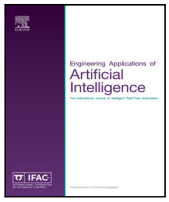
Engineering Applications of Artificial Intelligence 133 (2024) 108589

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Research paper

# Human in the loop active learning for time-series electrical measurement data

Tamara Sobot *, Vladimir Stankovic, Lina Stankovic

*Department of Electronic and Electrical Engineering, University of Strathclyde, United Kingdom*

ARTICLE INFO

Dataset link: https://pureportal.strath.ac.uk/en/datasets/refit-electrical-load-measurements-cleaned

*Keywords:*
Active learning
Deep learning
Human in the loop
Time series data
Labelling

ABSTRACT

Advanced machine learning algorithms require large datasets, along with good-quality labels to reach state-of-the-art performance. Although measurements themselves can often be easily available, the labelling process is usually a bottleneck. To address this, active learning approaches exploit the fact that different samples provide varying levels of information to the algorithm. However, these approaches often rely on several unrealistic assumptions — an oracle is assumed to provide error-free labels, all at the same cost and effort. We propose novel active learning-based methods for classification of time series measurements, typically obtained from sensors continuously measuring highly fluctuating environmental conditions including electricity consumption, and demonstrate their effectiveness for home energy management applications, where data labelling is a challenge. A new acquisition function is proposed, which accounts for both model and labelling uncertainty and class balancing. A stopping criterion is designed to stop the active learning process after an optimal point is achieved, to reduce labelling effort. We assess the effect of labelling errors on classification performance and propose two ways of mitigating their effects: (i) a re-labelling mechanism based on similarity of provided labels; (ii) a revised loss function based on confidence levels provided by experts. We validate our contributions for energy disaggregation task in a real-world scenario with three application domain experts. Our results show that the proposed methodology significantly improves performance of algorithms transferred to unseen domains with reduced number of labelled samples — from 61% reduction for dishwasher to 93% reduction for kettle.

## 1. Introduction

Advanced inference approaches, especially for dynamic time-varying measurements, require large, well-labelled datasets to achieve good performance when training a model in a supervised manner. In many applications, although raw measurements are easy to collect, the labelling process is time consuming or expensive, thus hindering use of the data. One way to significantly reduce labelling effort is via active learning.

Active learning (Settles, 2009; Ren et al., 2021) is designed to minimise the amount of data that needs labelling, by intelligently selecting (a small amount) of most valuable data samples to label among all the available data. Active learning builds on the fact that not all data samples are equally important for training of the model, and that there is redundancy within data, i.e., many samples are highly correlated, so providing labels for some of them eliminates the need to label the rest.

With the recent need towards trustworthy AI (European Commission and Directorate-General for Communications Networks, Content and Technology, 2019) and the need for humans to be involved in the learning process, human in the loop machine learning (HITL-ML) has grown in popularity. During active learning, one of the approaches towards HITL-ML, the AI system remains in control of the learning process and humans are treated as oracles to annotate unlabelled data (Mosqueira-Rey et al., 2023). However oracles are assumed to provide absolutely true labels, without any errors, all with the same effort and at the same cost (see Budd et al., 2021 for a review of active learning approaches for medical image analysis — most of active learning methods assume an oracle). This is a very unrealistic assumption — human error during labelling will be (unintentionally) introduced, especially for challenging to label samples (e.g., noisy samples) and time-series samples that are not always visually interpretable. Only a few studies have reported active learning system results where users/experts are recruited to provide labels during the active learning process. For example, Ghai et al. (2021) includes people in the labelling process, for an income prediction task using linear regression, investigating if active learning could boost their trust and

* Corresponding author.
*E-mail addresses:* tamara.todic@strath.ac.uk (T. Sobot), vladimir.stankovic@strath.ac.uk (V. Stankovic), lina.stankovic@strath.ac.uk (L. Stankovic).

https://doi.org/10.1016/j.engappai.2024.108589
Received 26 June 2023; Received in revised form 6 March 2024; Accepted 5 May 2024

confidence in AI, depending on their level of familiarity with AI, and their willingness to engage with the process. However, studies that actually deploy active learning concept focus mainly on social aspects of active learning and human–computer interaction, e.g., trust, while using toy active learning algorithms. The closer interaction between users and learning systems, especially where human users select and annotate examples to modify model features in an incremental fashion is termed interactive machine learning (Mosqueira-Rey et al., 2023). HITL-ML has not been explored in energy management related applications despite the acknowledged role of consumers on energy end use in order to meet European Green Deal Ambition goals related to bringing greenhouse gas emissions to the levels of 1990 by 2030 (Anon, 2022).

In this paper we propose a novel HITL-ML approach which sits between active learning and interactive learning, where the machine selects examples to query, then through a user interface which shows the time-series electrical signal under questions, a human expert manually labels such examples. Due to the nature of the variable electrical signals belonging to the same class, we show that human uncertainty is possible and the HITL-ML learns incrementally until a stopping criterion is met. Our approach is demonstrated for classification of time-series electrical data for smart home energy management application. More specifically, we tackle the problem of energy disaggregation from widely available smart meter aggregate measurements, but suffers from unavailability of labelled samples (i.e., labelled appliances contributing to the aggregate at each sampling point).

Energy disaggregation or Non-Intrusive Load Monitoring (NILM), is a useful tool for inferring fine-grained, appliance-level information from smart meter measurements. It consists of separating the aggregate energy consumption of a building into its sub-components, i.e., electricity consumption of individual appliances. This fine-grained information is used to provide energy conservation recommendations, automating load shifting to minimise carbon footprint and inform demand response programmes. Numerous approaches to NILM have been proposed and as per recent review papers (Huber et al., 2021; Angelis et al., 2022) deep learning-based NILM algorithms dominate the landscape due to ease of implementation (bypassing feature engineering) and state-of-the-art performance. However, deep learning methods require huge amounts of labelled data for training to achieve good performance, and acquiring reliable labelled data is resource-intensive. Labelling is either performed via submetering of individual appliances, which is intrusive and costly, or visually by recognising appliance signatures in the aggregate sample obtained from smart meter measurements. Moreover, once trained with labelled data from one domain, deep learning algorithms usually underperform when conditions change (i.e., appliance signatures change due to wear-and-tear; new appliances are introduced into the house; the number of occupants changes, etc.) or when transferred to a new domain (i.e., a new, unseen house where labelled data to train the models is unavailable) (Kaselimi et al., 2022). To overcome this issue, transfer learning approaches have been adopted, but, they often assume availability of new high-quality labelled data from the target domain to fine-tune the models (Li et al., 2023), which is resource-intensive to obtain.

Building on our prior work (Todic et al., 2023) that proposes a framework for active learning for low-frequency model-based NILM, assuming perfect error-free labelling, in this paper, several contributions are made to minimise labelling effort while accounting for possible errors during labelling process.

Namely, the main contributions of this paper are:

- Design of a new acquisition function based on maximum a posteriori hypothesis testing, that balances classes while taking into account model uncertainty (Section 3.1).
- A new stopping criterion once the optimal performance is approached to minimise labelling effort (Section 3.2).
- Loss function weighted with experts confidence to control the impact of potentially erroneous labels (Section 3.3).
- A mechanism for returning potentially wrongly labelled samples to be relabelled by the expert (Section 3.4).
- Assessment of the impact of erroneous labels introduced during expert labelling (Section 5.1.2).
- Evaluation of human-in-the-loop (expert) presence during model training and evaluation, providing labels and their confidence, in a real-world scenario using a designed user-friendly interface (Section 5.1.3).

We use the well documented and widely used, publicly available REFIT dataset (Murray et al., 2017), and an efficient transformer-based NILM architecture, ELECTRIcity (Sykiotis et al., 2022), though other deep learning NILM models, including the one used in Todic et al. (2023), can be used. Section 2 provides the background on active learning and NILM. The methodology, including acquisition functions used, exploiting experts' confidence and re-labelling mechanism, is presented in Section 3. Section 4 describes experiments — dataset and Deep Neural Network (DNN) NILM model used, and user interface. Results and discussion are presented in Section 5, before the conclusion in Section 6.

## 2. Related work/background

### 2.1. Active learning

The goal of active learning (Settles, 2009) is to reduce the amount of labelled data needed to train models. It is an iterative process, where an initial model $m_0$ is trained using a limited set of labelled data $\boldsymbol{D}_{\text{pt}}$. The prediction is then performed on a large pool of data $\boldsymbol{D}_{\text{pool}}$ where labels are not available, and the acquisition function $q(\cdot)$ is used to select samples $\boldsymbol{Q} \subseteq \boldsymbol{D}_{\text{pool}}$ that are worth including in training, i.e., that satisfy some informativeness criteria, as in Todic et al. (2023), diversity criteria (Ash et al., 2019), or both (Prabhu et al., 2021; Kothandaraman et al., 2023). Labels are requested for the chosen samples, and after they are available, those samples are included into a new fine-tuning (or re-training) set $\boldsymbol{D}_{\text{ft}}$. When retrained or fine-tuned on $\boldsymbol{D}_{\text{ft}}$, the model uses new knowledge to query more data. The loop runs until a stopping criterion has been met, as shown in Algorithm 1, where algorithm *train* performs either re-training of the entire model or fine-tuning the last layers.

An overview of deep active learning, explored recently for various types of problems, such as medical image analysis (Budd et al., 2021), and natural language processing (Zhang et al., 2022), is provided in a recent survey (Ren et al., 2021).

