and runtime adaptability. The prospect of MoE-LOF fulfilling both filtering and classification roles in the NILM process warrants further exploration. In future work, we plan to explore incorporating the filtered events directly into the classifiers to extend the system's ability to distinguish a larger number of devices.



(a) Unknown Devices



(b) Known Devices

Fig. 4. The recall of different filters for unknown (Scenario 3) and known devices (Scenario 4) in a new domain.

6 CONCLUSION

In addressing the challenge of accurately disaggregating energy consumption in dynamic household environments, our study introduces a novel filtering mechanism to the NILM pipeline, significantly enhancing its ability to differentiate between known and unknown device events. Through the innovative use of confidence-based clustering ensembles and density-based techniques, we achieved a notable reduction of over 93% in false positives from unknown devices, with a minimal loss of known device events (under 7%). This advancement increases NILM systems' accuracy and practical applicability in energy monitoring and conservation efforts.

Our findings underscore the potential of refined event detection and classification methodologies to improve the effectiveness of NILM technologies. By enhancing the system's ability to adapt to new and changing appliance signatures, we pave the way for more reliable energy consumption insights, contributing to the broader objectives of reducing energy waste and aiding in climate change mitigation.

REFERENCES

[1] Kyle D Anderson, Mario E Berges, Adrian Ocneanu, Diego Benitez, and Jose MF Moura. 2012. Event detection for non intrusive load monitoring. In *IECON 2012-38th Annual Conference on IEEE Industrial Electronics Society*. IEEE, 3312–3317.

[2] Georgios-Fotios Angelis, Christos Timplalexis, Stelios Krinidis, Dimosthenis Ioannidis, and Dimitrios Tzovaras. 2022. NILM Applications: Literature Review of Learning Approaches, Recent Developments and Challenges. *Energy and Buildings* 261 (2022), 111951.

[3] Elnaz Azizi, Mohammad TH Beheshti, and Sadegh Bolouki. 2021. Appliance-level Anomaly Detection in Nonintrusive Load Monitoring via Power Consumption-based Feature Analysis. *IEEE Transactions on Consumer Electronics* 67, 4 (2021), 363–371.

[4] Justus Breyer, Sparsh Jauhari, René Glebke, Muhammad Hamad Alizai, Markus Stroot, and Klaus Wehrle. 2024. Investigating Domain Bias in NILM. In *Proceedings of the 11th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. 333–336.

[5] K. Carrie Armel, Abhay Gupta, Gireesh Shrimali, and Adrian Albert. 2013. Is Disaggregation the Holy Grail of Energy Efficiency? The Case of Electricity. *Energy Policy* 52 (2013), 213–234.

[6] Xiaomin Chang, Wei Li, Chunqiu Xia, Qiang Yang, Jin Ma, Ting Yang, and Albert Y Zomaya. 2022. Transferable Tree-Based Ensemble Model for Non-Intrusive Load Monitoring. *IEEE Transactions on Sustainable Computing* 7, 4 (2022), 970–981.

[7] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[8] Suryalok Dash and NC Sahoo. 2022. Electric Energy Disaggregation via Non-intrusive Load Monitoring: A State-of-the-Art Systematic Review. *Electric Power Systems Research* 213 (2022), 108673.

[9] Leen De Baets, Chris Develder, Tom Dhaene, and Dirk Deschrijver. 2019. Detection of Unidentified Appliances in Non-intrusive Load Monitoring Using Siamese Neural Networks. *International Journal of Electrical Power & Energy Systems* 104 (2019), 645–653.

[10] Karen Ehrhardt-Martinez, Kat A Donnelly, Skip Laitner, et al. 2010. Advanced Metering Initiatives and Residential Feedback Programs: A Meta-Review for Household Electricity-Saving Opportunities. American Council for an Energy-Efficient Economy Washington, DC.

[11] Jessie M Gillis and Walid G Morsi. 2022. A Novel Flexible and Scalable Nonintrusive Load Monitoring Approach Using Wavelet Design and Machine Learning. In *2022 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*. IEEE, 441–445.

[12] Xiaochao Guo, Chao Wang, Tao Wu, Ruiheng Li, Houyi Zhu, and Huaiqing Zhang. 2023. Detecting the Novel Appliance in Non-Intrusive Load Monitoring. *Applied Energy* 343 (2023), 121193.

[13] Yinghua Han, Keke Li, Chen Wang, Fangyuan Si, and Qiang Zhao. 2023. Unknown Appliances Detection for Non-Intrusive Load Monitoring Based on Conditional Generative Adversarial Networks. *IEEE Transactions on Smart Grid* (2023).

[14] George William Hart. 1992. Nonintrusive Appliance Load Monitoring. *Proc. IEEE* 80, 12 (1992), 1870–1891.

