Fostering non-intrusive load monitoring for smart energy management in industrial applications: an active machine learning approach

Lukas Fabri, Daniel Leuthe, Lars-Manuel Schneider, Simon Wenninger

Angaben zur Veröffentlichung / Publication details:

Fabri, Lukas, Daniel Leuthe, Lars-Manuel Schneider, and Simon Wenninger. 2025. "Fostering non-intrusive load monitoring for smart energy management in industrial applications: an active machine learning approach." *Energy Informatics* 8 (1): 54. https://doi.org/10.1186/s42162-025-00517-5.

Energy Informatics

RESEARCH Open Access

Fostering non-intrusive load monitoring for smart energy management in industrial applications: an active machine learning approach

Lukas Fabri[1], Daniel Leuthe[1*], Lars-Manuel Schneider[2,3] and Simon Wenninger[1]

*Correspondence:
Daniel Leuthe
daniel.leuthe@fim-rc.de
[1]Research Center Finance & Information Management, Branch Business & Information Systems Engineering of the Fraunhofer FIT, Technical University of Applied Sciences Augsburg, Alter Postweg 101, 86159 Augsburg, Germany
[2]University of Augsburg, Universitätsstr. 2, 86159 Augsburg, Germany
[3]Reonic GmbH, Provinostr. 52, 86153 Augsburg, Germany

Abstract

Non-intrusive load monitoring (NILM) is a promising and cost-effective approach incorporating techniques that infer individual applications' energy consumption from aggregated consumption providing insights and transparency on energy consumption data. The largest potential of NILM lies in industrial applications facilitating key benefits like energy monitoring and anomaly detection without excessive submetering. However, besides the lack of feasible industrial time series data, the key challenge of NILM in industrial applications is the scarcity of labeled data, leading to costly and time-consuming workflows. To overcome this issue, we develop an active learning model using real-world data to intelligently select the most informative data for expert labeling. We compare three disaggregation algorithms with a benchmark model by efficiently selecting a subset of training data through three query strategies that identify the data requiring labeling. We show that the active learning model achieves satisfactory accuracy with minimal user input. Our results indicate that our model reduces the user input, i.e., the labeled data, by up to 99% while achieving between 62 and 80% of the prediction accuracy compared to the benchmark with 100% labeled training data. The active learning model is expected to serve as a foundation for expanding NILM adoption in industrial applications by addressing key market barriers, notably reducing implementation costs through minimized worker-intensive data labeling. In this vein, our work lays the foundation for further optimizations regarding the architecture of an active learning model or serves as the first benchmark for active learning in NILM for industrial applications.

Keywords Active learning, Energy efficiency, Machine learning, Non-intrusive load monitoring, Smart energy management

Introduction

The ongoing digitalization and the use of digital technologies such as machine learning has allowed smart service systems to become a core of technological developments [12, 13, 24]. Notably, the potential of smart energy management is increasingly recognized and finds its way into research and practice [1, 37, 38, 80]. This is due to the sustained increasing electrical energy demand in today's applications, ranging from personal devices to industrial applications, which is expected to increase by approximately 85% over the next two decades [1, 55]. Especially energy-intensive industrial nations such as Germany, where the industrial sector is responsible for around 45% of the electrical energy consumption [11], need to act to achieve efficient energy usage [64]. Consequently, companies introduce smart energy management to enforce energy reduction [1, 78]. Non-Intrusive Load Monitoring (NILM) is an emerging approach to enable more sophisticated smart energy management [4, 18, 48]. NILM estimates the power consumption of individual applications from aggregated power measurements [29, 31]. In contrast to intrusive load monitoring, NILM equalizes the implementation of expensive submetering while providing valuable information for energy-saving decision-making [18, 25]. Use cases of NILM include energy monitoring, demand side management, peak shaving, job scheduling and anomaly, and fault detection for saving energy [6, 29, 37]. For example, early detection of those faults or anomalies with the help of NILM avoids economic, energy, and environmental losses due to, e.g., maintenance cost reduction, machine fault reduction, increased spare part life, or the identification of inefficient applications [29].

Current NILM research focuses on technical and data science dimensions, including mathematical optimization and supervised and unsupervised machine learning algorithms [5, 18]. Unsupervised algorithms are suitable when labeled data is absent, but their accuracy remains typically lower than supervised algorithms as those are trained on labeled data [18, 29]. The primary application area of NILM is smart energy monitoring in the residential sector [25, 46, 67, 81]. However, although NILM has existed for decades, it has not been widely adopted in industrial applications [18, 29, 32]. The reasons are NILM's complexity and the lack of industrial data labels for promising supervised algorithms [21]. Nevertheless, labeling new data in an industrial environment requires domain-specific knowledge, is time-consuming and expensive to obtain—thus often unattractive in practice [21, 76]. Labeling in the context of NILM means determining the disaggregated applications from the total electrical energy consumption and to what extent they were used at the given timestamp.

Consequently, the objective of this work is to reduce labeling efforts. Therefore, research introduced the concept of active learning [30, 65], which is applied in various disciplines [4, 19, 33, 45, 62]. Active learning is a machine learning approach in which experts query the unlabeled data with the highest level of informativeness to subsequently provide targeted feedback to the learning algorithm to improve its performance [19, 51]. In the context of NILM, the labeled data is used for training the disaggregation algorithms that predict the energy consumption of the individual applications. In other domains besides NILM such as anomaly detection in industrial time series data or quality control for visual defect inspection, previous research has shown the potential of active learning to reduce the labeling effort [33, 63, 71, 79]. In this regards, previous work using active learning for NILM has mainly focused on residential loads and residential buildings, whereby active learning was able to achieve a reduction in data

labeling of up to 90% [30, 68, 70]. Nevertheless, research considering industrial applications has so far fallen short, as industrial applications present complex consumption patterns due to the varying number of heterogeneous loads and NILM algorithms necessitate extensive labeled datasets, which are scarce in industrial settings [21, 33]. Therefore, this work integrates an active learning model into NILM for industrial applications. Reducing the effort, especially in an industrial environment, increases the attractiveness of NILM. This enables smart energy management and therefore favors companies ecologically and economically. Hence, we pose the following research question (RQ):

> *How does an active learning model perform compared to established supervised learning models in NILM systems for industrial applications?*

To answer our RQ, we develop an active learning model using the HIPE data set based on a real-world production plant [14]. We implement our model to validate the underlying concept of reducing the labeling effort. Our methodology is based on the Cross-Industry Standard Process for Data Mining (CRISP-DM). The proposed model does not replace existing NILM techniques. Instead, it complements its techniques by significantly reducing data labeling and user effort while maintaining appropriate disaggregation accuracies. Hence, we reduce the concerns about the cost of NILM and make them more applicable for reducing energy consumption.

This work contributes to research and practice in three ways:

- First, we combine the two research streams on NILM for industrial applications [29, 38] and the research on active learning [19, 36].
- Second, we develop an active learning model, define the best-performing architecture depending on the disaggregation algorithms to separate individual applications' energy consumption, and compare three query strategies to select a subset of the training data efficiently.
- Third, the resulting active learning model fosters the introduction of a cost-effective NILM system that enables practitioners to save energy through smarter energy management and contribute to cleaner production.

The remainder is structured as follows: In Section "Theoretical background", we explain the foundations of active learning and examine previous research. In Section "Method development", we present our research approach and the design of the active learning model. In Section "Data and evaluation", we introduce the data set and the evaluation metrics. Afterward, in Section "Results", we show the results based on a comparison with a benchmark and a sensitivity analysis. We discuss the results obtained and point out further research in Section "Discussion" before concluding in Section "Conclusion".

Theoretical background

Active learning

Due to the limits of NILM, active learning is a promising approach to enhance NILM´s usage, especially in industrial applications with high electrical energy demand [1, 36]. Active learning is a subfield of machine learning that excels through interactively querying input of an expert system named as an oracle (often a human annotator), that provides the corresponding labels for model training [19]. A defined algorithm, called an active learner, queries the data based on an unlabeled data pool from which it expects better learning

success. An active learning model consists of two main components, as shown in Fig. 1: a query engine that selects the data instances from the unlabeled pool and the oracle that provides the labels [22, 47]. Hence, the idea is to select a minimum amount of unlabeled data instances to learn from that are annotated by a domain expert, maximizing the learning ability [65, 75]. In the context of active learning, the goal of the query is to select the data instances with the greatest information content so that the resources of the expert system (i.e., oracle) are used efficiently [33]. Goernitz et al. [27] identify the appropriate query strategy as the central challenge of active learning. In this way, the domain expert is not occupied with the distinction of data that contribute less to the success of the defined machine learning problem but can focus on data that significantly advance the training process [59]. Common use cases for active learning are speech recognition and medical classification. Thus, active learning is adopted in scenarios where data is abundant, but labels are complex, time-consuming, or expensive to obtain [65].

Literature provides many query strategies for unlabeled instance selection, divided into three streams [45]: Membership query synthesis, pool-based sampling, and stream-based sampling. Membership query synthesis is a strategy for generating queries based on some membership criteria for labeling. However, this approach is unsuitable for NILM due to the complexity of accurately simulating realistic appliance signatures [18, 36]. In pool-based sampling, queries are selectively drawn from an unlabeled data pool (Fig. 1). Based on the level of informativeness, all samples are ranked and then used for training. Stream-based sampling is similar to pool-based sampling, besides it queries the samples individually [65]. Overall, pool-based sampling, which involves assessing and ranking the informativeness of all samples within a substantial unlabeled dataset before selecting the most valuable ones for labeling, is best suited for NILM. It effectively manages extensive datasets, prioritizes the most informative samples for model training, and targets scenarios where many unlabeled samples are collected simultaneously [19]. Within the pool-based sampling, several specific query strategies have been identified as useful for NILM: naive query, extreme query, and cluster-based query strategy [19]. The naive query strategy involves randomly selecting samples from the unlabeled pool for labeling. While it does not actively prioritize informative instances, it serves as a baseline for evaluating the effectiveness of more sophisticated strategies. The extreme query strategy focuses on selecting samples with extreme or outlier feature values. In the context of NILM, these extreme instances often correspond to unique or rare appliance events and labeling them can significantly enhance the model's ability to recognize and disaggregate such events. Lastly, the cluster-based query strategy involves clustering the unlabeled data based on feature similarity and selecting representative samples

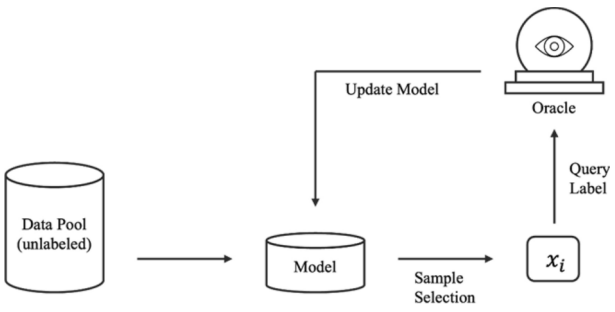

Fig. 1 Active learning—Pool-based sampling scenario

from each cluster for labeling. This ensures that the model learns from a diverse set of appliance signatures, improving its generalization across different appliance types [27, 33, 65].