**Algorithm 1** Active learning

$i = 1$ - active learning iteration
$m_i$ - DNN-based model at iteration $i$ ▷ $m_0$ - pre-trained DNN model
$q(\cdot)$ - acquisition function
$\boldsymbol{Q}_i$ - set of samples queried at iteration $i$
$\boldsymbol{D}_{\text{pool}}$ - query pool
$\boldsymbol{D}_{\text{ft}} = \emptyset$ - fine-tuning set
$S$ - stopping criterion met (Boolean flag)
**while** not $S$ **do**
  $\boldsymbol{Q}_i \leftarrow q(m_{i-1}, \boldsymbol{D}_{\text{pool}})$
  $\boldsymbol{D}_{\text{pool}} \leftarrow \boldsymbol{D}_{\text{pool}} \setminus \boldsymbol{Q}_i$
  $\boldsymbol{D}_{\text{ft}} \leftarrow \boldsymbol{D}_{\text{ft}} \cup \boldsymbol{Q}_i$
  $m_i \leftarrow train(m_0, \boldsymbol{D}_{\text{ft}})$
  $i \leftarrow i + 1$
**end while**

#### 2.1.1. Acquisition functions

Acquisition function is used to select the most worthy data samples from $\boldsymbol{D}_{\text{pool}}$ to be queried, labelled and added to $\boldsymbol{D}_{\text{ft}}$, by ranking samples belonging to query pool $\boldsymbol{D}_{\text{pool}}$ based on informativeness or diversity criteria (Ren et al., 2021). For the classification problem, the model

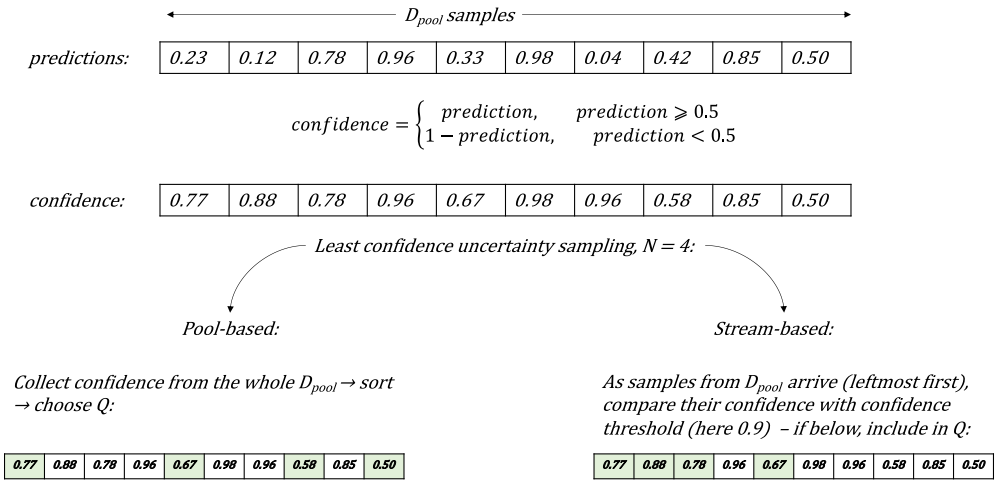



**Fig. 1.** Difference between pool- and stream-based uncertainty sampling on an example of binary classification with 10 data samples in the query pool out of which 4 should be selected for query. Samples belonging to $\boldsymbol{Q}$ are highlighted in green.

produces a vector containing probabilities that a data sample belongs to each of the possible classes/labels. Common approaches use those class probabilities to estimate model uncertainty (e.g., as in Todic et al., 2023). This approach is commonly referred to as *least confidence uncertainty sampling*, and can be implemented in pool- and stream-based fashion. If in the pool-based fashion, then all samples from query pool $\boldsymbol{D}_{\text{pool}}$ are evaluated and then the best subset, $\boldsymbol{Q}$, are selected. That is, it is assumed that the whole query pool is available at the moment of query. If in the stream-based fashion, then data samples are considered to arrive in a stream, and the whole query pool is not available at query time. Therefore, a predefined informativeness threshold is applied to each data sample as it arrives, and if informativeness of the sample exceeds the threshold, then the sample is considered informative enough and it is included in query $\boldsymbol{Q}$, and otherwise it is not. An example in Fig. 1 shows how uncertainty sampling works in the pool- and stream-based fashion — the task in the example is binary classification, query pool contains 10 samples out of which 4 are selected for the query.

Selecting a batch of data samples to label independently leads to redundancy because many similar highly-correlated samples would be queried. Therefore, acquisition strategies that account for both informativeness and diversity among queried data samples have been developed. For example, BatchBALD (Kirsch et al., 2019) looks at mutual information between a sequence of samples and model parameters. Although it works well with small datasets, it underperforms for large ones (Sener and Savarese, 2017; Todic et al., 2023). BADGE (Ash et al., 2019) queries samples that give high-magnitude penultimate layer gradients of different directions if the predicted label would be the true one (i.e., if pseudo-labels are used to compute gradients). Samples are chosen via k-means++ initialisation algorithm on the obtained gradient embeddings. This approach needs computation of gradients for each sample in the query pool, which is resource-intensive. CLUE (Prabhu et al., 2021) scales the activations of the penultimate layer of the network with the entropy of the output as uncertainty measure. Obtained embeddings are clustered using k-means algorithm, and then samples closest to cluster centres are chosen. This method depends heavily on the clustering algorithm initialisation, and also on the convergence of the clustering algorithm. Acquisition function used in SALAD (Kothandaraman et al., 2023) combines l-2 norms of gradients computed using pseudo-labels as in BADGE (Ash et al., 2019), and entropy of the prediction as an uncertainty measure. Sum of the two components is greedily maximised to choose samples for query. This approach avoids clustering, but the whole SALAD framework contains pre-trained network, target network, as well as guided attention transfer network, which are all used throughout the process, and which can be demanding.

#### 2.1.2. Stopping criterion

Active learning is usually performed in an iterative manner, where, in each iteration, the user provides a set of new labels that are used to retrain the model. At some point, newly labelled data supplied to the model will either not anymore improve the performance, or even worse it can start degrading the performance due to overfitting. Hence, it is important to stop the iterative labelling process on time. Setting a threshold on the achieved performance, or observing performance improvement smaller than a threshold (Ueno et al., 2021), can be used to determine when to stop if this is practically possible. Also, confidence levels of the model can be exploited (Zhu et al., 2010), or agreement between the models from a couple of previous iterations (Bloodgood and Vijay-Shanker, 2014).

If active learning is conducted in small steps (i.e., in each iteration a small number of labelled samples are passed to the model), which is the case in near real-time applications and is the case in this paper, it is difficult to use stopping criteria based on measuring the improvement between two successive iterations, because small to no improvement can be observed long before the optimal point of active learning is achieved. Furthermore, measuring model agreement requires saving either several models from previous iterations, or their outputs for the data used to determine when to stop, which is resource inefficient.

#### 2.1.3. Human-in-the-loop

Although active learning has gained in popularity recently, there are still very few papers where users or experts are included in the loop, to verify the use of active learning in a real-world scenario — user input is usually simulated (Ren et al., 2021). Several studies including human-in-the-loop (Budd et al., 2021) use crowdsourcing platforms to obtain annotations for biomedical image processing related tasks. Although not implementing active learning, they offer several useful conclusions. For example, Cheplygina et al. (2016) points out that the annotation tool should not offer too many degrees of freedom, and that instructions should be simple and clear to the annotators. Labels provided by non-expert annotators can show medium to high correlation with labels provided by experts, especially if crowdsourced labels are aggregated. A study from Tinati et al. (2017) argues that gamification of annotation tasks can be very beneficial, but, the method is of limited generalisability.

### 2.2. Active learning for NILM

NILM consists of extracting per-appliance electricity consumption from the aggregate electricity consumption of a building. To ensure

wider adoption, NILM that considers practical challenges, such as transferability, reliability, scalability, safety, privacy, and trustworthiness, are required (Kaselimi et al., 2022). Active learning can address many of these practical challenges (Todic et al., 2023). By smartly selecting small amounts of data for training, scalability and transferability can be improved; new data can be added to the model by the end-user, without the need to export the data out of the house, tackling privacy and security issues; and, moreover, human-in-the-loop algorithm design can improve their trust and boost confidence when using AI algorithms in everyday life (Ghai et al., 2021).

Despite the fact that active learning is a very popular and effective approach for relaxing labelling effort as demonstrated, e.g., in Wang et al. (2022) for anomaly detection in time-series data, in Gu et al. (2021) for transfer to a domain with known data distribution, and in Martins et al. (2023) based on a meta-learning approach for feature extraction and uncertainty threshold tuning, there have only been a few attempts of leveraging active learning for the NILM problem (see Liebgott and Yang, 2017, Fatouh et al., 2018, and Guo et al., 2020). As reviewed in detail in Todic et al. (2023), prior work that explored active learning for event-based, high-frequency NILM (Liebgott and Yang, 2017, Fatouh et al., 2018, and Guo et al., 2020), have not considered stopping criteria and batch-aware acquisition functions, and has exclusively assumed an oracle providing always-correct labels.

A deep active learning framework for model-based, low-frequency NILM was been proposed in Todic et al. (2023). The ability of active learning for reducing labelling effort and improving performance of models pre-trained with large, publicly available datasets when transferred to a new environment was demonstrated. However, labels are considered to be provided by an oracle — they are error-free, all obtained at the same cost. Stopping criteria was also not considered, and the samples, once labelled were not returned for re-labelling.

In this paper, a detailed analysis of the active learning for model-based low-frequency NILM is provided. The pitfalls of active learning are explored, such as its vulnerability to errors in the labels, and strategies to overcome this are investigated. A real-world scenario with human-in-the-loop is deployed, where experts provide labels. Additionally, we quantify the effect of errors potentially injected throughout the active learning process, and consider mitigating measures such as exploitation of user confidence when giving a label and returning possibly wrong samples back, offering the possibility for re-labelling.