[15] Jana Huchtkoetter and Andreas Reinhardt. 2019. A Study on the Impact of Data Sampling Rates on Load Signature Event Detection. *Energy Informatics* 2 (2019), 1–12.

[16] Matthias Kahl, Thomas Kriechbaumer, Anwar Ul Haq, and Hans-Arno Jacobsen. 2017. Appliance Classification Across Multiple High Frequency Energy Datasets. In *2017 IEEE International Conference on Smart Grid Communications (SmartGridComm)*. IEEE, 147–152.

[17] Matthias Kahl, Anwar Ul Haq, Thomas Kriechbaumer, and Hans-Arno Jacobsen. 2017. A Comprehensive Feature Study for Appliance Recognition on High Frequency Energy Data. In *Proceedings of the Eighth International Conference on Future Energy Systems*. 121–131.

[18] Jack Kelly and William Knottenbelt. 2015. The UK-DALE Dataset, Domestic Appliance-level Electricity Demand and Whole-house Demand from five UK Homes. *Scientific data* 2, 1 (2015), 1–14.

[19] Weicong Kong, Zhao Yang Dong, Jin Ma, David J Hill, Junhua Zhao, and Fengji Luo. 2016. An Extensible Approach for Non-Intrusive Load Disaggregation with Smart Meter Data. *IEEE Transactions on Smart Grid* 9, 4 (2016), 3362–3372.

[20] Angshul Majumdar. 2022. Disaggregating a New Appliance On-the-Fly Without Data Acquisition and Retraining. *IEEE Transactions on Instrumentation and Measurement* 71 (2022), 1–4.

[21] Stephen Makonin. 2016. Investigating the Switch Continuity Principle Assumed in Non-Intrusive Load Monitoring (NILM). In *2016 IEEE Canadian conference on electrical and computer engineering (CCECE)*. IEEE, 1–4.

[22] Geoffroy Peeters. 2004. A Large Set of Audio Features for Sound Description (Similarity and Classification) in the CUIDADO Project. *CUIDADO Ist Project Report* 54, 0 (2004), 1–25.

[23] Lucas Pereira. 2017. Developing and Evaluating a Probabilistic Event Detector for Non-intrusive Load Monitoring. In *2017 Sustainable Internet and ICT for Sustainability (SustainIT)*. IEEE, 1–10.

[24] JG Roos, IE Lane, EC Botha, and Gerhard P Hancke. 1994. Using Neural Networks for Non-intrusive Monitoring of Industrial Electrical Loads. In *Conference Proceedings. 10th Anniversary. IMTC/94. Advanced Technologies in I & M. 1994 IEEE Instrumentation and Measurement Technolgy Conference (Cat. No. 94CH3424-9)*. IEEE, 1115–1118.

[25] Nasrin Sadeghianpourhamami, Joeri Ruyssinck, Dirk Deschrijver, Tom Dhaene, and Chris Develder. 2017. Comprehensive Feature Selection for Appliance Classification in NILM. *Energy and Buildings* 151 (2017), 98–106.

[26] Giulia Tanoni, Emanuele Principi, Luigi Mandoliniy, and Stefano Squartini. 2023. Appliance-Incremental Learning for Non-Intrusive Load Monitoring. In *2023 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*. IEEE, 1–6.

[27] Benjamin Völker, Marc Pfeifer, Philipp M Scholl, and Bernd Becker. 2020. Annoticity: A Smart Annotation Tool and Data Browser for Electricity Datasets. In *Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring*. 1–5.

[28] Benjamin Völker, Marc Pfeifer, Philipp M Scholl, and Bernd Becker. 2020. FIRED: A Fully-labeled hIgh-fRequency Electricity Disaggregation Dataset. In *Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. 294–297.

[29] Benjamin Völker, Philipp M Scholl, and Bernd Becker. 2021. A Feature and Classifier Study for Appliance Event Classification. In *International Conference on Sustainable Energy for Smart Cities*. Springer, 99–116.

[30] Shirantha Welikala, Chinthaka Dinesh, Roshan Indika Godaliyadda, Mervyn Parakrama B Ekanayake, and Janaka Ekanayake. 2016. Robust Non-Intrusive Load Monitoring (NILM) with Unknown Loads. In *2016 IEEE international conference on information and automation for sustainability (ICIAfS)*. IEEE, 1–6.

[31] Niccolò Zangrando, Piero Fraternali, Marco Petri, Nicolò Oreste Pinciroli Vago, and Sergio Luis Herrera González. 2022. Anomaly Detection in Quasi-Periodic Energy Consumption Data Series: A Comparison of Algorithms. *Energy Informatics* 5, 4 (2022), 1–22.

[32] Jianjun Zhang, Xuanqun Chen, Wing WY Ng, Chun Sing Lai, and Loi Lei Lai. 2019. New Appliance Detection for Nonintrusive Load Monitoring. *IEEE Transactions on Industrial Informatics* 15, 8 (2019), 4819–4829.