Related work

Hitherto, research regarding NILM focuses on technical aspects (e.g., mathematical optimization, pattern recognition, and machine learning) to improve NILM applications [5, 37, 67]. An open-source NILM Toolkit was proposed to enable an empirical comparison across the disaggregation algorithms [7]. Providing a complete pipeline from data sets to performance metrics, the NILM Toolkit lowers the entry barrier for NILM research [e.g., 32, 38, 75]. Nevertheless, the application domains and the temporal resolution of the available data show only slight variations over the data sets. Table 1 shows that research and data on NILM focus on household settings and mainly consider residential buildings [e.g., 3, 40, 43]. The reasons for the lack of industrial data are three-fold: First, most companies do not collect granular electricity data on an application level. Second, even if they do so, they rarely publish their data [38]. Third, household data is more accessible to aggregate due to less complex applications and, therefore, more straightforward to transfer between households. While researchers acknowledge NILM´s importance in industry, data collected from industrial applications are rare and mainly comprise data with a low-frequency sample rate [50, 83]. Only a few articles provide an industrial perspective [e.g., 9, 32, 38, 50]. In the industrial sector, the assumptions made for NILM differ from residential houses. In commercial buildings, the base load of applications tends to be higher than in households [9]. Hence, this implies that most low-frequency methods would fail to disaggregate the respective portion of electrical energy.

Nevertheless, research shows that NILM algorithms can replace resource-consuming submeter hardware in the industrial context [32]. However, this does not tackle the issue of having too much data lacking class labels in the industrial sector. Against this backdrop, and often having limited resources for manual labeling, a combination of active learning with NILM systems is a promising approach [30, 36]. This is mainly due to the

Table 1 Focus of NILM's application and the combination of active learning in research

References	Residential focus	Industrial focus	Active learning	Public data set
Kolter and Johnson [43]	✓	–	–	✓
Anderson et al. [3]	✓	–	–	✓
Kelly and Knottenbelt [40]	✓	–	–	✓
Holweger et al. [34]	✓	–	–	✓
Wu et al. [75]	✓	–	–	✓
Langevin et al. [46]	✓	–	–	✓
Werthen-Brabants et al. [73]	✓	–	–	✓
Timplalexis et al. [69]	✓	–	–	✓
Jin [36]	✓	–	✓	✓
Guo et al. [30]	✓	–	✓	✓
Todic et al. [70]	✓	–	✓	✓
Tanoni et al. [68]	✓	–	✓	✓
Batra et al. [9]	✓	✓	–	✓
Holmegaard and Baun Kjaergaard [32]	–	✓	–	–
Martins et al. [50]	–	✓	–	✓
Kalinke et al. [38]	–	✓	–	✓

– Not fulfilled; ✓ fulfilled

fact that NILM data in industrial applications refers to easy-to-obtain energy data (e.g., using installed sub-meters or in-system meters used for energy monitoring) as opposed to common hardly accessible mechanical data or process data [33]. Hence, NILM combined with active learning is gaining attention for industrial energy management [67]. In this vein, Guo et al. [30] present an active deep learning NILM method using discrete wavelet transform features. It targets large-scale datasets by actively selecting the most valuable power signal samples to label, either in a pool-based or stream-based query strategy. The authors form a mixed dataset from three public NILM datasets (covering residential loads) to evaluate the active learning query strategies. The approach reduces the labeling costs, i.e., 33% fewer labeled samples are needed than the supervised baseline model. Similarly, Jin [36] proposes an active learning framework to enhance NILM by interactively querying users for minimal class label information based on a k-Nearest Neighbors classifier. The framework is validated using the BLUED dataset covering residential loads. This approach reduces the manual labeling burden, achieving similar disaggregation performance while requiring up to 90% fewer user inputs. Compared to Guo et al. [30], focusing on feature extraction and sample selection strategies, Jin [36] emphasizes minimal user interaction to decrease labeling costs. On step further, Todic et al. [70] adapt a deep learning disaggregation model for NILM using active learning and compare different query strategies. The authors use the public available REFIT dataset which consists of four residential appliances. The results show that with active learning, the model achieves comparable accuracy using only 5–15% of the training data, greatly lowering labeling costs. Similar results are achieved by Tanoni et al. (2024b) as they propose a combined weakly supervised and active learning approach. They validate their architecture using two public datasets, each covering residential loads from different households. The combined approach reduced the amount of labeling required by almost 90%. All in all, the combination of NILM and active learning could reduce the labeling effort by up to 90% while delivering similar disaggregation results [30, 36, 68].

In sum, research shows the potential of using active learning to reduce labeling costs. While existing research on NILM and active learning focuses on residential buildings, only a few works consider an industrial perspective as described in the previous paragraph [67]. The reasons are that, first, industrial environments present intricate aggregate consumption patterns due to the vast number of heterogeneous electrical loads, complicating the accurate disaggregation of individual load signatures. Second, supervised NILM algorithms necessitate extensive labeled datasets, which are scarce in industrial settings as they require domain-specific expertise, making the process both time-consuming and costly, thereby deterring practical implementation [21, 33]. However, given the promising approaches, combined methods could foster NILM's dissemination within smart energy management in the industrial sector to enforce energy reduction. Consequently, this work investigates the performance of active learning in combination with NILM for industrial applications.

Method development

Research approach

A suitable methodology is necessary to address the RQ and benchmark our model against established NILM models for industrial applications. To meaningfully compare different models, we derived a five-step process from the CRISP-DM and the guidelines

by Müller et al. [53] for extensive data analysis. Generally, the CRISP-DM provides a standardized process in six steps: "Business Understanding", "Data Understanding", "Data Preparation", "Modeling", "Evaluation", and "Deployment" [74]. We explain our derived process steps in the following:

Business understanding We adapt the initial first step of the CRISP-DM to compare different models and, hence, infer a benchmark problem in terms of the underlying business understanding. This is in line with the RQ of how an active learning model performs compared to established supervised learning models in NILM systems for industrial applications. Relevant to our benchmark problem, we collect domain-specific knowledge about model performance in NILM systems for industrial applications. We introduce domain-specific knowledge in Section "Theoretical Background", providing the theoretical background for NILM and active learning and examining previous research approaches. For the benchmark, we use the NILM Toolkit API as described in Sect. 2.3, a standard for reproducible comparisons of energy disaggregation algorithms [7, 61]. We further compare it with three different applied algorithms (i.e., Mean algorithm, Combinatorial Optimization algorithm, Sequence to Sequence algorithm) for the load disaggregation. Those three algorithms and their selection justification are further explained in Section "Active learning architecture". After selecting the benchmark model for comparison, we modify the CRISP-DM again by introducing our performance evaluation measures, as explained in Section "Data description and preparation", before proceeding with data understanding.[1]

Data understanding This step has not been modified. The underlying dataset consists of smart meter readings from ten industrial applications and the readings of the main terminal of a power electronics plant. Data is available for three months with a 5 s resolution. Our study includes exploratory data analysis, explained in Section "Data and Evaluation".

Data preparation: This step was not modified either. We prepare the data by converting it to HDF format and slicing the time period considered. Afterward, we interpolate the missing data and reduce noise in the main terminal. Last, we shrink the data set (Section "Data description and preparation"). This procedure ensures high data quality for reliable results.

Modeling, comparison & evaluation In these steps, we implement and train our active learning model (see Section "Results"). Within this phase, we define the active learning architecture with its components, as illustrated in Section "Active learning architecture". It includes query strategies and algorithm training. The modeling step is performed iteratively and completed when all models'parameters have been optimized. To increase reproducibility, the resulting parameters of the models are in the respective Table 5 (i.e., machine learning report card according to Kühl et al. [44]) in Appendix A. Subsequently, we validate the designed active learning model. Afterward, we evaluate the results against a benchmark using four established performance evaluation measures (Sect. 4.3). The benchmark is based on all available training data with the same sampling rate as the active

[1] Note, the dataset was already at our disposal when we started with our study. Hence, we did not include "Data Collection" to our derived process. Nevertheless, this step could be set in parallel with the definition of the performance measure to allow a general process application and to enable further studies starting without available datasets.

learning model. Hence, within the benchmark, we assume that 100% of the initially unlabeled training data is labeled and then used for the subsequent disaggregation predictions. We use the same three disaggregation algorithms as in the active learning model for these predictions.

Deployment: This step largely coincides with the deployment step in the original CRISP-DM, which involves the framework of the productive operation. Besides the discussion chapter, we disregard the last step for most of this work because it is out of scope. We discuss our results and present derived implications for policy, research, and commercial application in Section "Discussion".
In this vein, our results solve the defined problem with our RQ and close the CRISP-DM process cycle.

Active learning architecture

To realize the active learning architecture within the research step "Modeling, Comparison & Evaluation", we combine an unsupervised model with a supervised model involving a feedback module by building on Das et al. [19] and Pimentel et al. [59]. We extract further information from the data with the unsupervised model's help, increasing the data's efficient usage. Figure 2 illustrates the architecture's components.