## 3. Methodology

In this section we describe the proposed active learning approach, illustrated in Fig. 2. As in Todic et al. (2023), Algorithm 1, described in Section 2.1, is used to select samples to query. Four main contributions to Todic et al. (2023) are made. First, a new acquisition function $q(\cdot)$ is proposed based on hypothesis testing to ensure diversity of labels in terms of reliability and classes (see Section 3.1). Second, a stopping criterion is introduced when all "uncertain" samples are exhausted (see Section 3.2). Third, confidence levels are included during model learning within the fine-tuning step (Section 3.3), to account for experts' confidence about provided labels and mitigate the effect of errors introduced for hard-to-label samples. Finally, after the fine-tuning step, an additional step for returning potentially wrongly labelled data samples back to experts for re-labelling is proposed (Section 3.4).

### 3.1. Acquisition function

Traditional uncertainty-based acquisition strategies for selecting samples to label tend to first query windows of samples containing appliance activations, i.e., positive samples (Todic et al., 2023). This leads to a very unbalanced set after labelling, containing predominantly positive samples. To keep the diversity of queried samples, both in terms of classes (all classes should be well represented) and model uncertainty (most uncertain samples should be queried), a new acquisition function based on maximum a posteriori (MAP) hypothesis testing is proposed next.

Let $\hat{y}$ be a realisation of a random variable $\hat{Y} \in [0,1]$ denoting the model output (0 = appliance if off; 1 = appliance is on). Let us consider two hypotheses: hypothesis $H_0$ corresponding to the appliance being in off-state, and hypothesis $H_1$ corresponding to the appliance being in on-state. Suppose that prior probabilities of both states are known, i.e., $P(H_0)$ and $P(H_1)$, as well as probability density distributions of model output $\hat{y}$ under the two hypotheses, i.e., $f_{\hat{Y}}(\hat{y}|H_0)$ and $f_{\hat{Y}}(\hat{y}|H_1)$.

Then, after applying Bayes' rule, posterior probabilities of hypotheses $H_0$ and $H_1$ are obtained as:

$$P(H_i|\hat{Y}=\hat{y}) = \frac{f_{\hat{Y}}(\hat{y}|H_i)\cdot P(H_i)}{f_{\hat{Y}}(\hat{y})}, i \in \{0,1\}. \tag{1}$$

Using the MAP test, the winning hypothesis will be the one that maximises (1). Since the denominator is the same for both hypotheses, hypothesis $H_0$ is chosen if and only if:

$$f_{\hat{Y}}(\hat{y}|H_0)\cdot P(H_0) > f_{\hat{Y}}(\hat{y}|H_1)\cdot P(H_1). \tag{2}$$

Otherwise, hypothesis $H_1$ is chosen.

The model output value $\hat{y}^*$ for which posterior probabilities of the two hypotheses, $H_0$ and $H_1$, are the same, i.e.,

$$f_{\hat{Y}}(\hat{y}^*|H_0)\cdot P(H_0) = f_{\hat{Y}}(\hat{y}^*|H_1)\cdot P(H_1) \tag{3}$$

is considered the most challenging model output value to make a decision. Therefore, model output space $[0,1]$ is divided into three regions: likely negative model predictions ($H_0$ chosen; model output value close to 0), likely positive model predictions ($H_1$ chosen; model output value close to 1), and uncertain model predictions (model output value close to $\hat{y}^*$ where posterior probabilities for $H_0$ and $H_1$ are equal). See Fig. 3 for illustration. Each point in the model output space is assigned to one of three regions depending on its proximity to 0, $\hat{y}^*$, and 1. Samples to be queried are taken from all three regions as per equation:

$$\begin{aligned}
\boldsymbol{Q}_i &= \boldsymbol{Q}_{i,\ likely\ negative} \cup \boldsymbol{Q}_{i,\ uncertain} \cup \boldsymbol{Q}_{i,\ likely\ positive} \\
\boldsymbol{Q}_{i,\ likely\ negative} &\subset \{s \in \boldsymbol{D}_{pool} | y = m_{i-1}(s) \in (0, \frac{\hat{y}^*}{2})\} \\
\boldsymbol{Q}_{i,\ uncertain} &\subset \{s \in \boldsymbol{D}_{pool} | y = m_{i-1}(s) \in (\frac{\hat{y}^*}{2}, \frac{1+\hat{y}^*}{2})\} \\
\boldsymbol{Q}_{i,\ likely\ positive} &\subset \{s \in \boldsymbol{D}_{pool} | y = m_{i-1}(s) \in (\frac{1+\hat{y}^*}{2}, 1)\}
\end{aligned} \tag{4}$$

where query from the current iteration $i$ is denoted by $\boldsymbol{Q}_i$. $\boldsymbol{D}_{pool}$ is query pool, $s$ denotes samples belonging to the query pool, $m_{i-1}$ is the model from previous active learning iteration, and $y$ is the model output for sample $s$. The number of samples from each region is controlled by hyper-parameters.

Since off-state of an appliance is more frequent than on-state (that is, most appliance are not used continuously), point $\hat{y}^*$ is expected to be closer to 1 than to 0 (see the example in Fig. 3), so samples containing measurements while appliance is turned on are favoured by this strategy, which is beneficial to NILM algorithms, as discussed later in Section 5.1. Most of queried samples therefore come from the uncertain region as defined above (the number is controlled by a hyper-parameter), but to prevent model from forgetting, samples are also taken from the two likely (positive/negative) regions.

### 3.2. Stopping criterion

Stopping criteria usually rely on comparison of performance across subsequent active learning iterations (e.g., in Ueno et al., 2021) or on agreement of models in subsequent iterations (e.g., in Bloodgood and Vijay-Shanker, 2014 and Zhu et al., 2010). To avoid the need to store the models from multiple iterations and compare them, or to store the model outputs or uncertainty levels from multiple iterations, which can be resource-intensive, a stopping criterion relying on confidence of the model from a single, current iteration is designed.

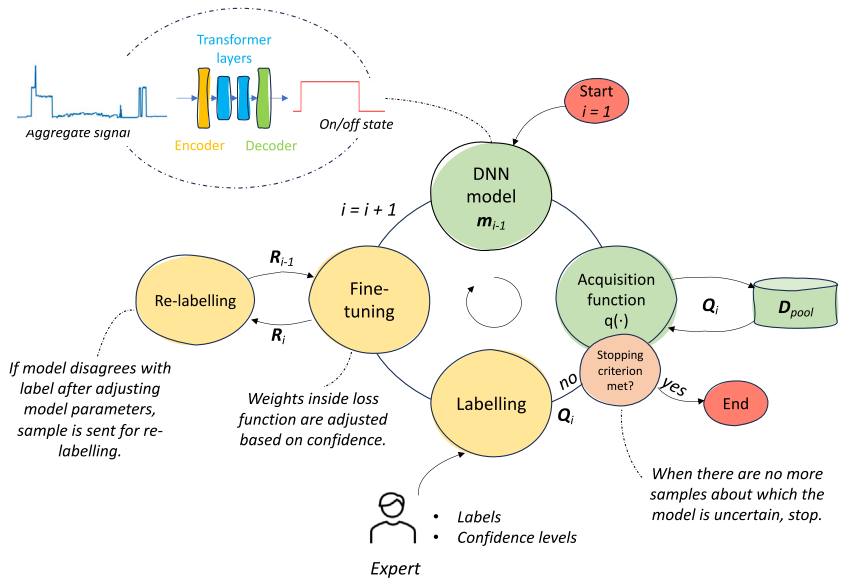



**Fig. 2.** Active learning framework.

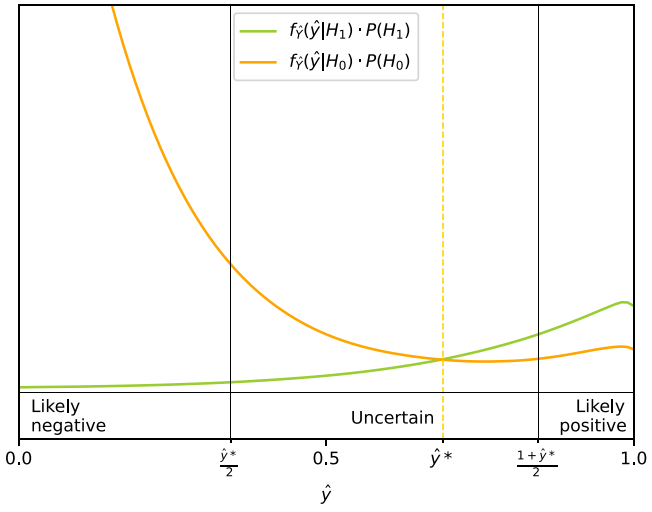



**Fig. 3.** Acquisition strategy — an illustration (for appliance kettle): Distributions of the model output under hypotheses $H_0$ and $H_1$, and three model output space regions.

When using the proposed acquisition function, as described in Section 3.1, there is a region in the model output space where model predictions are considered uncertain. During the active learning process, the uncertain region is quickly exhausted, but, as the model changes during the process, the model output for some samples can shift from likely positive or negative regions to uncertain. When samples from the uncertain region are exhausted, the process is meant to stop — it means that the uncertain samples have been already included in training and only samples for which the model has high level of certainty remain. To ensure that the model is consistently certain in its predictions, patience for a few epochs can be introduced - i.e., active learning can stop when the uncertain region is empty, or does not contain enough samples to fill $\boldsymbol{Q}$ for a few consecutive epochs, as per Eq. (5).

$$patience_i = \begin{cases} patience_{i-1} + 1, & \{s \in \boldsymbol{D}_{pool} | y = m_{i-1}(s) \in (\frac{\hat{y}^*}{2}, \frac{1+\hat{y}^*}{2})\} = \emptyset \\ 0, & \{s \in \boldsymbol{D}_{pool} | y = m_{i-1}(s) \in (\frac{\hat{y}^*}{2}, \frac{1+\hat{y}^*}{2})\} \neq \emptyset \end{cases}$$

$$S = \begin{cases} False, & patience_i < max\_patience \\ True, & patience_i \geq max\_patience \end{cases} \tag{5}$$

Patience in current iteration $i$ is denoted by $patience_i$, while $patience_{i-1}$ denotes the patience from the previous active learning iteration. $S$ is a boolean variable denoting if the stopping criterion has been met or not. This strategy offers timely stopping of the active learning process without the need to store and compare performance of the models from earlier stages of the process. In addition, this strategy eliminates the need for setting a predefined threshold on model performance, which can be a challenging task since it is not always straightforward to estimate the level of expected performance if the model is deployed in a new previously unseen environment.

