Investigating Domain Bias in NILM

Justus Breyer
Chair of Communication and Distributed Systems
RWTH Aachen University
Aachen, Germany
breyer@comsys.rwth-aachen.de

Sparsh Jauhari
Chair of Communication and Distributed Systems
RWTH Aachen University
Aachen, Germany
sparsh.jauhari@rwth-aachen.de

René Glebke
Chair of Communication and Distributed Systems
RWTH Aachen University
Aachen, Germany
glebke@comsys.rwth-aachen.de

Muhammad Hamad Alizai
SBA School of Science and Engineering
LUMS
Lahore, Pakistan
hamad.alizai@lums.edu.pk

Markus Stroot
IAEW
RWTH Aachen University
Aachen, Germany
m.stroot@iaew.rwth-aachen.de

Klaus Wehrle
Chair of Communication and Distributed Systems
RWTH Aachen University
Aachen, Germany
wehrle@comsys.rwth-aachen.de

Abstract

Enhancing household energy efficiency is crucial, and Non-intrusive Load Monitoring (NILM) offers a valuable solution by giving consumers insights into their energy use without individual device monitoring. However, the deployment of NILM models in new settings is challenging due to their training on domain-specific data.

To effectively use public data for training NILM models for identifying individual appliances, understanding the challenges of model transfer is crucial. This study explores several factors that could hinder successful model transfer and highlights the challenges in broader NILM system deployment. We developed and tested various NILM models, both event-based and eventless, across multiple household domains and found that domain bias, e.g., noise and line frequency, does not significantly impact model performance.

CCS Concepts

• **Hardware** → **Energy metering**; • **Computing methodologies** → *Supervised learning*; **Cross-validation**.

Keywords

machine learning, non-intrusive load monitoring (NILM), energy disaggregation, load signatures, model transfer

ACM Reference Format:
Justus Breyer, Sparsh Jauhari, René Glebke, Muhammad Hamad Alizai, Markus Stroot, and Klaus Wehrle. 2024. Investigating Domain Bias in NILM. In *The 11th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BUILDSYS '24), November 7–8, 2024, Hangzhou, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3671127.3699532

BUILDSYS '24, November 7–8, 2024, Hangzhou, China

ACM ISBN 979-8-4007-0706-3/24/11
https://doi.org/10.1145/3671127.3699532

1 Introduction

Climate change necessitates lowering greenhouse gas emissions which calls for reducing energy consumption across various sectors. Consumers are more inclined to save energy when they have access to detailed information about their consumption patterns [2].

This information can be provided by installing meters at every single appliance, imposing extensive maintenance burdens on the user. In contrast, nonintrusive load monitoring (NILM) offers a promising alternative by analyzing aggregate energy consumption data to infer the usage of individual appliances. This approach relies on machine learning (ML) models to disaggregate the total household energy consumption into appliance-specific insights [5]. However, the diversity of household appliances, coupled with the vast amounts of data required for training these ML models, presents significant challenges. This complexity makes it difficult for average consumers to implement NILM techniques without substantial external support. Facilitating the adoption of NILM technologies could be significantly enhanced by the availability of shared appliance consumption data and pre-trained models through public repositories. While there is a variety of datasets available for NILM research and the benchmarking of proposed ML models (e.g., [8, 12]), the task of transferring ML models across various domains often results in marked performance declines (e.g. [6]). Consequently, there has been a growing focus on research aimed at reducing the need for model adaptation when deploying in new environments [1, 3, 4]. This study delves into the transferability challenges of NILM models, aiming to uncover the underlying issues and propose solutions to enhance their applicability and effectiveness in real-world settings.

Contributions.

- We developed NILM models using publicly available data, aimed at unfamiliar environments without dataset-specific adjustments.
- We established a novel lab testing methodology to simulate real-world conditions with various devices, evaluating NILM models' performance in an environment akin to a typical household. These tests confirmed the models' ability to identify device types accurately, highlighting challenges with devices of significantly different signature curves and low-power devices.
- We performed "in-the-wild" testing using data from different domains to assess NILM models in real-world conditions. These

tests revealed consistent model performance and minimal domain bias, underscoring the feasibility of deploying pre-trained NILM models in new environments.