Unsupervised model

We base the selection of the unsupervised model on the taxonomy of deep learning approaches for anomaly detection [57]. Thus, we utilize an Autoencoder (AE) which outperforms the compared methods [57]. AEs are trained using unsupervised learning on unlabeled training data. The AE consists of two different neural network types, an encoder and a decoder, with a so-called bottleneck in between. The encoder takes the unlabeled data $D = \{x_1, ..., x_n\}$, where D is a data set with n observations and $x_i \in \mathbb{R}^d$, as an input vector to compress it from a high-dimensional input space to a lower dimension, called latent space. The decoder reconstructs the input vector from the low-dimensional latent representation [17]. They learn the reconstructions close to their original input, thus ignoring the noise in the data [28]. The difference between the original input

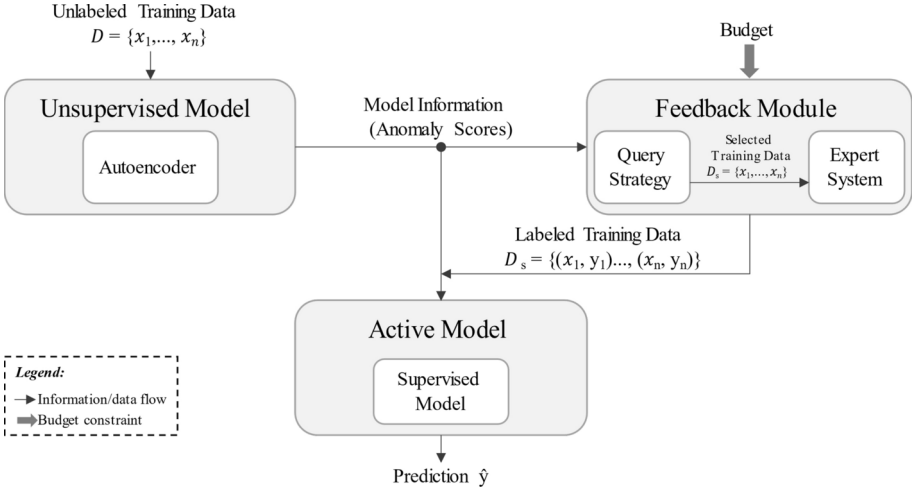

Fig. 2 Components of the active learning architecture depicting an unsupervised model, a budget constraint feedback module, and a supervised model

vector and the reconstructed vector is called reconstruction error. For AE-based anomaly detection, we use the reconstruction error as the anomaly score [2, 46]. Data with high reconstruction errors are considered anomalies. The purpose of the unsupervised model is to provide model information that rates each data point's anomaly score [4].

Feedback module

The feedback module serves as a labeling entity for the training data. The module's output represents a set of labeled data (x, y) as shown in Fig. 2. The unlabeled training data and the described unsupervised model form the input. Hence, the feedback module consequently has access to the anomaly scores. The feedback module consists of the query strategy and an expert system, i.e., oracle (Fig. 2). The budget represents the capacity constraint that determines the output of labeled data and hence the capacity of the expert system [19]. First, the query strategy selects a subset $D_s = \{x_1, ..., x_n\}$ of the input training data based on the information provided by the unsupervised model. Subsequently, the data are labeled by the expert system. To achieve the highest impact of the limited labeling capacity, active learning aims to maximize the informativeness of the labeled data for the following supervised model [19, 36]. Again, labeling means determining the connected applications and to what extent they were used at the given timestamp. Within this work, the expert system is simulated. We assume the expert system is flawless, always assigning the correct data labels [82].

We compare naive, extreme, and cluster queries introduced by [27], as illustrated in Fig. 3 and previously described in Sect. 2.1. Naive queries describe the random selection of data instances as part of the unlabeled training data. Additional information like anomaly scores is not considered. Consequently, only a few anomalies are queried, proportional to the budget. Extreme queries select the data most likely to be anomalies based on their anomaly score. Hence, this query strategy uses the information provided by the unsupervised model and queries them in descending order until the budget is reached. This strategy is advantageous if the proportion of anomalies in the training data set is small since this allows for finding as many supposedly informative anomalies

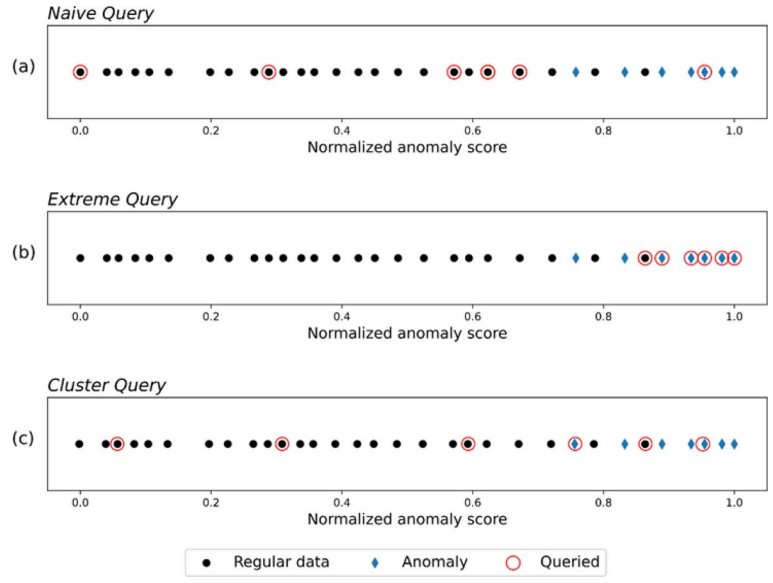

Fig. 3 Graphical comparison of the naive, extreme, and cluster query strategies

as possible [19, 59]. Cluster queries focus on selecting data equally from different clusters of the provided data set regardless of the density of each cluster. The latent space is divided into clusters using, e.g., a k-means clustering algorithm [19]. Thus, this query prevents a disproportionate sample selection [27].

Supervised model

The supervised model uses the labeled data for training. We use three different algorithms (i.e., Mean algorithm, Combinatorial Optimization algorithm, Sequence to Sequence algorithm) for the load disaggregation. First, the Mean algorithm estimates the output usage of an application to be the mean computed over the training data. As stated by previous work, the Mean algorithm is a strong benchmark for evaluating other disaggregation algorithms [20, 32]. Second, we use the Combinatorial Optimization (CO) algorithm, which also serves as a baseline algorithm in existing NILM literature [10, 31, 60]. The assumption is that each application can be in a given state (1 of K where K is a small number), where each state has an associated power consumption [8]. The objective of the CO algorithm is to assign states to applications so that the difference between the aggregate power reading and the sum of power usage distributed over the different applications is minimized. Third, we consider Sequence to Sequence (Seq2Seq) learning. Seq2Seq learning is a neural network that uses 1-dimensional convolutions to map an input sequence to a variable length output sequence and delivers promising results in research [26, 77]. This approach trains one deep neural network per application and learns its behavior based on input data sequences. Seq2Seq is suitable for load disaggregation in NILM and outperforms classical algorithms [16, 77].

Data and evaluation

Public datasets

Publicly available energy-related data sets primarily consist of energy data of residential or office buildings, with a granularity of ten or more minutes [14, 58] such as the UK-DALE [40], the BLUED [3], or the REDD-dataset [43].

Commercial data sets are far less publicly available [14]. To the best of our knowledge, only four have been released to date—two for commercial buildings and two for industrial applications. The EnerNOC data set collected data across different commercial buildings with a 5-min resolution [52]. The BERDS data set provides the energy consumption data of applications at a 15-min resolution in a commercial building [49]. Neither data set contains high-frequency sampled data. Hence, both have limited applicability for NILM. For industrial applications, only two publicly accessible data sets suitable for NILM research exist. First, the IMBELD data set contains six industrial applications [50]. The monitored machines are solely applications in terms of energy consumption with two possible states, "on" and "off". Second, the HIPE data set is an industrial energy data set and the most comprehensive one released so far [14]. Since this work focuses on removing barriers regarding NILM's adoption in industrial applications, we use the HIPE data set due to its high data quality and realistic applicability.

Data description and preparation

The HIPE data set contains smart meter readings of ten industrial applications and the main terminal's readings over three months at a 5-s resolution from a power-electronics

plant. The plant produces electronic systems for battery systems in small batches, i.e., less than 1000 pieces. The data is collected via high-resolution smart meters and sampled with a frequency of 0.2 Hz containing various electrical quantities like active, reactive, and apparent power. The load signature of the monitored applications ranges from simple, with two possible states, to complex, with continuously variable states. Table 2 gives an overview of the applications considered in the HIPE dataset and their respective energy consumption. The different applications' usage and energy consumption vary strongly, and their usage depends on each other (Appendix B). Such dependencies are unusual for a household setting and are characteristic of industry data [38, 43].

The raw data set has been pre-processed to remove measurement errors. However, the measurements are not equidistant, i.e., the time between subsequent measurements varies. Usually, the temporal resolution of the data points is about 5 s. Nevertheless, there are cases where the time between successive data points is several minutes due to maintenance. To tackle that issue, the data points between these outliers have been interpolated to achieve equal distances between the measurements. This ensures uniformity in the dataset, facilitating more accurate analyses. To reduce existing noise included in the load profile, the main terminal and the submetered applications are leveled by subtracting the baseload from the main terminal. This enhances the clarity of the data by reducing fluctuations unrelated to the primary signals of interest. The data was collected from 2017–10-01 to 2018- 01–01. The machines did not run at the end of December, and the 3rd of October is a German holiday. In turn, we limit the period from 2017–10-04 to 2017–12–21, matching the used time frame in previous research [14].