### 3.3. Exploiting experts' confidence

To account for possible wrong labels introduced by humans during labelling, a method to incorporate their confidence about a label is introduced. Expert confidence levels are used to set weights inside the loss function during training – instead of treating all samples equally – by applying weighted average when calculating the loss as:

$$Loss = \frac{1}{N} \cdot \sum_{i=1}^{N} c_i \cdot Loss_i. \tag{6}$$

Here, $N$ denotes the total number of samples, and $Loss_i$ is the model's loss value for the $i$th sample. The higher the expert certainty, the higher the sample confidence weight $c_i$. A lower weight means that the effect of a sample to the calculated loss is attenuated, thus it contributes less to model learning.

### 3.4. Re-labelling samples

To reduce likelihood of training the model with wrong labels, a mechanism for returning samples with possibly erroneous labels, $\boldsymbol{R}$, for

re-labelling is implemented as:

$$\boldsymbol{R} = \{s \in \boldsymbol{Q} : MR(y, \hat{y}) < T_{\text{return}}\} \tag{7}$$

where

$$MR(y, \hat{y}) = \frac{\sum_{i=1}^{N} \min\{y_i, \hat{y}_i\}}{\sum_{i=1}^{N} \max\{y_i, \hat{y}_i\}}, \tag{8}$$

and $N$ is the signal window length.

Namely, after the loss function has been applied to each newly added sample $s_i \in \boldsymbol{Q}$, match rate (Eq. (8)) between the correct label $y_i$ of sample $s_i$ and soft model prediction $\hat{y}_i$ is calculated – if $MR$ is below a threshold $T_{\text{return}}$ even after the loss function is applied, it means that the sample possibly deviates from the rest of the training set, and that the label is possibly wrong; thus this sample is sent back for re-labelling, enabling the expert to re-consider and change their original decision.

## 4. Experimental setup

### 4.1. Data & DNN model

To facilitate reproducibility, we use the well documented public REFIT (Murray et al., 2017) and UK-DALE (Kelly and Knottenbelt, 2015) real-world electrical load measurements datasets as these two datasets are among the most widely used datasets for evaluation of NILM algorithms mimicking well real-world conditions (Angelis et al., 2022; Huber et al., 2021; Kaselimi et al., 2022). For example, both REFIT and UK-DALE datasets are used in Sykiotis et al. (2022) for complexity reduction and transferability via transformer-based architecture, in D'Incecco et al. (2020) for cross-domain and cross-appliance transfer, and in Murray et al. (2019) for evaluation of transferability of DNN architectures. REFIT consists of 2-year long (2013–2015) continuous time series electricity consumption recordings from 20 houses in the United Kingdom. Each house data contains aggregate electricity consumption time series measurements (see Fig. 5), as well as consumption of 9 individual appliances, measured at an 8-sec interval. The large number and diversity of appliance waveforms or signatures across 20 houses makes the REFIT dataset one of the most challenging NILM datasets and a good exemplar for robust evaluation of active learning methodologies.

To align with the widespread smart meter roll-out with in-house recording granularity of about 10 sec (Anon, 2013), the data is re-sampled to 10-sec sampling interval. Appliance types used in this study are kettle and microwave – resistive loads with short activation times – as well as washing machine and dishwasher – inductive (and also resistive) loads, with long cycle duration and multiple states. Measured aggregate electricity consumption expressed in Watts (W) is normalised using Z-normalisation technique: $Z = \frac{x-\mu}{\sigma}$, where $x$ denotes the original measurement, and $\mu$ and $\sigma$ stand for mean value and standard deviation of $x$ across the training dataset, respectively. To determine the ON–OFF state of appliances, thresholds are applied to measured electricity consumption of each appliance, according to Table 1.

In all experiments, as in Sykiotis et al. (2022), REFIT house 5, and UK-DALE house 1, which contain all four targeted appliances with many activations, are used for testing. A continuous period without missing data from 1st March 2014 to 1st September 2014 is chosen — first 2 months for the query pool and the rest for testing, to ensure that there is enough diversity among testing data, and that the query pool is of reasonable size since manual labelling is included in experiments. Continuous recordings from the query pool and testing data are sliced into non-overlapping windows before being fed to the model. As explained in Section 2.1, labels are not available for the query pool data, so, in the query pool, only aggregate electricity consumption measurements are used. Labels are provided later after the model makes a query, either by an oracle (Experiment 1), or by an expert (Experiment 2). For testing, submetering measurement labels are used to quantify model performance. Houses and time periods used for pre-training of each appliance are shown in Table 1 - for washing machine and dishwasher as in Sykiotis et al. (2022), and for microwave and kettle as in Murray et al. (2019). It is worth mentioning that in NILM, like in many other real-world applications based on time-series data where class-balance depends on the frequency of events, even though raw measurements are highly imbalanced (home appliances are turned off most of the time), it is possible to create balanced training datasets through continuous recording over long periods of time, without data augmentation.

**Table 1**
On-state power thresholds [W], REFIT houses and time periods used for training for each target appliance.

| Appliance | Training houses (REFIT) | "On-power" threshold [W] |
|---|---|---|
| Kettle | 6 (28.11.2013–28.06.2015.)<br>8 (01.11.2013–10.05.2015.)<br>17 (06.03.2014–19.06.2015.) | 2000 |
| Microwave | 6 (28.11.2013–28.06.2015.)<br>8 (01.11.2013–10.05.2015.)<br>17 (06.03.2014–19.06.2015.) | 200 |
| Washing machine | 2 (17.09.2013–28.05.2015.)<br>3 (25.09.2013–02.06.2015.)<br>16 (10.01.2014–08.07.2015.) | 20 |
| Dishwasher | 2 (17.09.2013–28.05.2015.)<br>3 (25.09.2013–02.06.2015.)<br>16 (10.01.2014–08.07.2015) | 10 |

The DNN model used in this paper is the ELECTRIcity transformer (Sykiotis et al., 2022), designed to work well with unbalanced data. The model architecture is presented in Fig. 4. It is trained in two phases: an unsupervised pre-training phase followed by a supervised training phase. The model shows superior performance to other state-of-the-art algorithms (Sykiotis et al., 2022). In experiments in this paper, for creating pre-trained models to be transferred to a new house, both training phases are used, but during the active learning process, only supervised fine-tuning phase is used. A *sigmoid* activation function has been added to the final layer of the network to perform on/off-state binary classification (instead of regression as in Sykiotis et al., 2022). One DNN model is created per monitored appliance — for example, if 4 different appliances are monitored in a house, then 4 different models will be created, for determining the state of each appliance separately. Therefore, each DNN model performs classification to 2 classes — on and off state. Since the model works in a sequence-to-sequence fashion, a pooling function is applied to the model output to get a single uncertainty value, by taking the maximum value of the model prediction window, with a reasoning that signal window is considered positive if there is at least one sample in that window where the appliance is active.

### 4.2. Evaluation metrics

The classification performance of the DNN-based NILM algorithm is evaluated using the standard $F_1$-score, which is calculated as the harmonic mean of precision and recall:

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} = \frac{TP}{TP + \frac{1}{2} \cdot (FP + FN)} \tag{9}$$

where TP denotes true positives – both model prediction and ground truth are positive; FP for false positives – prediction is positive but ground truth is negative; and FN for false negatives – prediction is negative while ground truth is positive.

AL performance is usually presented as a curve showing model accuracy against the number of labelling iterations, i.e., the number of samples queried and labelled. If a point with no labelling effort (i.e., iteration 0), and the maximum possible model performance (i.e., $F_1$-score equal to 1) is considered as an "ideal" point, as proposed in Todic

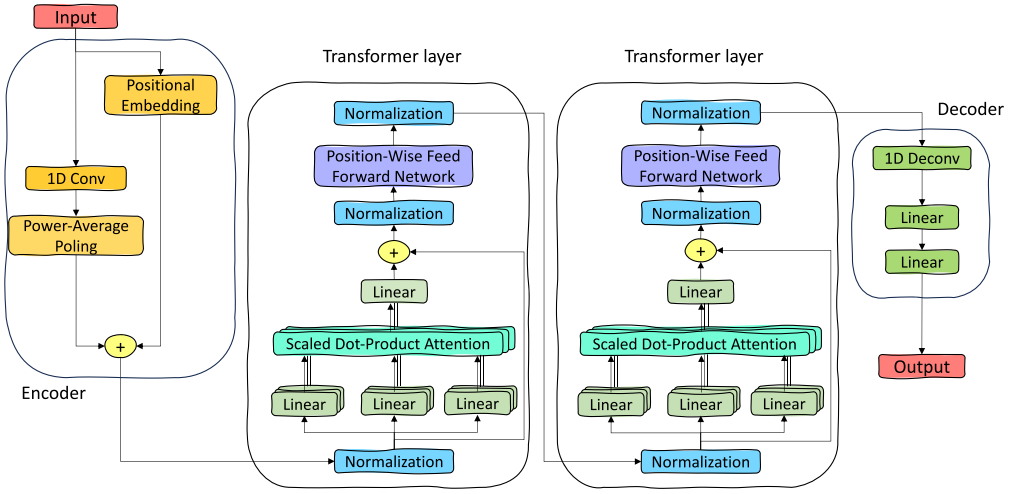



**Fig. 4.** Architecture of ELECTRIcity transformer model (Sykiotis et al., 2022).

et al. (2023), then the optimal point of the active learning process can be calculated as the point with minimum Euclidean distance from the ideal point:

$$dist = \sqrt{(1 - F_1)^2 + \left(\frac{|\boldsymbol{D}_{\text{ft}}|}{|\boldsymbol{D}_{\text{pool}}|}\right)^2} \tag{10}$$

### 4.3. Experiments

Two experimental settings were considered in this paper, as described next.