Through a comprehensive examination of various factors that could potentially impact the performance of NILM models, this study concludes that domain bias, with the exception of variations in device signatures, is not a major barrier to model transfer. This insight fosters optimism regarding the wider adoption of NILM as an effective method for monitoring individual appliances.

2 Related Work

NILM research, with a history spanning over three decades, focuses on the distinct characteristics of appliance classes during startup and operation, enabling device classification with high accuracy. The release of various datasets [12], encompassing controlled laboratory measurements and real-world residential data, has sharpened the focus on algorithmic challenges and enhanced comparability.

Yet, the adaptation of machine learning models beyond their training domains—a key for wider deployment—has been sparingly addressed. This oversight is critical given the effort required to gather comprehensive data for each consumer setting, underscoring the importance of model generalization and transferability.

Kahl et al.'s investigation into model generalizability evaluated feature performance across different contexts to identify universally effective features [7]. The exploration of NILM model transferability has seen the application of various transfer learning techniques, including decision trees [4], generative adversarial networks [1], and sequence-to-point models [3, 10], aiming to enhance model applicability in unknown environments. Despite these advancements, achieving effective model transfer typically necessitates some degree of model retraining, fine-tuning, or adaptation [9].

The observed decline in model performance during initial attempts at unadjusted model transfer [6, 9] underscores the need for strategies to enhance NILM model adaptability. Our research aims to systematically examine these performance challenges to identify viable solutions, thereby facilitating NILM's broader deployment.

3 General NILM Pipeline & Configuration

To deploy an NILM setup, multiple steps need to be considered. There have been two main approaches to this problem: *event-based* and *eventless* NILM, differing in the necessary components.

Data collection is crucial for both NILM approaches, leveraging sources like smart meters to gather consumption data. This data supports model training pre-deployment and informs energy disaggregation post-deployment. Understanding how data collection from one environment translates to another is key in NILM model transferability research. **Event detection**, specific to event-based NILM, identifies appliance state changes (e.g., on/off) through variations in power consumption. It requires higher sampling rates to prevent simultaneous event occurences. **Feature calculation** in event-based NILM involves analyzing data around detected events to differentiate appliances or states, often using data transformations like Fourier or wavelet transforms, which has been extensively reviewed [6, 7]. **Classification** uses the extracted features to identify the appliance or its state using ML models, ranging from simple algorithms like RF or SVM [7] to more complex neural networks [13]. **Energy disaggregation**, the final step, varies between approaches. Event-based NILM matches events to appliance states and their expected consumption for power estimation, whereas eventless NILM directly estimates power from collected data, bypassing event detection and feature calculation.

Table 1: Considered Devices of FIRED [12]

Device	Events	Consumption	Duration
Espresso Machine	1074	High	Fixed
Fridge	318	High	Fixed
Coffee Grinder	172	Low	Manual
Oven	132	High	Manual
TV	23	Low	Manual
HiFi-System	16	Low	Manual
Kettle	15	High	Fixed

Designing an NILM setup is a complex task, requiring careful consideration of various factors to achieve optimal performance, particularly when intending to test in unknown environments. To examine the factors affecting NILM system performance in new consumer settings, we have segmented the NILM pipeline into several modular components. Using insights from prior research, we trained various pipeline configurations on a publicly available dataset to achieve satisfactory results. We detail the different configurations of our NILM pipeline and explain our choice of dataset.

Dataset Selection. We selected the FIRED Dataset [12] for the training and validation data of our study. It covers 66 residential appliances, including 21 with isolated monitoring, and offers high sampling rates. Additionally, its precise labeling guarantees ample data for training. Being a recent dataset from 2020, FIRED's appliance mix is likely reflective of current consumer use.

Device Selection. Our device selection (cf. Table 1) includes devices with both high power (> 200 W) and low power consumption. This enables us to investigate whether the power consumption of devices has an impact on the classification performance. Moreover, we included a mix of automatic devices and devices that depend on user behavior, with varying durations of operation, to consider possible effects on the classification performance.

Model Selection. We chose to train and evaluate ML models from both the eventless and event-based NILM domains, allowing for representative comparison with regards to transferibility. In the event-based NILM domain, we selected four different models that have been widely used and studied (e.g. [6, 7]): SVM, kNN, RF, and XGBoost, the latter being an ensemble learning method. For eventless NILM, we opted for the seq2point approach, particularly for its potential for model transfer [3, 10]. To construct the underlying NN of our seq2point model, we followed related work [13].