Performance evaluation measures

Evaluating the disaggregation performance for NILM requires performance evaluation measures–PEMs [4, 41, 66]. Nalmpantis and Vrakas [54] and Klemenjak et al. [42] provide an overview of the most commonly-used PEMs split into two categories. The first category is based on the comparison between the observed aggregate power signal and the reconstructed signal after disaggregation. Within this category, they mention the

Table 2 Overview of the applications, their description and their respective energy consumption within the HIPE dataset [14]

Application	Description	Total energy consumption [kWh]
Chip press	Heat treatment of surfaces under high pressure, e.g., for multi-layered printed circuit board	246.26
Chip saw	Separation of chips of a silicon wafer	51.40
High-temperature oven	Fixing layers for thick-film technology (heats up to 1200 °C)	250.14
Pick and place unit	Placement of electronic components such as resistors and microcontrollers on a printed circuit board	60.64
Screen printer	Printing of material layers to interconnect electronic components	63.46
Soldering oven	Components soldering to the printed circuit board	186.95
Vacuum oven	Oven with a vacuum chamber	10.83
Vacuum pump 1	Auxiliary machine to generate vacuum for other machines such as the Pick and Place Unit	271.41
Vacuum pump 2	Auxiliary machine to generate vacuum for other machines such as the Pick and Place Unit	65.71
Washing machine	Cleaning of the printed circuit board at the end of the production line	81.74

Root-Mean-Square Error (RMSE), the Mean Absolute Error (MAE), and the Normalized Disaggregation Error (NDE). The second category describes how effectively the disaggregated signal signatures are assigned to the application signatures. The main PEMs used are Precision, Recall, and the resulting F-Score. The F-Score provides a balance between precision, i.e., the ratio of correctly assigned signatures to the total number of assigned signatures, and recall, i.e., the ratio of correctly assigned signatures to the total number of actual signatures [33]. Table 3 shows the PEMs, including their formal definitions, units, value ranges, and optima.

F_i and A_i are the predicted and actual values for a data instance i, N is the sample size, and $\overline{A}$ is the mean of all actual values A_i. The value range represents a right-hand infinite closed interval including the value "0" for each metric. Each PEM exhibits different characteristics, leading to different outcomes of prediction accuracy. With the RMSE, comparing applications with high differences in power consumption (i.e., a high-temperature oven and a LED light bulb) is challenging [34, 54]. Hence, we also consider the NDE as it normalizes the squared error of a single application by the total energy of the signal (Klemenjak et al. [42]). The MAE is well-known in signal processing and is similar to the RMSE [41, 42]. However, unlike the RMSE, the MAE does not penalize outliers with quadratic weight. Outlier sensitivity is essential, as high deviations between predicted and actual values are not beneficial for NILM. To further evaluate the performance of the proposed active learning algorithm, it is necessary to detect whether an application is on or off. Hence, as a fourth PEM, we apply the F-Score [8]. Furthermore, as the F-Sore is unitless, it provides an intuitive understanding of the PEMs for readers unfamiliar with this subject. As selecting the best-suited PEM is not trivial, comparing several PEMs is preferable [66].

Results

Comparison to benchmark

First, we show the active learning model's results before introducing the benchmark model's prediction results. Following, we contrast the scores of the three tested disaggregation algorithms of the active learning model with the results of the benchmark model. For this comparison we only consider the values of the best-performing query strategy.

In general, the cluster query strategy outperforms both the naive and extreme query strategies over all tested permutations and across all considered PEMs, as shown in Appendix E. Looking at the individual mean values across all ten appliances, we can surmise significant differences between them. A one-way ANOVA test, which tests the means of several groups for equality [23], shows at a significance level of 0.1 that at least one mean

Table 3 Overview of the performance evaluation measures used for the disaggregation performance

Measures	Abbreviation	Equation	Unit, value range	Optimal value
Root-mean-square error	RMSE	$\sqrt{\frac{\sum_{i=1}^{N}(F_i-A_i)^2}{N}}$	$kWh, [0, \infty)$	0
Mean absolute error	MAE	$\frac{1}{N}\sum_{i=1}^{N}\lvert F_i - A_i\rvert$	$kWh, [0, \infty)$	0
Normalized disaggregation error	NDE	$\frac{\sum_{i=1}^{N}(F_i-A_i)^2}{\sum_{i=1}^{N}{A_i}^2}$	$kWh, [0, \infty)$	0
F-score	F-score	$\frac{2\times Precicion\times Recall}{Precicison\times Recall}$	[0, 1]	1

is statistically significantly different from the others (see Appendix E for details). Limiting the number of queried samples per cluster prevents the feedback module from picking too much data from one cluster while still considering outliers with high anomaly scores and thus using samples from the entire latent space. In contrast, the naive query selects a majority of the samples from high-density clusters with low anomaly scores and little benefit to the active learning model, the extreme query selects isolated outliers which are less common and lead to poor reconstruction. Hence, for the comparison to the benchmark, we solely consider the results of the cluster query strategy.

Figure 4 displays the disaggregation performance of the active learning model using a budget size of 1% of the data for training the disaggregation algorithm. The Seq2Seq algorithm achieves the best average prediction performance across all considered PEMs (RMSE = 335.88, F-Score = 0.33, MAE = 99.75, NDE = 0.83) compared to the CO (RMSE = 530.95, F-Score = 0.32, MAE = 129.74, NDE = 1.62) and the Mean (RMSE = 554.09, F-Score = 0.25, MAE = 343.76, NDE = 1.09). Furthermore, conducting another one-way ANOVA test for all three algorithms, a 1% budget, the cluster query strategy and MAE as a PEM for each application, resulted in a p-value of 0.03 which shows strong statistical significance of the algorithm selection assuming again a significance level of 0.1. We note that the standard deviation of the algorithms varies considerably. Across all PEMs, the Seq2Seq algorithm shows the widest spread of possible results. In contrast, the Mean algorithm barely shows any volatility. By design, the Seq2Seq algorithm has more flexibility than the optimization problem of CO and the Mean disaggregation. The boxplots visualize this behavior of the algorithms.

Table 6 in Appendix C contains the average benchmark values displaying the best in bold print. This benchmark only consists of the results of the Seq2Seq algorithm since it received the best disaggregation performance across all considered PEMs. The disaggregation performance of the benchmark model shows that even with the best possible preconditions, i.e., the fully labeled training data and the most suitable algorithm, the data set is difficult to disaggregate. These observations support previous research findings from other studies [38]. The scores for the respective PEMs differ among the individual applications mainly depending on the pattern of their load profile. As visualized in Appendix B, certain applications stand out because of their especially good or poor disaggregation performance. As with the benchmark model, the active learning model

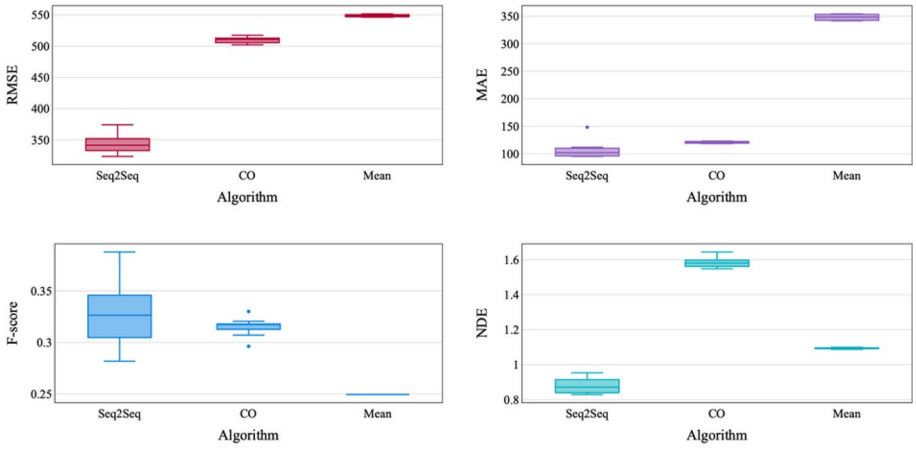

Fig. 4 Active learning prediction results using 1% budget size and ten iterations presented as RMSE, MAE, F-score, and NDE performance evaluation metrics

Table 4 Mean prediction results for 1% budget size comparing the benchmark model with the active learning model and considered algorithms

	Algorithm	RMSE	F-score	MAE	NDE
Benchmark model	Seq2Seq	279.53	0.47	72.41	0.65
Active learning model	Seq2Seq	335.88	0.34	99.75	0.83
Active learning model	CO	530.95	0.32	129.74	1.62
Active learning model	Mean	554.01	0.25	343.76	1.09

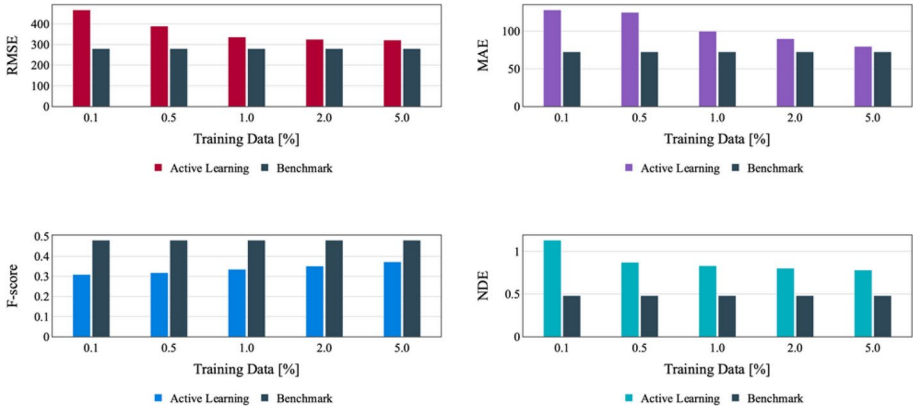

Fig. 5 Comparison of the active learning model with a varying budget size (0.1, 0.5, 1, 2 and 5%) and benchmark prediction results on the whole data set

benefits some applications more than others. Appendix D shows the ratio when comparing the results of the active learning model against the benchmark model. In most cases, the scores of the active learning model are lower than the benchmark model. However, in the case of the high-temperature oven, scores are similar with a ratio above 0.9 or even equal in the case of the Pick and Place Unit.

We focus on comparing the average disaggregation results across all measured applications, thus the arithmetic mean. Table 4 shows the scores of both the benchmark and active learning models. The benchmark model achieves better scores on all PEMs, with the active learning model using the best algorithm (Seq2Seq) reaching around 70% of its accuracy. The CO and Mean algorithm scores display the difficulty of disaggregating the load profile of the considered data set.

Sensitivity analysis

Figure 5 compares the disaggregation performance of the proposed active learning model against the benchmark. Analogous to the benchmark, we solely pick the best-performing algorithm of the active learning model for comparison against the benchmark. We use a sensitivity analysis varying the budget size between 0.1% and 5% of the total training data to investigate to what extent a variation in the budget size influences the performance of the active learning model. Total data points correspond from 1,054 to 52,705 queries of unlabeled data that are then labeled through the expert system. Essentially, the budget can be considered a special hyperparameter of the active learning model that needs to be optimized. However, the goal of this work is not to minimize the number of queried samples or to achieve the best possible prediction performance but to reduce the use of the expert system to a feasible level while still achieving satisfactory disaggregation results. For most PEMs, the first three budget sizes significantly impact

the disaggregation results, then slowing down towards the benchmark. As anticipated, the disaggregation results of the active learning model get closer to the benchmark as the budget size increases without fully reaching it.