- Experiment 1: Transfer learning with labels obtained via submetering, with simulated labelling errors and re-labelling mechanism, and simulated confidence levels
  In this experiment, samples from the query pool are labelled using submetering electricity consumption measurements. The effect of balancing of queried batches using different acquisition functions is explored using several balanced acquisition functions. Stopping criterion is also applied to reduce labelling effort after the optimal point is achieved, as explained in Section 3.1. To study the effect of possible labelling errors and mimic a real-world active learning process when labels are provided by humans, different levels of false positive and false negative errors are simulated. Namely, if the model prediction for a sample in the query pool contains appliance activation, but the ground truth does not, false positive error is introduced to that sample by accepting model prediction as ground truth label, with a predefined probability. On the other hand, if model prediction for a sample does not contain appliance activation, but the ground truth does, false negative error (missing appliance activation; setting ground truth label to 0) is introduced with a predefined probability. The proposed re-labelling mechanism (Section 3.4) is then applied to detect possibly wrong labels and send them back for re-labelling. Also, simulated confidence levels in correlation with simulated errors were utilised throughout the process to attenuate negative effects of errors (Section 3.3).
- Experiment 2: Transfer learning with expert labelling, exploiting expert confidence levels
  In this experiment, the best setup obtained from the first experiment is verified in a real-world scenario, where experts provide labels during the active learning process. As those labels can be erroneous, expert's confidence level is considered during the training phase, assuming that if an expert is not confident about a label, the label is more likely to be wrong, and should be used with caution. A graphical user interface enabling experts to quickly provide labels together with their confidence was developed and used (see Section 4.4).

All DNN and active learning hyper-parameters are shown in the Table 2. Parameters for the DNN used are set as in Sykiotis et al. (2022). Although in Sykiotis et al. (2022), a window length of 480 samples is used for all appliances, here the window length is shortened for kettle and microwave to 120 samples instead of 480, because those appliances have very short activation times. Therefore query pool sizes differ for kettle and microwave (4416 samples) from those for washing machine and dishwasher (1104 samples), although the same time period of two months is used for the query pool. Learning rate and the number of epochs are different in the pre-training and fine-tuning phases — they are set lower in the fine-tuning phase within the active learning process to mitigate effects of overfitting due to a small number of labelled samples, especially in the beginning. At each labelling iteration, one batch of samples is queried. Confidence threshold for stream-based uncertainty acquisition function is set to be the same as in Todic et al. (2023). The number of uncertain samples coming from the uncertain region for the proposed acquisition function is set to 56 so that the majority of queried samples come from the uncertain region, and the rest - 8 samples per iteration from the likely positive and likely negative prediction regions — for the purpose of preserving diversity among queried data and preventing forgetting of the model. A PC with the following specifications is used in the experiments: Intel(R) Core(TM) i7-7800X CPU @ 3.50 GHz, 32 GB RAM, and a NVIDIA TITAN Xp GPU.

### 4.4. User interface

In order to facilitate experts' participation in the active learning process, a graphical user interface, shown in Fig. 5, is developed. Queried samples (windows of electric load measurements) from one labelling iteration are shown to the expert in a sequence, one by one. The aggregate signal in Watts is shown on the left vertical axis — this value can help experts decide if the appliance in question is on or off. Model prediction is shown together with aggregate signal (the values of the prediction can be seen from the right vertical axis, in range 0–1), to inform experts of model's behaviour and possibly help them make a decision. Horizontal axis shows time, which also can help an expert make a decision - e.g., some appliances are more likely to be operated during a particular time of a day. Experts are asked to mark the part of the window where they think the appliance of interest is active, by simply drawing a rectangle over that area, as shown in Fig. 5. Apart from labels, experts are asked to provide their confidence level associated with each label - i.e., they are asked to select one of three offered options — low confidence, medium confidence or high

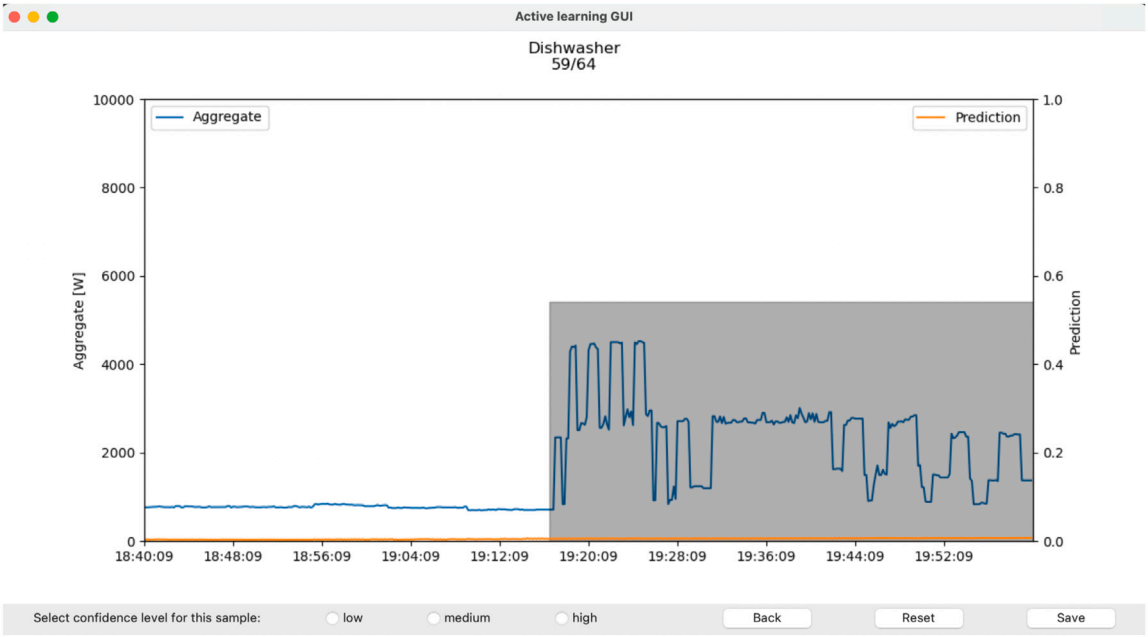



**Fig. 5.** User interface that facilitates quick labelling by experts participating in the active learning process.

**Table 2**
Hyper-parameters used in the experiments.

| | |
|---|---|
| DNN model | |
| Input window size | kettle, microwave: 120<br>washing m., dishwasher: 480 |
| Heads, hidden, layers | 2, 256, 2 |
| Dropout rate | 0.2 |
| tau | 0.1 |
| Learning rate | pre-training: 1e−3<br>fine-tuning: 1e−4 |
| Epochs | pre-training: 100<br>fine-tuning: 10 |
| Batch size | 64 |
| Model threshold | 0.3 |
| Active learning | |
| Queries per iteration | 64 |
| Query pool size | 2 month worth of samples:<br>kettle, microwave: 4416<br>washing m., dishwasher: 1104 |
| Confidence threshold<br>(stream-based unc. acq. function; Exp.1) | 0.9 |
| # of samples<br>for the proposed<br>acquisition function | 4 likely neg.<br>56 uncertain<br>4 likely pos. |

confidence. High confidence is then mapped in the back end to a coefficient $k = 3$, mid confidence to $k = 2$, and low confidence to $k = 1$, which are then converted into sample weight according to:

$$c_i = \frac{N}{\sum_{j=1}^{N} k_j} \cdot k_i, \tag{11}$$

calculated at a batch level. This way, the samples with higher confidence have triple the weight of samples with lower confidence, and samples with mid confidence double, but the sum of weights in a batch remains the same as before the weights were adjusted. Obtained weights are then included in the loss function (see Eq. (6)), as described in Section 3.3.

## 5. Results & discussion

In this section we report our experimental results. The goal of the experiments is to: (1) evaluate performance of the proposed acquisition function against state-of-the-art benchmarks without labelling errors; (2) test effectiveness of the proposed stopping criteria; (3) test if the proposed re-labelling leads to performance gains, and (4) show usefulness of the introduced expert confidence scores.

We organise the section into two parts: first we report the results related to Experiment 1 as described in the previous section; then, we evaluate the proposed system with three NILM experts using the designed user interface.