Feature Selection. In the event-based approach, the ML models require a feature vector as input. To ensure scalability, we aim to keep the dimensionality at a minimum, possibly trading off classification performance. Active and reactive power do not only serve as common choices but also demonstrate strong performance as standalone features [7]. Additionally, features incorporating information about higher-order harmonics tend to perform best in cross-dataset

validation [6, 7]. Hence, our feature vector encompasses the three dimensions of the Tristimulus (Tri) [7] as well as active (P) and reactive (Q) power, resulting in a total of five dimensions.

Pipeline Reduction. To focus on the foundational aspects of model transfer, we simplify the event-based NILM pipeline by excluding two components. Firstly, *event detection* will be replaced by utilizing ground truth data, as the detection is separately optimizable. Secondly, we omit *energy disaggregation* due to its reliance on precise event classification. Misaligned start or end times can significantly skew power estimates, hence an energy estimation error could result from either the classification or the disaggregation step.

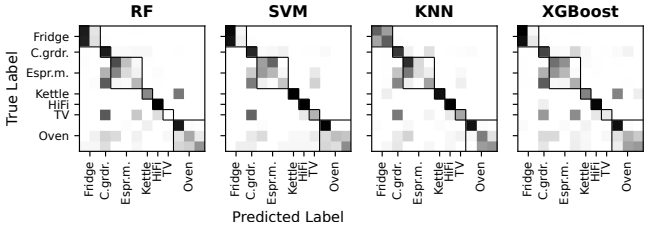

Figure 1: The performance of state-based models trained on the isolated and tested on the aggregate data of FIRED.

4 Model Training & Validation

Event-based models (i.e., supervised ML) necessitate labeled ground truth events, using isolated measurements from a specific period in 2020 for training. Conversely, the eventless approach, leveraging seq2point models, used FIRED's isolated, unlabeled readings as ground truth, allowing for increased training data volume. Seq2point required downsampling data for longer input sequences.

Applying z-score normalization was pivotal for both model types, aligning with established practices that suggest normalized data significantly benefits neural network training and standardizing feature inputs based on their statistical distribution removes biases.

The event-based models were trained for both device-based and state-based classification. Training and validation data were created using an 80-20 split. The event-based models were tuned utilizing GridSearchCV. For the eventless models, early stopping was applied and the ADAM optimizer was used to mitigate overfitting.

Model Performance. Upon validating our models with the FIRED dataset's aggregated measurements, we observed comparable scores of device-based models (up to 80%) to other approaches using similar models [7]. The reduced scores of state-level classification were mostly caused by inner-device confusion (Figure 1). The eventless approach on the other hand generally produced a higher MSE for high consuming devices than for low consuming devices.

Model Selection. For further evaluation, we chose the best performing out of the supervised models after a comprehensive feature validation across all possible feature subsets, which was kNN with the full feature set. Notably, our hyperparameter configuration ($k = 4$, using city block distance metric) matched the findings of related work for cross-dataset classification [6], that found this configuration to work best in scenarios without model adaptation.

5 Lab Testing

To investigate the impact of device variations and domain differences, we firstly simulate an environment reflective of the FIRED dataset's domain in a lab. During our test runs, we collected data from all device types known to our models, with the exception of the oven. We managed to obtain the same model of a coffee grinder as in FIRED, while all other device types were different models.

Scenarios. We devised three test scenarios to replicate typical user behavior with multiple devices operating simultaneously, based on patterns from the FIRED dataset. Each scenario was conducted five times to factor in timing variations and device behavior inconsistencies. Fridge events were excluded due to insufficient support.

Data Acquisition. Our lab setup mirrored a standard household configuration except for the noise. A class A 100 kW linear power amplifier, free from external noise, supplied a clean three-phase 230 V_{RMS} 50 Hz signal, replicating an optimal European low-voltage power grid. Voltage and current were measured using a power analyzer with a 20 kHz sampling frequency. For comprehensive scenario testing, all devices were connected to a single phase.

Similarities between lab and FIRED's devices were noted for the TV and coffee grinder in P, with a notable mismatch in Q for the TV. Further, the P signatures of the Hifi-Systems did not match.