Depending on the budget, disaggregation accuracies span from 23.1% to 90.1% for all considered PEMs. The initial budget of 0.1% receives poor scores showing the difficulty of disaggregating when only training on a very small sample of labeled data. We obtain a strict monotonically increasing shape of the accuracy with an increased budget for all PEMs. With RMSE and NDE, the disaggregation accuracy increases at the beginning but then reaches a plateau a little below 80% and a little above 80%, respectively. From there on, both scores increase only slowly. In contrast, the MAE indicates a different behavior with a slow increase initially but enhancing between 0.5% and 5%. The growth rate slows down marginally afterward with a continuously increasing budget size. For the F-score, it appears that the increase of the score is somewhat slow and steady compared to the other PEMs. Yet, as with all other PEMs, a positive trend can be recognized as the budget increases.

Discussion

Results interpretation

Researchers are aware of the importance of NILM in industrial applications [18, 50, 61]. Although the potential benefits of applying this technology to industrial devices have been recognized since the field's inception, most NILM studies focus on the residential sector [42, 56]. This focus is mainly due to the complexity of industrial applications and the lack of available labeled energy data.

Based on the stated RQ on how an active learning model performs compared to established supervised learning models in NILM systems for industrial applications, our results show that the present active learning model can effectively reduce the labeling effort for NILM by up to 99% while still achieving between 62 and 80% of the prediction performance compared to a benchmark with 100% labeled training data. This trade-off between prediction performance and labeling effort is remarkable, as in industrial settings, these labels typically require domain-specific knowledge, are time-consuming, and are expensive to obtain; thus, they are often unattractive in practice [33]. The design of the active learning model that allows this statement consists of an unsupervised model, a feedback module, and a supervised model. We show that the cluster query strategy is superior, achieving up to 50% better results than the other two strategies (i.e., naive and extreme queries). We found the Seq2Seq algorithm significantly better for load disaggregation than the CO and Mean algorithms. To examine the shape of the disaggregation results over different budget constraints for the feedback module, we use sensitivity analysis and vary the budget size in discrete steps between 0.1% and 5.0%. The sensitivity analysis shows no obvious answer as to which tested budget size is the most feasible for the given use case. Hence, the larger the budget size, the higher the cost and the better the disaggregation results. Furthermore, if the goal is to minimize the budget by 0.1%, corresponding to 1,054 queried data points in our case, the disaggregation results are comparatively poor. In addition, the maximum slope of the ratio of the active learning model to the benchmark lies at different budget sizes across the PEMs used. However, to indicate which budget size is appropriate in this context, we define a budget size of 1%, corresponding to 10,541 data points, since at this level, a plateau is reached for both the RMSE and the NDE. Using this budget of only 1% of the training data for this active

learning model, we achieve over 70% disaggregation accuracies compared to the benchmark using 100% labeled training data in 3 out of 4 PEMs.

Theoretical implications

First, this work contributes to the existing literature by consolidating two research streams: The research on NILM for industrial applications [29, 38] and the research on active learning as a novel machine learning method [19, 36, 54]. Yet, no research has been conducted on active learning for NILM using industrial applications. Therefore, we attempt to close this gap by combining these two technologies. Second, we develop an active learning model and establish the best-suited disaggregation algorithms. Based on previous research on anomaly detection for industrial applications using active learning, we adjust the existing model for NILM and design the respective architecture [22]. For this particular use case, the cluster query strategy is superior to naive and extreme queries and works best for training the disaggregation algorithm. The best-performing algorithm is the Seq2Seq disaggregation, beating both CO and Mean. Both decisions align with previous findings from similar research [38]. Third, we contribute to an energy-efficient industrial production by overcoming the barrier of the lack of industrial data labels for promising supervised algorithms. In this vein, we enable broader use of NILM and, consequently, energy conservation by tracking energy consumption patterns, identifying inefficient applications, and enabling smarter energy management [1]. Fostering the usage of NILM in industrial applications equalizes the implementation of expensive submetering, which is more economical while providing valuable information for energy-saving decision-making [18]. Fourth, our work can be discussed from a higher and more abstract level in the context of Watson et al. [72]'s energy informatics framework. Our active learning model represents an information system itself positioned between the framework's supply and demand side. However, one could argue that our active learning model does not necessarily manage the balancing of supply and demand side on its own, but it can be effectively applied on either the demand or supply side. The derived insights into demand or supply side characteristics, i.e., submeter consumption, subsequently enable other information systems to manage supply and demand with demand response measures for example. Doing that, our model helps digitizing the framework's 'sensor networks' and 'sensitized objects' by providing insights that formerly where limited to installing additional hardware. Further, our model incorporates the framework's 'stakeholders' not only as consumers, suppliers, or the government but as actual input providers through labeling tasks that effectively influence the information systems performance and behavior. In sum, our work contributes to closing the gap of hitherto scarcely researched and applied NILM in industrial applications. We provide valuable information on how to select the disaggregation algorithm in this case and the reasons for the usually more difficult disaggregation. Yet, the most significant contribution is the active learning model's architecture which contributes to implementing NILM efficiently. This developed model can be the foundation for further optimizations regarding the architecture of an active learning model or serve as a benchmark.

Practical implications

Although NILM is recognized in the industrial sector, no NILM systems are implemented for this use case [38]. Therefore, our active learning approach provides practical guidance

for its efficient implementation. Our research is conducted under near-real-world conditions. We use the HIPE dataset, which consists of real production plant measurements, not laboratory conditions. The dataset includes the main terminal and does not rely on virtual main meters. However, our research setting implies that we have to deal with a lot of noise in the data, which makes accurate load disaggregation difficult. To clarify, the disaggregation of industrial loads with only one metering point is unrealistic [32]. Therefore, sub-metering is necessary according to the size of the industrial plant.

First, the proposed active learning model can provide intelligent advice on which devices should be submetered. The active learning model selects the most informative data points based on an unlabeled data set. During the initial queried data points, the active applications indicate a high value to the disaggregation results, providing a solid basis for additional submetering. Making smart decisions about which devices to submeter directly impacts the effort required to implement NILMs. For example, it reduces the equipment cost and the time required for the complex installation of submeters in an industrial setting [32]. Second, our analysis of the HIPE dataset shows that the equipment in the production plant is only in use during the working week. The weekend energy demand is only the base load. Therefore, there is a significant difference between the energy consumed on weekends and weekdays. Thus, the day of the week is an important feature to consider when performing load disaggregation. This explains why the disaggregation results are worse when considering the whole data set including several more holidays and weekends, as previous results with different data splits show. Industrial sites behave similarly across industries, with only about 10% running on weekends or holidays. Hence, these usage patterns can be exploited by considering separate models for weekdays and weekends, thus improving the overall disaggregation performance of active learning models. Third, the designed active learning model leads to a cost-effective NILM system in practice. In addition to the possibility of using easily collected energy consumption data (e.g., with installed sub-meters or in-system meters used for energy monitoring), the expert system combined with human annotators can be efficiently used to optimize both the prediction performance and the budget for the feedback module. When implementing an active learning system, it can consequently be deduced that, in addition to the technical aspects, the involvement of employees and their expertise in the daily workflow should also be considered. Hence, we expect these results to increase the interest and implementation of such systems as pilot projects. Industrial sites often consume several orders of magnitude more energy than residential buildings, and thus the expected return on investment for NILM is greater. The presented active learning architecture reduces the labeling effort, removing one of the key barriers to adopting NILMs in industrial applications. For example, practitioners could use NILM for failure prediction, leading to cost savings and increased equipment efficiency, or for emission reduction and energy savings through smarter energy management [13]. Consequently, a wider use of NILM could help practitioners to save energy and costs and contribute to a cleaner and more sustainable production. Fourth, our work fits well in the broader context of energy management systems and digitals being used for industrial (energy) data monitoring. Energy management systems typically support a functionality known as'virtual meters'or'virtual datapoints', where non-existent or unconnected meters are extrapolated through basic mathematical calculations (e.g., calculating the energy consumption of a particular floor based on the overall building

consumption minus the consumption of other floors). Our proposed approach represents a natural extension of the virtual meters concept, offering over time varying, thus more accurate, insights obtained through NILM techniques. This extension can serve as a facilitator for various monitoring applications, including submetering of equipment and the derivation of consumption, costs, and emissions associated with products in a manufacturing setting. Such information is essential for reporting practices as required for sustainability reporting and compliance with ISO 50001 energy management systems standards [35]. Moreover, by incorporating our identified consumption data into digital twins, our approach has the potential to enhance production planning, asset management, and other related processes. We conclude that the proposed active learning is suitable to replace traditional submetering, although it has deficits in disaggregation compared to established supervised learning models. The significant time and resource savings compensate for this shortcoming.

Limitations and further research

Naturally, our work is subject to limitations but offers prospects for further research.

- First, although the proposed active learning model has shown its potential to reduce labeling efforts significantly, the prediction is still inferior to established supervised learning models in NILM systems. Providing accurate predictions to support decision-making processes is essential, potentially leading to energy savings. Yet, even though we have created a solid basis for algorithm comparison, further research can apply different (hyperparameter) tuning techniques, such as halving grid search to improve the current prediction performance. For instance, we focus on three query strategies and three disaggregation algorithms. Also, we split the data set into train and test data to ensure comparability with previous studies on the data, allowing a high bias and variance since we are not using cross-validation [38]. Testing additional algorithms or using cross-validation may increase the prediction performance and, thus, impact the current design of the active learning model.
- Second, by design, active learning models assume that the expert system is infallible and indefatigable. This assumption may be unrealistic in many real-world settings, especially when utilizing a human annotator as an oracle. Human experts working under quality and time pressures in a heavily efficiency-driven setting may make human errors, distorting the labels used to train the supervised model to correctly learn and disaggregate the loads. In many real-world applications, multiple imperfect predictors may have differing qualities. Hence, to increase the feasibility of the active learning model, future research could sprinkle random misjudgments. Given this, future research could build on Zhu and Yang [82], who developed a concept that distinguishes between human expert systems of different levels of reliability. While such approaches can represent real-world circumstances in a more detailed manner, they also increase the complexity of designing an optimal query strategy. Hence, in addition to the selection of the examples to be queried, it is necessary to assign them to the respective expert systems [33]. Further, instead of adding randomness, a proactive learning model could be proposed, bridging the gap between traditional active learning and more practical real-world scenarios [36]. Extending active learning to proactive learning aims to predict true labels given the risk estimates and the noisy output of predictors.