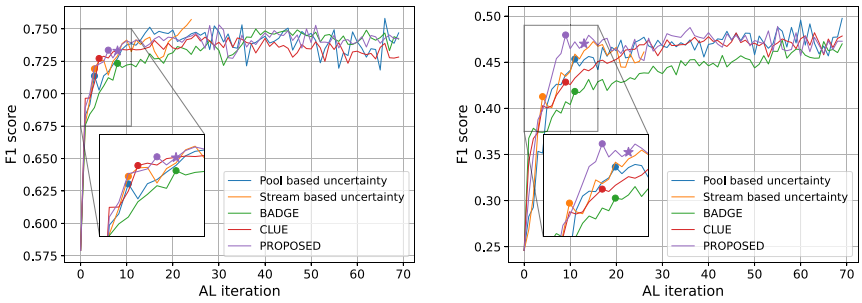
### 5.1. Experiment 1

#### 5.1.1. Acquisition function

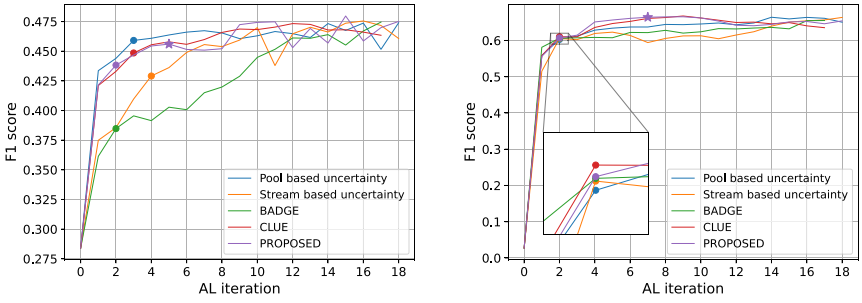
In this subsection we compare the performance of the proposed acquisition function against state-of-the-art benchmarks. Acquisition functions used for benchmarking are pool- and stream-based uncertainty acquisition functions, as they are lightweight algorithms and demonstrate good performance for the NILM problem (Todic et al., 2023).

For the stream-based uncertainty acquisition function an informativeness threshold is used to make a decision if samples are sent for labelling or not (see Section 2.1.1). Since in Todic et al. (2023), it was demonstrated that low values of informativeness threshold provide higher improvement in the beginning of the active learning processes, the starting threshold is set to 0.9, and then as the process progresses, it is increased if the number of selected samples is lower than the batch size. This way the active learning process experiences both high performance improvement in the beginning and longer lasting process which includes samples with a higher confidence at later stages.

Two additional benchmarks are used that attempt to diversify samples and balance the classes: BADGE acquisition function (Ash et al., 2019) that diversifies queried samples to avoid redundancy by looking at gradient embeddings, and CLUE acquisition function (Prabhu et al., 2021), that diversifies queried samples by looking at penultimate layer activations, but also includes least confidence uncertainty, i.e., it takes advantage of both uncertainty and diversification of queried batch of samples.

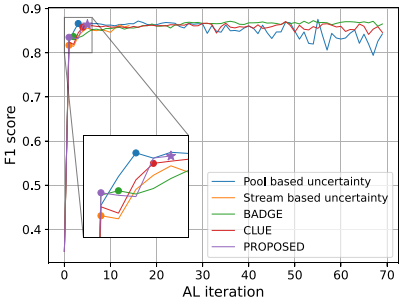



(a) Kettle (b) Microwave

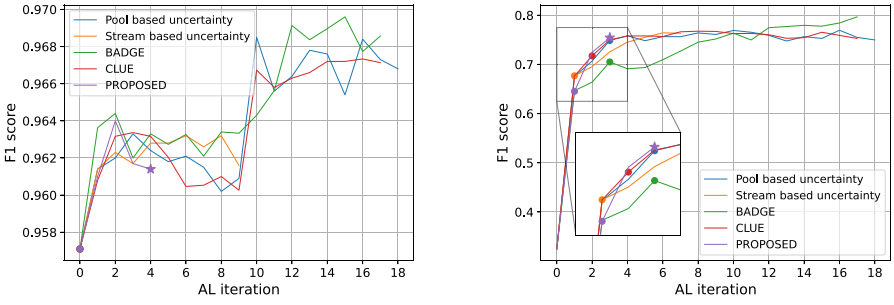



(c) Washing Machine (d) Dishwasher

**Fig. 6.** Comparison between different acquisition functions — transfer to REFIT house 5: the proposed one based on the optimal thresholding strategy; pool-based uncertainty (as in Todic et al., 2023); stream-based uncertainty (Todic et al., 2023); BADGE (Ash et al., 2019); CLUE (Prabhu et al., 2021). Dots denote the optimal points and stars the stopping point for the proposed strategy.




(a) Kettle




(b) Washing Machine (c) Dishwasher

**Fig. 7.** Comparison between different acquisition functions — pre-training on the REFIT dataset and transfer to UK-DALE house 1: The proposed acquisition function based on the optimal thresholding strategy; pool-based uncertainty (as in Todic et al., 2023); stream-based uncertainty (Todic et al., 2023); BADGE (Ash et al., 2019); CLUE (Prabhu et al., 2021). Dots denote the optimal points and stars the stopping point for the proposed strategy.

Results of the comparison for the four appliances from REFIT house 5 are shown in Fig. 6, and from UK-DALE house 1 in Fig. 7. Horizontal axis shows the AL, i.e., labelling, iteration, and vertical axis the achieved $F_1$-score. Optimal points calculated based on Eq. (10) are marked as dots, and stopping points for the proposed acquisition function as proposed in Section 3.2 are marked with stars. Results for BADGE (Ash et al., 2019) and CLUE (Prabhu et al., 2021) acquisition functions are averaged over 3 independent runs, because those algorithms depend on cluster initialisation.

Numerical results of this experiment for REFIT house 5 - achieved $F_1$-scores and percentage of query pool samples queried at optimal point, maximum performance point and stopping point (for the proposed acquisition function) - are presented in Table 3, and for UK-DALE house 1 in Table 4.

As shown in Figs. 6 and 7 (and in accordance with findings of Todic et al., 2023), pool- and stream-based acquisition functions both demonstrate high and stable performance. Batch-aware acquisition function BADGE (Ash et al., 2019) performs slightly worse than pool- and stream-based uncertainty (except for dishwasher in Fig. 6(d), and washing machine in Fig. 7(b)), which indicates that the dataset does not benefit from batch balancing during acquisition, and that some types of samples (windows containing activations in this case) are more significant for model improvement. Although CLUE (Prabhu et al., 2021) diversifies queried samples as well, it exploits model uncertainty, so its performance is on par with pool-based acquisition function.

It is observed that with pool- and stream-based uncertainty acquisition functions, in the beginning of the process, mostly samples containing appliance activations are being queried and added to the training set, due to high uncertainty associated with them. That usually results in a large jump in performance. After all samples containing activation have been exhausted, samples without activation, but with high aggregate consumption, are being queried, and finally, samples without appliance activation and with low aggregate values are being queried.

Our proposed acquisition function favours low- and mid-certainty signal windows containing appliance activation, but also chooses samples without appliance activation, as well as high-certainty samples containing activation. That way, it keeps diversity among queried data, but also ensures that sufficient number of samples important for learning of new patterns are regularly selected. This strategy performs the best for kettle and microwave in REFIT house 5 (Figs. 6(a) and 6(b)), since these two appliances are often confused with washing machine, since they have similar wattage. Moreover, those two appliances have very short duration times, hence a small number of samples within a window contain an activation. Thus, it is important to choose enough samples where the model predicts there is an activation, but also high-certainty samples help in preventing forgetting of patterns of interest, and correcting wrong behaviour caused by confusions with other appliances as described above.

For washing machine and microwave in REFIT house 5, with multi-state relatively more complex signatures, high-certainty samples are usually correctly predicted and the model benefits mostly from low-certainty samples. Therefore, the pool-based acquisition function performs well. Transferability of the washing machine and microwave models, compared to more distinct kettle signature, is relatively poor in general (D'Incecco et al., 2020; Li et al., 2023) and often excluded in the NILM literature.

However, for washing machine in UK-DALE house 1, initial performance is very good, due to much lower background noise levels in this house. The active learning curve, hence, does not have the usual shape, but its range covers only the $F_1$-score from 0.96 to 0.97 — there is not much room for improvement if starting performance is so good, as opposed to other appliances from this house and from REFIT house 5.

For the dishwasher, the starting performance is poor in REFIT house 5 and UK-DALE house 1, indicating that the dishwasher model in the test houses is very different from those present in the pre-training dataset, but only two (Fig. 6(d)) and three (Fig. 7(c)) active learning labelling iterations are sufficient to significantly improve the performance. All query strategies perform equally well in REFIT house 5 - due to a very low starting performance, all acquisition functions provide a highly informative fine-tuning set that contributes to significant model improvement. Nevertheless, it can be seen that the proposed strategy (purple star in Figs. 6(d) and 7(c)) led to the highest performance in both test houses.

Based on the proposed stopping criteria, stopping is applied after 3 consecutive iterations with less than a half of the required high-uncertainty samples present in the query pool, to ensure consistent certainty of the model. Stopping points are therefore always located several iterations after the optimal points. It can be seen from Figs. 6 and 7, as well as from the numerical results presented in Tables 3 and 4, that the proposed early stopping significantly saves the labelling effort with negligible performance loss. Indeed, the gap between the point where the maximum performance is achieved and the stopping point is always very small.