Evaluation. The lab data was downsampled accordingly to match each model: 2 kHz for the kNN and low seconds for the seq2point models. For the seq2point models, we assessed device usage predictions (above 10 W) against ground truth, categorizing overlaps as True Positives (TP), missed detections as False Negatives (FN), and incorrect predictions as False Positives (FP).

The kNN models showed F1-scores of 50.73% and 57.87% for appliance and state classifications, respectively. Misidentifications, such as the TV being confused with a coffee grinder, were similar to patterns from the analysis on FIRED. The seq2point approach demonstrated challenges in identifying low-power devices, failing to recognize the Hifi-System, TV, and coffee grinder in all instances, as outlined in Table 2. This discrepancy, particularly with devices that share signatures with FIRED counterparts, suggests issues with model sensitivity to operation duration and power thresholds.

Conclusions. By managing external factors, such as line noise and frequency, the models overall exhibit a commendable capability to identify similar device types. Although we see a slight dip in performance, most of the difficulties in the new environment originate from devices having significantly distinct signature curves. This issue can be addressed by using an extended feature set and expanding the inner-class diversity of devices during training.

6 In-the-Wild Testing

While our lab environment emulates ideal conditions, it is clear that these conditions do not fully capture the complexity of real-world NILM model deployment. Therefore, we utilized segments of real household measurements from diverse domains for evaluation.

Dataset Selection. We selected the SustDataED2 [11] and the UK-DALE [8] datasets. Both datasets offer ground truth in the low seconds range and aggregate data above 12 kHz. We downsampled the data to match the rates to which our models were calibrated.

Evaluation. Table 2 depicts the overlap of devices between our models and the datasets, as well as their corresponding scores.

Table 2: Precision/Recall per Considered Device of the FIRED [12], Lab, UK-DALE [8] and SustDataED2 [11] Datasets.

		Fridge	C.Grinder	Esp.Machine	Kettle	HiFi-System	TV	Oven	Average
FIRED	kNN (App.)	0.96/0.99	0.49/0.76	0.96/0.84	0.88/1.0	0.8/1.0	0.25/0.65	0.95/0.84	0.76/0.87
	kNN (State)	0.61/0.63	0.45/0.78	0.71/0.63	0.88/0.93	0.76/1.0	0.26/0.35	0.79/0.7	0.64/0.72
Lab	kNN (App.)	0.0/0.0*	0.23/0.6	0.56/1.0	1.0/0.8	0.0/0.0	0.6/0.6	-	0.48/0.60
	kNN (State)	0.5/1.0*	0.29/0.8	0.83/1.0	1.0/0.6	1.0/0.1	0.67/0.6	-	0.76/0.62
	seq2point	0.38/1.0	0.0/0.0	0.32/0.6	1.0/0.92	0.0/0.0	0.0/0.0	-	0.28/0.42
SustDataED2	kNN (App.)	1.0/0.4	-	0.56/1.0	1.0/0.7	-	[1]	[1]	0.85/0.70
	kNN (State)	1.0/0.5	-	0.48/1.0	1.0/0.2	-	[1]	[1]	0.83/0.57
	seq2point	0.53/0.6	-	1.0/0.03	0.92/0.15	-	1.0/0.66	0.32/0.39	0.75/0.37
UK-DALE	kNN (App.)	0.67/0.2	-	0.56/0.9	0.0/0.0	-	0.0/0.0	-	0.31/0.28
	kNN (State)	0.67/0.4	-	0.53/0.9	0.0/0.0	-	0.0/0.0	-	0.30/0.33
	seq2point	0.55/0.34	-	0.65/0.13	0.34/0.29	-	0.92/0.55	-	0.62/0.33

* = The support of these devices was too low, so their scores were excluded from the averages

[1] = The events of these devices could not be recognized due to low/no change in power consumption

We excluded the TVs and the stove-oven from the **SustDataED2** dataset in the kNN analysis as they offered few detectable events. Moreover, the fridge-freezer and stove-oven in the SustDataED2 dataset differ in their operational purpose from the types of devices we were previously detecting, altering their signatures. However, we found that the kNN performed as well on the fridge-freezer as on the fridge from other domains, yielding a precision of 100%. Similarly, both espresso machine and kettle were accurately identified, with the former achieving 100% recall for both models.