- Third, we could not achieve maximally precise results due to hardware limitations. The computation power allowed only ten iterations, four PEMs, and smaller budget sizes. Further research might address these issues by replicating our study on high-performance hardware.
- Fourth, as with any data-driven endeavor, the availability and use of data are a limitation of our study. Thus, we evaluated and tested our active learning approach only on a single data set of ten industrial applications, limiting the transferability and generalization of our results. The same as for the data set holds true for various other industrial applications in which the explanatory power of energy data for disaggregation might differ. Although Chen et al. [15] state that "data on energy consumption of manufacturing machines also contains the information on the conditions of [manufacturing] machines" and Kaymakci et al. [39] identify particular useful results for energy data of a laser punching machine, we call for extending our research to other industrial applications.
- Fifth, our work compared a subset of available NILM approaches and neglected others like unsupervised clustering or self-supervised learning as well as the integration into overarching approaches such as federated learning. We recommend for future research to expand the scope of considered NILM approaches to receive a more holistic picture about performance improvements.

Conclusion

The industrial sector is a major consumer in terms of electrical energy consumption. Hence, there is significant potential for realizing cost and energy savings through NILM. However, the implementation of NILM is hindered by the availability of data labels and high labeling costs. In this work, we present an active learning model to reduce the labeling effort for NILM in industrial applications and implement it using the HIPE dataset based on a real production plant using a derived research process based on CRISP-DM [14]. In this sense, we are the first to combine active learning with NILM for an industrial setting. We evaluated the proposed method by applying it to the HIPE dataset to answer the stated RQ of how an active learning model performs compared to established supervised learning models in NILM systems for industrial applications. We apply budget sizes ranging from 0.1% to 5% of the available data as training data for the active learning model. We compared the disaggregation predictions with a benchmark using NILMTK and established supervised learning models trained on 100% of the available data. To tune the active learning model, we selected the best-performing query strategies and algorithms for the architecture of our models. By allowing our model to choose the data it learns from, we significantly reduce the number of labeled training instances required while achieving comparable disaggregation predictions. Our results indicate that a budget size of 1% is a good fit for real-world applications since the labeling effort is significantly reduced while maintaining an average (i.e., by up to 99% reduction) of over 70% accuracy compared to the benchmark. These results demonstrate practical relevance as the cost and accuracy concerns of NILM can be addressed and consequently may lead to broader adoption of NILM in industrial applications. This fosters energy efficiency by tracking energy consumption patterns and the identification of inefficient applications enabling smarter energy management. Further, our work can serve as a foundation for more active learning methods being applied in the context of NILM.

Appendices

Appendix A

See Table 5.

Table 5 Model card based on Kühl et al. [44]

General information	Problem statement	Performance comparison of an active learning model and established supervised learning models in NILM systems for industrial applications based on a classification problem		
	Data gathering	Pre-defined data set called HIPE, which contains smart meter readings of ten industrial applications and the main terminal over three months from a power-electronics plant run by the Karlsruhe Institute of Technology [14]		
	Sampling	Pool-based sampling (c.f. 2.1) using time frame from 2017–10-04 to 2017–12-21		
	Data quality	Generally high, it does not contain any missing values such as NULL or similar error values. But the data measurements are not equidistant		
	Data pre-processing methods	c.f. Section "Data description and preparation"		
	Feature engineering and vectorizing	c.f. Section "Data description and preparation"		
	Hardware and configuration for calculation	Linux Virtual Machine running Ubuntu 20.04 32 Cores, 32 GB RAM Python 3.10.4 using JupyterLab		
All	Performance evaluation measure for hyperparameter tuning	Disaggregation performance of all applications considered in the HIPE data set		
Unsupervised model (AE)	Parameter optimization	Yes	Search space	Layers [1, 10]; Batch size [2^2, 2^{10}]; Epochs [5, 100]
	Data split	Not applicable		
	Algorithm	Feed Forward Neural Network		
	Performance evaluation measure for training	MSE		
	Package	Keras API using Tensorflow 2.9.1		
	Additional information	Adam as optimizer; Exponential Linear Unit as activation functions for the encoder layers and a linear activation function for the bottleneck; Sigmoid activation function for the last layer of the decoder		
Supervised model (Seq2Seq)	Parameter optimization	Sensitivity analysis c.f. 5.2.; search space: 0.1%—5.0% budget size		
	Data split	78/22 split to ensure comparability to previous studies on the same data set		
	Algorithm	Seq2Seq, CO, Mean		
	Performance evaluation measure for training	RMSE, MAE, F-score, NDE		
	Package	NILMTK API v0.4.2		
	Additional information	Not applicable		

Appendix B

See Figs. 6, 7, 8.

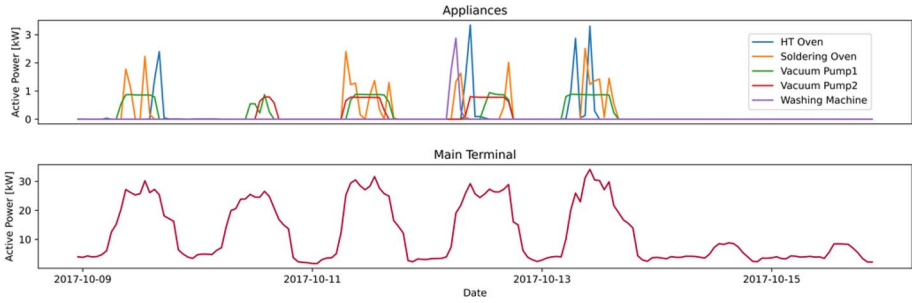

Fig. 6 Week load profile of HIPE applications

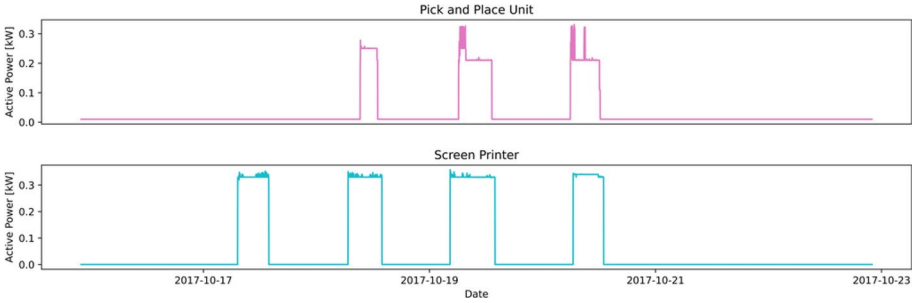

Fig. 7 Strong disaggregation of HIPE applications

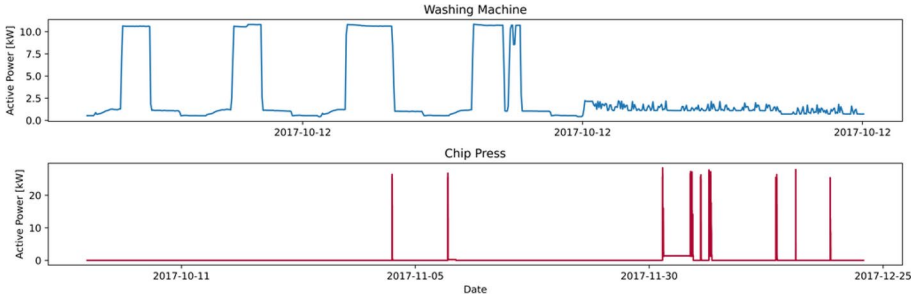

Fig. 8 Weak disaggregation of HIPE applications

Appendix C

See Table 6.

Table 6 Benchmark prediction results

Application	RMSE	F-score	MAE	NDE
Chip press	1012.80	0.26	285.91	0.48
Chipsaw	105.67	0.45	26.17	0.78
High temperature oven	275.97	0.50	72.19	**0.28**
Pick and place unit	**60.29**	**0.99**	32.02	0.81
Screen printer	83.45	0.68	35.59	0.69
Soldering oven	376.50	0.53	52.14	0.45
Vacuum oven	164.64	0.15	**6.50**	0.95
Vacuum pump 1	208.42	0.80	121.11	0.58
Vacuum pump 2	124.15	0.25	30.96	1.01
Washing machine	383.37	0.16	61.50	0.51
Mean	279.53	0.47	72.41	0.65

The bold printed value per column indicates the best average benchmark value in terms of the underlying PEM

Appendix D

See Fig. 9.

Fig. 9 Active learning model compared to benchmark using F-score

Appendix E

Further statistical significance testing conducting a one-way ANOVA test for all query strategies, a 1% budget, using the Seq2Seq algorithm and MAE as a PEM for each application, resulted in a p-value of 0.11 which shows a slight significance for a significance level of 0.1 (see Fig. 10).

In contrast, conducting another one-way ANOVA test for all algorithms, a 1% budget, the cluster query strategy and MAE as a PEM for each application, resulted in a p-value of 0.03 which shows strong statistical significance of the algorithm selection assuming again a significance level of 0.1.