**Table 3**
Comparison between five acquisition functions for 4 appliances from REFIT house 5: kettle, microwave, washing machine and dishwasher. The optimal points (Opt.), stopping points (Stop) and maximum performance (Max) are all included. Note that Maximum point is a point where the curves reach their maximum, which is unknown in practice and cannot be used to stop. $\frac{|D_{\text{fi}}|}{|D_{\text{pool}}|}$ is the percentage of samples being labelled.

| Acquisition function | | Kettle | | Microwave | | Washing M. | | Dishwasher | |
|---|---|---|---|---|---|---|---|---|---|
| | | $F_1$ | $\frac{\lvert D_{\text{fi}}\rvert}{\lvert D_{\text{pool}}\rvert}$ | $F_1$ | $\frac{\lvert D_{\text{fi}}\rvert}{\lvert D_{\text{pool}}\rvert}$ | $F_1$ | $\frac{\lvert D_{\text{fi}}\rvert}{\lvert D_{\text{pool}}\rvert}$ | $F_1$ | $\frac{\lvert D_{\text{fi}}\rvert}{\lvert D_{\text{pool}}\rvert}$ |
| Pool based unc. | Opt. | 0.71 | 4% | 0.45 | 16% | 0.46 | 17% | 0.60 | 11% |
| | Max | 0.76 | 96% | 0.49 | 100% | 0.47 | 100% | 0.66 | 78% |
| Stream based unc. | Opt. | 0.72 | 4% | 0.41 | 6% | 0.43 | 22% | 0.60 | 12% |
| | Max | 0.76 | 35% | 0.47 | 22% | 0.48 | 89% | 0.66 | 100% |
| BADGE (Ash et al., 2019) | Opt. | 0.72 | 12% | 0.42 | 16% | 0.38 | 12% | 0.60 | 12% |
| | Max | 0.75 | 55% | 0.47 | 81% | 0.47 | 100% | 0.66 | 100% |
| CLUE (Prabhu et al., 2021) | Opt. | 0.73 | 6% | 0.43 | 13% | 0.45 | 18% | 0.61 | 12% |
| | Max | 0.75 | 36% | 0.48 | 91% | 0.47 | 71% | 0.67 | 53% |
| PROPOSED | Opt. | 0.73 | 9% | 0.48 | 13% | 0.44 | 11% | 0.61 | 11% |
| | Stop | 0.73 | 12% | 0.47 | 19% | 0.46 | 28% | 0.66 | 39% |
| | Max | 0.75 | 43% | 0.49 | 80% | 0.48 | 83% | 0.66 | 50% |

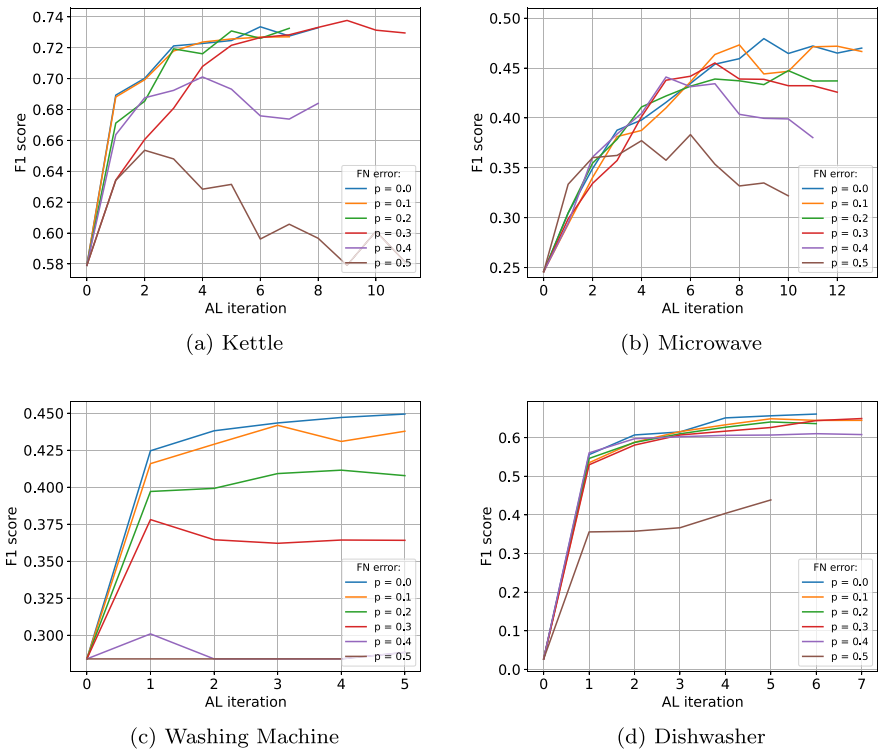
**Table 4**
Comparison between five acquisition functions for 3 appliances from UK-DALE house 1: kettle, washing machine and dishwasher. The optimal points (Opt.), stopping points (Stop) and maximum performance (Max) are all included. Note that Maximum point is a point where the curves reach their maximum, which is unknown in practice and cannot be used to stop. $\frac{|D_{\text{fi}}|}{|D_{\text{pool}}|}$ is the percentage of samples being labelled.

| Acquisition function | | Kettle | | Washing M. | | Dishwasher | |
|---|---|---|---|---|---|---|---|
| | | $F_1$ | $\frac{\lvert D_{\text{fi}}\rvert}{\lvert D_{\text{pool}}\rvert}$ | $F_1$ | $\frac{\lvert D_{\text{fi}}\rvert}{\lvert D_{\text{pool}}\rvert}$ | $F_1$ | $\frac{\lvert D_{\text{fi}}\rvert}{\lvert D_{\text{pool}}\rvert}$ |
| Pool based unc. | Opt. | 0.87 | 4% | 0.96 | 0% | 0.75 | 17% |
| | Max | 0.88 | 80% | 0.97 | 56% | 0.77 | 89% |
| Stream based unc. | Opt. | 0.81 | 1% | 0.96 | 0% | 0.68 | 6% |
| | Max | 0.86 | 17% | 0.96 | 33% | 0.76 | 33% |
| BADGE (Ash et al., 2019) | Opt. | 0.84 | 3% | 0.96 | 0% | 0.72 | 18% |
| | Max | 0.87 | 78% | 0.97 | 88% | 0.80 | 100% |
| CLUE (Prabhu et al., 2021) | Opt. | 0.86 | 6% | 0.96 | 0% | 0.72 | 12% |
| | Max | 0.87 | 70% | 0.97 | 94% | 0.77 | 47% |
| PROPOSED | Opt. | 0.83 | 1% | 0.96 | 0% | 0.65 | 6% |
| | Stop | 0.86 | 7% | 0.96 | 22% | 0.75 | 17% |
| | Max | 0.86 | 7% | 0.96 | 11% | 0.75 | 17% |

### 5.1.2. The impact of errors and re-labelling mechanism

Next, we evaluate the performance when labelling errors are present in REFIT house 5 and assess usefulness of the proposed re-labelling strategy with the proposed acquisition function and the proposed stopping criteria.

Fig. 8 shows the results when false negative errors are introduced into labels, i.e., positive labels are set as negative. Blue line corresponds




(a) Kettle (b) Microwave

(c) Washing Machine (d) Dishwasher

**Fig. 8.** Active learning with simulated false negative errors in the labels for kettle (a), microwave (b), washing machine (c) and dishwasher (d) from REFIT house 5.

to correct labels, without any errors introduced. Note that the number of iterations differ across the appliances due to the proposed stopping criteria. As expected, as error probability $p$ increases, the performance decreases — lower $F_1$-score is achieved. Kettle is sensitive to high levels of error, especially at later stages — it has a signature of short duration that is easily forgotten by the model if the error rate is high. Lower error rates do not impact the performance significantly. Microwave and washing machine are sensitive to this type of labelling errors even with lower error probabilities, which is reasonable since they have signatures that are already challenging to disaggregate even without any errors in labels.

Fig. 9 shows the results when false positive errors are introduced into labels, i.e., negative labels are set as positive. Since samples with appliance activations are more likely to be queried first as described above, the impact of false positive errors is expected to be less pronounced than the impact of false negative errors, at least in the beginning, which can be confirmed in Fig. 9. Namely, since the dataset is already highly imbalanced in favour of sample windows without appliance activation, with false negative errors, we introduce even more negative samples, and the model starts to 'forget' the pattern it learnt to recognise. On the other hand, false positive errors are likely to be introduced for samples where the aggregate signal looks as if there is appliance activation, so the model retains the ability to recognise important patterns.

Fig. 10 demonstrates the usefulness of the proposed re-labelling mechanism. Performance is compared between the case with and without re-labelling, with false negative errors occurring with the probability of 0.3. Match rate threshold $T_{\text{return}}$ in Eq. (7) is heuristically set to $1e-4$ for kettle and microwave, and $5e-5$ for washing machine and dishwasher, since these appliances have longer lasting cycles and the match rate is expected to be lower even for the good predictions. The assumption is that once a sample is returned for re-labelling, a correct label is provided. Improvement in performance when using the re-labelling mechanism is observed for all four appliances, and it is most pronounced for washing machine, which is very sensitive to this type of error (see Fig. 8(c)). This means that the mechanism successfully captures the samples which were wrongly labelled, and enables correcting labels by taking another look at them.

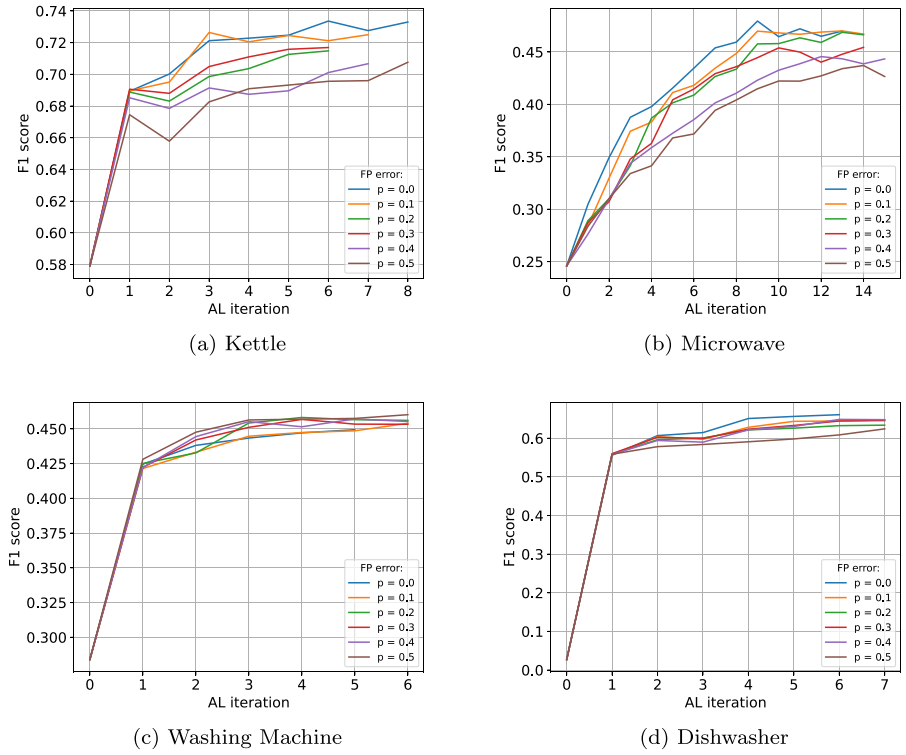
Results show that more samples are returned in active learning iterations where a drop in performance is observed (e.g. iterations 6 and 7 for kettle, iterations 6–9 for microwave), indicating that the model started to adopt wrong labels, but still has not forgotten the pattern of interest, and still can detect suspicious labels. Due to the complex pattern of washing machine, the model is less confident in its predictions, and relies more and adapts to provided labels, making the predictions similar to labels, even if those are wrong. However, samples re-labelled in the beginning do improve the performance, and the improvement achieved in the beginning does not decline in later stages.