Since the seq2point model is not reliant on event detection, it had a wider array of devices to classify. Interestingly, the low-power consuming state of the stove-oven went unrecognized, while the high power consuming state was detected. However, the high-power devices generally struggled with a high volume of FPs.

In **UK-DALE**, the fridge consumes significantly less power than FIRED's fridge, leading to a high volume of FNs for the kNN. As in our validation on FIRED, the kNN misclassified the kettle as an oven and the TV as a coffee grinder. These issues reflect in the average F1-score, which, when disregarding the unrecognized devices, would be comparable to the score achieved on the lab data.

The seq2point model achieved high scores for the fridge and TV, with lower results for the espresso machine and kettle (again due to many FPs), overall recognizing more devices than in the lab data.

Conclusions. Our studies suggest that domain bias does not constitute a significant hindrance to the deployment of NILM models in unfamiliar environments without retraining. We observed some difficulties with certain devices which could be attributed to the relatively sparse number of training samples. However, overall, the patterns of success and inter-device confusions appeared consistent across different domains. Importantly, the deterioration in results was not as marked as observed in prior studies [6].

7 Conclusion

In this paper, we explored the complexities of NILM transfer. We developed both event-based and eventless NILM models using public data and tested them in various new environments. Laboratory tests under simulated household conditions revealed that a key challenge in model transfer is the introduction of new signatures for the same device type. Testing with real household data across various domains led to consistent model behavior, showing that domain bias did not significantly hinder deployment further.

As an initial exploration of domain bias, our evaluation is not complete. Mitigating model performance degradation will require future work, involving an increase in training data for certain devices and the inclusion of more diverse datasets and features in training. However, the initial results are promising, suggesting that further research in this area will indeed be worthwhile.

References

[1] Awadelrahman MA Ahmed, Yan Zhang, and Frank Eliassen. 2020. Generative Adversarial Networks and Transfer Learning for Non-Intrusive Load Monitoring in Smart Grids. In *IEEE SmartGridComm.*

[2] 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).

[3] Michele D'Incecco, Stefano Squartini, and Mingjun Zhong. 2019. Transfer Learning for Non-Intrusive Load Monitoring. *IEEE Trans. Smart Grid* 11, 2 (2019).

[4] Xiao Chang et al. 2022. Transferable Tree-Based Ensemble Model for Non-Intrusive Load Monitoring. *IEEE Trans. Sust. Comp.* 7, 4 (2022).

[5] George William Hart. 1992. Nonintrusive Appliance Load Monitoring. *Proc. IEEE* 80, 12 (1992).

[6] Matthias Kahl, Thomas Kriechbaumer, Anwar Ul Haq, and Hans-Arno Jacobsen. 2017. Appliance Classification Across Multiple High Frequency Energy Datasets. In *IEEE SmartGridComm.*

[7] 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 *ACM e-Energy.*

[8] 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).

[9] Christoph Klemenjak, Anthony Faustine, Stephen Makonin, and Wilfried Elmenreich. 2019. On Metrics to Assess the Transferability of Machine Learning Models in Non-Intrusive Load Monitoring. *arXiv preprint arXiv:1912.06200* (2019).

[10] Yan Li, Yujiao Liu, Zhaoqing Zhang, Fang Shi, Guoliang Li, and Kun Wang. 2021. Non-intrusive Load Monitoring Method Based on Transfer Learning and Sequence-to-point Model. In *IEEE iSPEC.*

[11] Lucas Pereira, Donovan Costa, and Miguel Ribeiro. 2022. A Residential Labeled Dataset for Smart Meter Data Analytics. *Scientific Data* 9, 1 (2022).

[12] Benjamin Völker, Marc Pfeifer, P M Scholl, and B Becker. 2020. FIRED: A Fully-labeled hIgh-fRequency Electricity Disaggregation Dataset. In *ACM BuildSys.*

[13] Chaoyun Zhang, Mingjun Zhong, Zongzuo Wang, Nigel Goddard, and Charles Sutton. 2018. Sequence-to-Point Learning with Neural Networks for Non-Intrusive Load Monitoring. In *Proc. AAAI*, Vol. 32.