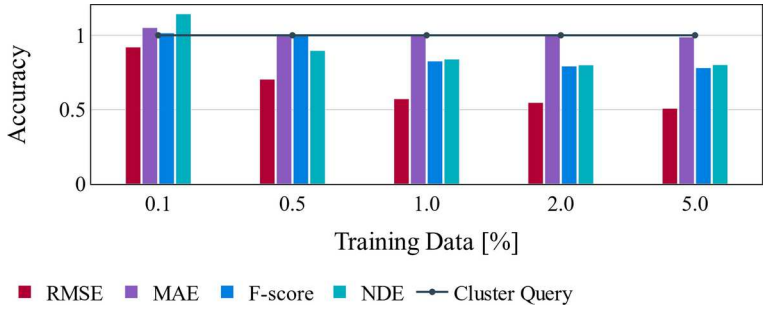

Fig. 10 Naive query prediction performance compared to cluster query

Author contributions
L.F.: Investigation, Data Curation, Methodology, Visualization, Writing—Original draft. D.L.: Conceptualization, Investigation, Data Curation, Supervision, Methodology, Visualization, Writing—Original draft, Writing—Reviewing & Editing. L.S.: Conceptualization, Investigation, Data Curation, Software, Supervision, Methodology, Visualization, Writing—Original draft, Writing—Reviewing & Editing. S.W.: Conceptualization, Investigation, Data Curation, Supervision, Methodology, Visualization, Writing—Original draft, Writing—Reviewing & Editing.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials
Data is provided within the manuscript or supplementary information files.

Declarations

Ethical Approval and Consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing Interests
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Received: 9 January 2025 / Accepted: 8 April 2025
Published online: 28 April 2025

References

1. Ali M, Prakash K, Hossain MA, Pota HR (2021) Intelligent energy management: Evolving developments, current challenges, and research directions for sustainable future. J Clean Prod 314:127904. https://doi.org/10.1016/j.jclepro.2021.127904
2. An J, Cho S (2015) Variational autoencoder based anomaly detection using reconstruction probability. Special Lecture on IE 2(1):1–18. http://dm.snu.ac.kr/static/docs/tr/snudm-tr-2015-03.pdf.
3. Anderson KD, Ocneanu A, Carlson DR, Rowe AG, Mario B (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 (7):1–5.
4. Angelis G-F, Timplalexis C, Krinidis S, Ioannidis D, Tzovaras D (2022) NILM applications: Literature review of learning approaches, recent developments and challenges. Energy and Buildings 261:111951. https://doi.org/10.1016/j.enbuild.2022.111951
5. Bao K, Ibrahimov K, Wagner M, Schmeck H (2018) Enhancing neural non-intrusive load monitoring with generative adversarial networks. Energy Inform. https://doi.org/10.1186/s42162-018-0038-y
6. Barth L, Hagenmeyer V, Ludwig N, Wagner D (2018) How much demand side flexibility do we need? In: Proceedings of the Ninth International Conference on Future Energy Systems; 12 06 2018 15 06 2018: Karlsruhe Germany. New York, NY, USA: ACM
7. Batra N, Kelly J, Parson O, Dutta H, Knottenbelt W, Rogers A et al. (2014a) NILMTK. In: Crowcroft J, Penty R, Le Boudec J-Y, Shenoy P. Proceedings of the 5th international conference on Future energy systems; 11 06 2014 13 06 2014: Cambridge United Kingdom. New York, NY, USA: ACM
8. Batra N, Kukunuri R, Pandey A, Malakar R, Kumar R, Krystalakos O et al (2019) Towards reproducible state-of-the-art energy disaggregation. In: Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation; 13 11 2019 14 11 2019: New York NY USA. New York, NY, USA: ACM
9. Batra N, Parson O, Berges M, Singh A, Rogers A (2014b) A comparison of non-intrusive load monitoring methods for commercial and residential buildings
10. Batra N, Singh A, Whitehouse K (2015) If You Measure It, Can You Improve It? Exploring The Value of Energy Disaggregation. In: Culler D, Agarwal Y, Mangharam R. Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments; 04 11 2015 05 11 2015: Seoul South Korea. New York, NY, USA: ACM

11. BDEW (2024) Nettostromverbrauch nach Verbrauchergruppen. Bundesverband der Energie- und Wasserwirtschaft
12. Bertolini M, Mezzogori D, Neroni M, Zammori F (2021) Machine Learning for industrial applications: A comprehensive literature review. Expert Syst Appl 175:114820. https://doi.org/10.1016/j.eswa.2021.114820
13. Beverungen D, Müller O, Matzner M, Mendling J, vom Brocke J (2019) Conceptualizing smart service systems. Electron Markets 29(1):7–18. https://doi.org/10.1007/s12525-017-0270-5
14. Bischof S, Trittenbach H, Vollmer M, Werle D, Blank T, Böhm K (2018) HIPE In: Proceedings of the Ninth International Conference on Future Energy Systems; 12 06 2018 15 06 2018: Karlsruhe Germany. New York, NY, USA: ACM;
15. Chen H, Fei X, Wang S, Lu X, Jin G, Li W, et al (2014) Energy Consumption Data Based Machine Anomaly Detection. In: 2014 Second International Conference on Advanced Cloud and Big Data; 20.11.2014–22.11.2014: Huangshan, China: IEEE
16. Chen K, Wang Q, He Z, Chen K, Hu J, He J (2018) Convolutional sequence to sequence non-intrusive load monitoring. J Eng 2018(17):1860–1864. https://doi.org/10.1049/joe.2018.8352
17. Chevrot A, Vernotte A, Legeard B (2022) CAE: Contextual auto-encoder for multivariate time-series anomaly detection in air transportation. Comput Secur 116:102652. https://doi.org/10.1016/j.cose.2022.102652
18. Cui J, Jin Y, Yu R, Okoye MO, Li Y, Yang J et al (2022) A robust approach for the decomposition of high-energy-consuming industrial loads with deep learning. J Clean Prod 349:131208. https://doi.org/10.1016/j.jclepro.2022.131208
19. Das S, Wong W-K, Dietterich T, Fern A, Emmott A (2016) Incorporating Expert Feedback into Active Anomaly Discovery. In: 2016 IEEE 16th International Conference on Data Mining (ICDM); 12.12.2016 - 15.12.2016: Barcelona, Spain: IEEE
20. Desai S, Alhadad R, Mahmood A, Chilamkurti N, Rho S (2019) Multi-state energy classifier to evaluate the performance of the NILM algorithm. Sensors (Basel, Switzerland). https://doi.org/10.3390/s19235236
21. Feng C, Tian P (2021) Time Series Anomaly Detection for Cyber-physical Systems via Neural System Identification and Bayesian Filtering. In: Zhu F. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining; 14 08 2021 18 08 2021: Virtual Event Singapore: ACM Special Interest Group on Management of Data; ACM Special Interest Group on Knowledge Discovery in Data. New York,NY,United States: Association for Computing Machinery
22. Finder I, Sheetrit E, Nissim N (2022) A time-interval-based active learning framework for enhanced PE malware acquisition and detection. Comput Secur 121:102838. https://doi.org/10.1016/j.cose.2022.102838
23. Fisher RA (1925) Statistical Methods for Research Workers. Oliver & Boyd, Edinburgh
24. Flores-García E, Hoon Kwak D, Jeong Y, Wiktorsson M (2024) Machine learning in smart production logistics: a review of technological capabilities. Int J Prod Res. https://doi.org/10.1080/00207543.2024.2381145
25. Garcia FD, Souza WA, Diniz IS, Marafão FP (2020) NILM-based approach for energy efficiency assessment of household appliances. Energy Inform. https://doi.org/10.1186/s42162-020-00131-7
26. Gehring J, Auli M, Grangier D, Yarats D, Dauphin YN (2017) Convolutional Sequence to Sequence Learning. Proceedings of the 34th International Conference on Machine Learning 1243–52. http://proceedings.mlr.press/v70/gehring17a.html?ref=https://githubhelp.com.
27. Goernitz N, Kloft M, Rieck K, Brefeld U (2013) Toward supervised anomaly detection. Journal of Artificial Intelligence Research 46:235–262. https://doi.org/10.1613/jair.3623
28. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. The MIT Press, Cambridge, Massachusetts, London, England
29. Gopinath R, Kumar M, Prakash Chandra Joshua C, Srinivas K (2020) Energy management using non-intrusive load monitoring techniques–State-of-the-art and future research directions. Sustain Cities Soc 62:102411. https://doi.org/10.1016/j.scs.2020.102411
30. Guo L, Wang S, Chen H, Shi Q (2020) A load identification method based on active deep learning and discrete wavelet transform. IEEE Access 8:113932–113942. https://doi.org/10.1109/ACCESS.2020.3003778
31. Hart GW (1992) Nonintrusive appliance load monitoring. Proc IEEE 80(12):1870–1891. https://doi.org/10.1109/5.192069
32. Holmegaard E, Baun Kjaergaard M (2016) NILM in an Industrial Setting: A Load Characterization and Algorithm Evaluation. In: 2016 IEEE International Conference on Smart Computing (SMARTCOMP); 18.05.2016 - 20.05.2016: St Louis, MO, USA: IEEE.
33. Holtz D, Kaymakci C, Leuthe D, Wenninger S, Sauer A (2025) A data-efficient active learning architecture for anomaly detection in industrial time series data. Flex Serv Manuf J. https://doi.org/10.1007/s10696-024-09588-0
34. Holweger J, Dorokhova M, Bloch L, Ballif C, Wyrsch N (2019) Unsupervised algorithm for disaggregating low-sampling-rate electricity consumption of households. Sustainable Energy, Grids and Networks 19:100244. https://doi.org/10.1016/j.segan.2019.100244
35. International Organization for Standardization (2018) ISO 50001:2018: Energy management systems-Requirements with guidance for use: Second Edition; 2.
36. Jin X. Active Learning Framework for Non-Intrusive Load Monitoring: Preprint; 2016. National Renewable Energy Lab. (NREL), Golden, CO (United States) NREL/CP-5500–66273.
37. Jing J, Di H, Wang T, Jiang N, Xiang Z (2025) Optimization of power system load forecasting and scheduling based on artificial neural networks. Energy Inform. https://doi.org/10.1186/s42162-024-00467-4
38. Kalinke F, Bielski P, Singh S, Fouché E, Böhm K (2021) An Evaluation of NILM Approaches on Industrial Energy-Consumption Data. In: Proceedings of the Twelfth ACM International Conference on Future Energy Systems; 28 06 2021 02 07 2021: Virtual Event Italy. New York, NY, USA: ACM
39. Kaymakci C, Wenninger S, Sauer A (2021) Energy anomaly detection in industrial applications with long short-term memory-based autoencoders. Procedia CIRP 104:182–187. https://doi.org/10.1016/j.procir.2021.11.031
40. Kelly J, Knottenbelt W (2015) The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes. Scientific data 2:150007. https://doi.org/10.1038/sdata.2015.7
41. Klemenjak C (2018) On performance evaluation and machine learning approaches in non-intrusive load monitoring. Energy Inform. https://doi.org/10.1186/s42162-018-0051-1
42. Klemenjak C, Makonin S, Elmenreich W (2021) Investigating the performance gap between testing on real and denoised aggregates in non-intrusive load monitoring. Energy Informatics. https://doi.org/10.1186/s42162-021-00137-9
43. Kolter ZJ, Johnson MJ (2011) REDD: A Public Data Set for Energy Disaggregation Research. Workshop on data mining applications in sustainability (SIGKDD) 25:59–62
44. Kühl N, Hirt R, Baier L, Schmitz B, Satzger G (2021) How to conduct rigorous supervised machine learning in information systems research the supervised machine learning report card. CAIS 48(1):589–615.https://doi.org/10.17705/1CAIS.04845

45. Kumar P, Gupta A (2020) Active learning query strategies for classification, regression, and clustering: a survey. J Comput Sci Technol 35(4):913–945. https://doi.org/10.1007/s11390-020-9487-4
46. Langevin A, Carbonneau M-A, Cheriet M, Gagnon G (2022) Energy disaggregation using variational autoencoders. Energy and Buildings 254:111623. https://doi.org/10.1016/j.enbuild.2021.111623
47. Li Y, Guo L (2007) An active learning based TCM-KNN algorithm for supervised network intrusion detection. Comput Secur 26(7–8):459–467. https://doi.org/10.1016/j.cose.2007.10.002
48. Liu Y, Ma J, Xing X, Liu X, Wang W (2022) A home energy management system incorporating data-driven uncertainty-aware user preference. Appl Energy 326:119911. https://doi.org/10.1016/j.apenergy.2022.119911
49. Maasoumy M, Sanandaji BM, Poolla K, Vincentelli AS (2013) BERDS-BERkeley EneRgy Disaggregation Data Set. Proceedings of the Workshop on Big Learning at the Conference on Neural Information Processing Systems (NIPS) (7).
50. Martins PBdM, Nascimento VB, Freitas ARd, Bittencourt e Silva P, Pinto RGD (2018) Industrial Machines Dataset for Electrical Load Disagreggation. IEEE DataPort.https://doi.org/10.21227/cg5v-dk02
51. Mehrotra KG, Mohan CK, Huang H (2017) Anomaly Detection Principles and Algorithms. Springer International Publishing, Cham
52. Miller C, Meggers F (2017) The Building Data Genome Project: An open, public data set from non-residential building electrical meters. Energy Procedia 122:439–444. https://doi.org/10.1016/j.egypro.2017.07.400
53. Müller O, Junglas I, vom Brocke J, Debortoli S (2016) Utilizing big data analytics for information systems research: challenges, promises and guidelines. Eur J Inf Syst 25(4):289–302. https://doi.org/10.1057/ejis.2016.2
54. Nalmpantis C, Vrakas D (2019) Machine learning approaches for non-intrusive load monitoring: from qualitative to quantitative comparation. Artif Intell Rev 52(1):217–243. https://doi.org/10.1007/s10462-018-9613-7
55. Newell R, Raimi D, Villanueva S, Prest B (2021) Global Energy Outlook 2021: Pathways from Paris. Ressources for the Future (8). https://media.rff.org/documents/RFF_GEO_2021_Report_1.pdf.
56. Norford LK, Leeb SB (1996) Non-intrusive electrical load monitoring in commercial buildings based on steady-state and transient load-detection algorithms. Energy and Buildings 24(1):51–64. https://doi.org/10.1016/0378-7788(95)00958-2
57. Pang G, Shen C, Cao L, van Hengel A, den (2022) Deep Learning for Anomaly Detection. ACM Comput Surv 54(2):1–38. https://doi.org/10.1145/3439950
58. Pereira L, Nunes N (2018) Performance evaluation in non-intrusive load monitoring: Datasets, metrics, and tools—A review. WIREs Data Min Knowl Discovery. https://doi.org/10.1002/widm.1265
59. Pimentel T, Monteiro M, Veloso A, Ziviani N (2020) Deep Active Learning for Anomaly Detection. In: 2020 International Joint Conference on Neural Networks (IJCNN): 2020 conference proceedings; 7/19/2020 - 7/24/2020: Glasgow, United Kingdom: Institute of Electrical and Electronics Engineers; IEEE Computational Intelligence Society; International Neural Network Society. Piscataway, NJ, USA: IEEE
60. Ramadan R, Huang Q, Bamisile O, Zalhaf AS (2022) Intelligent home energy management using Internet of Things platform based on NILM technique. Sustainable Energy, Grids and Networks 31:100785. https://doi.org/10.1016/j.segan.2022.100785
61. Reddy R, Garg V, Pudi V (2020) A feature fusion technique for improved non-intrusive load monitoring. Energy Inform. https://doi.org/10.1186/s42162-020-00112-w
62. Ren P, Xiao Y, Chang X, Huang P-Y, Li Z, Gupta BB et al (2022) A survey of deep active learning. ACM Comput Surv 54(9):1–40. https://doi.org/10.1145/3472291
63. Rožanec JM, Trajkova E, Dam P, Fortuna B, Mladenić D (2022) Streaming machine learning and online active learning for automated visual inspection. IFAC-PapersOnLine 55(2):277–282. https://doi.org/10.1016/j.ifacol.2022.04.206
64. Saba CS, Ngepah N (2022) Convergence in renewable energy sources and the dynamics of their determinants: An insight from a club clustering algorithm. Energy Rep 8:3483–3506. https://doi.org/10.1016/j.egyr.2022.01.190
65. Settles B (2010) Active Learning Literature Survey. University of Wisconsin, Madison, 52. https://minds.wisconsin.edu/handle/1793/60660.
66. Shmueli K, Koppius OR (2011) Predictive analytics in information systems research. MIS Q 35(3):553. https://doi.org/10.2307/23042796
67. Tanoni G, Principi E, Squartini S (2024a) Non-Intrusive Load Monitoring in industrial settings: A systematic review. Renew Sustain Energy Rev 202:114703. https://doi.org/10.1016/j.rser.2024.114703
68. Tanoni G, Sobot T, Principi E, Stankovic V, Stankovic L, Squartini S (2024b) A weakly supervised active learning framework for non-intrusive load monitoring. Integrated Computer-Aided Engineering 32(1):39–56. https://doi.org/10.3233/ICA-240738
69. Timplalexis C, Angelis G-F, Krinidis S, Ioannidis D, Tzovaras D (2022) Low frequency residential non-intrusive load monitoring based on a hybrid feature extraction tree-learning approach. Energy Sources, Part A: Recovery, Utilization, and Environmental Effects 44(1):493–514. https://doi.org/10.1080/15567036.2022.2046663
70. Todic T, Stankovic V, Stankovic L (2023) An active learning framework for the low-frequency non-intrusive load monitoring problem. Appl Energy 341:121078. https://doi.org/10.1016/j.apenergy.2023.121078
71. van Leeuwen R, Koole G (2023) Anomaly detection in univariate time series incorporating active learning. Journal of Computational Mathematics and Data Science 6:100072. https://doi.org/10.1016/j.jcmds.2022.100072
72. Watson, Boudreau, Chen (2010) Information Systems and Environmentally Sustainable Development: Energy Informatics and New Directions for the IS Community. MIS Quarterly 34(1):23.https://doi.org/10.2307/20721413
73. Werthen-Brabants L, Dhaene T, Deschrijver D (2022) Uncertainty quantification for appliance recognition in non-intrusive load monitoring using Bayesian deep learning. Energy and Buildings 270:112282. https://doi.org/10.1016/j.enbuild.2022.112282
74. Wirth R, Hipp J (2000) CRISP-DM: Towards a standard process model for data mining. In: Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining
75. Wu Z, Wang C, Peng W, Liu W, Zhang H (2021) Non-intrusive load monitoring using factorial hidden markov model based on adaptive density peak clustering. Energy and Buildings 244:111025. https://doi.org/10.1016/j.enbuild.2021.111025
76. Yu Z, Kaplan Z, Yan Q, Zhang N (2021) Security and privacy in the emerging cyber-physical world: a survey. IEEE Communications Surveys & Tutorials 23(3):1879–1919. https://doi.org/10.1109/COMST.2021.3081450

77. Zhang C, Zhong M, Wang Z, Goddard N, Sutton C (2018) Sequence-to-Point Learning With Neural Networks for Non-Intrusive Load Monitoring. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1). https://ojs.aaai.org/index.php/aaai/article/view/11873.https://doi.org/10.1609/aaai.v32i1.11873
78. Zhang J, Lyu Y, Li Y, Geng Y (2022) Digital economy: An innovation driving factor for low-carbon development. Environ Impact Assess Rev 96:106821. https://doi.org/10.1016/j.eiar.2022.106821
79. Zhao T, Zheng Y, Wu Z (2022) Improving computational efficiency of machine learning modeling of nonlinear processes using sensitivity analysis and active learning. Digital Chemical Engineering 3:100027. https://doi.org/10.1016/j.dche.2022.100027
80. Zheng Z, Shafique M, Luo X, Wang S (2024) A systematic review towards integrative energy management of smart grids and urban energy systems. Renew Sustain Energy Rev 189:114023. https://doi.org/10.1016/j.rser.2023.114023
81. Zhou X, Feng J, Li Y (2021) Non-intrusive load decomposition based on CNN–LSTM hybrid deep learning model. Energy Rep 7:5762–5771. https://doi.org/10.1016/j.egyr.2021.09.001
82. Zhu Y, Yang K (2019) Tripartite active learning for interactive anomaly discovery. IEEE Access 7:63195–63203. https://doi.org/10.1109/ACCESS.2019.2915388
83. Zoha A, Gluhak A, Imran MA, Rajasegarar S (2012) Non-intrusive load monitoring approaches for disaggregated energy sensing: a survey. Sensors (Basel, Switzerland) 12(12):16838–16866. https://doi.org/10.3390/s121216838

Publisher's Note