#### 5.1.3. Exploiting confidence during training

Fig. 11 shows the usefulness of the proposed modification of loss function (Eq. (6)) to take into account confidence levels related to labels. False negative errors with probability of 0.5 are simulated. Based on the assumption that confidence level is correlated with the quality of label, two confidence levels are assumed — high confidence for samples without labelling errors and low confidence for samples containing a labelling error. The improvement in performance when using confidence levels during training compared to not using them is observable for kettle, washing machine and dishwasher from the very beginning. Even though proposed strategy improves performance for microwave and washing machine, clear convergence is not reached as with kettle and dishwasher. This is due to the complex, multi-state signatures of microwave and washing machine, as opposed to distinct patterns of kettle and dishwasher.

### 5.2. Experiment 2

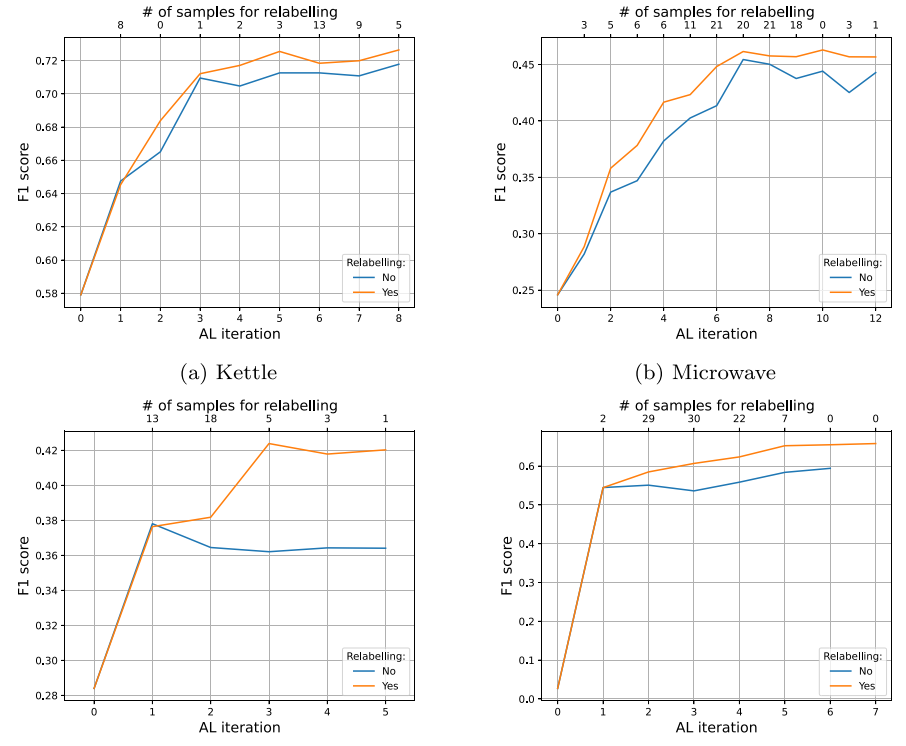
In this subsection we report the results when three experts are asked to label the samples using the user interface presented in Fig. 5. Each expert was asked to label one or more appliances. We used the proposed acquisition function, the stopping criteria and re-labelling mechanism.




(a) Kettle (b) Microwave

(c) Washing Machine (d) Dishwasher

**Fig. 9.** Active learning with simulated false positive errors in the labels for kettle (a), microwave (b), washing machine (c) and dishwasher (d) from REFIT house 5.

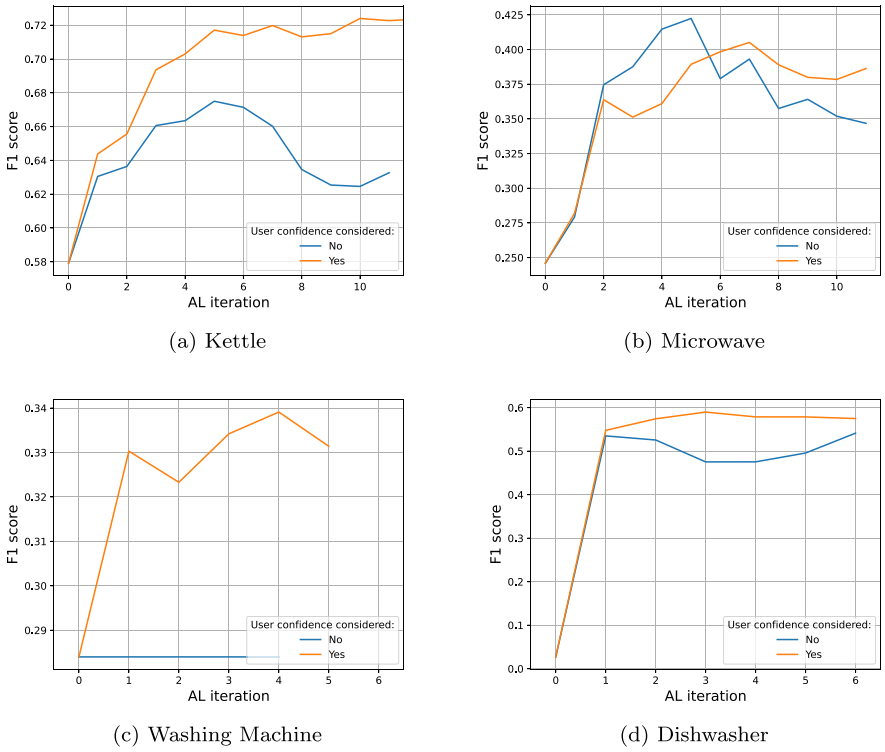



(a) Kettle (b) Microwave

(c) Washing Machine (d) Dishwasher

**Fig. 10.** The proposed active learning method with and without the re-labelling mechanism for kettle (a), microwave (b), washing machine (c) and dishwasher (d) from REFIT house 5.

Fig. 13 shows the results with and without using expert confidence levels. Horizontal axis represents the number of active learning labelling iterations, and vertical $F_1$-score achieved. The blue line corresponds to the case when labels are provided by an expert familiar with NILM, but without his/her confidence levels related to each label taken into account during training (i.e., all confidence levels are set




(a) Kettle (b) Microwave

(c) Washing Machine (d) Dishwasher

**Fig. 11.** Active learning with and without confidence taken into account during training for kettle (a), microwave (b), washing machine (c) and dishwasher (d) from REFIT house 5.

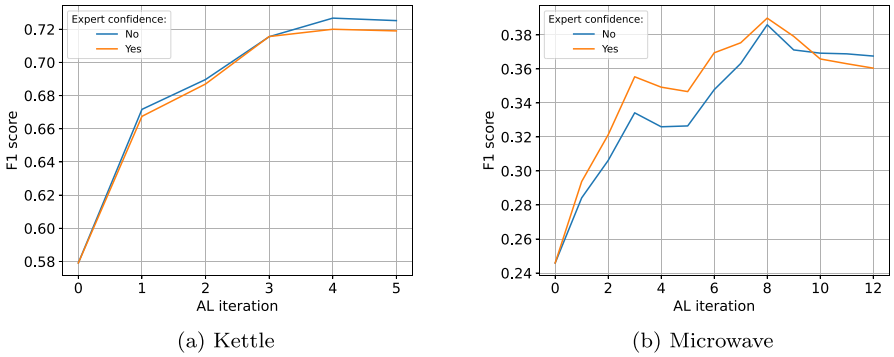
to 'high'); and the orange line corresponds to the case when labels are provided by an expert, and their confidence levels are included into the loss function (Eq. (6)) during training.

Examples of signal windows from REFIT house 5 labelled for washing machine by expert #3 and tagged with low and high confidence levels are shown in Fig. 14, showing that more noisy samples, with not so distinct signatures, are more challenging to be labelled by naked eye.

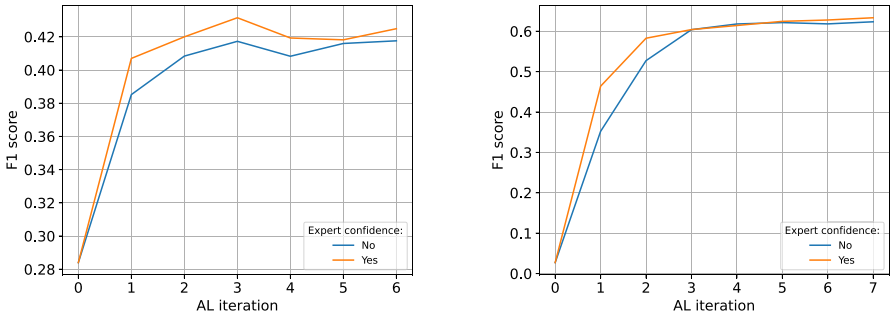
The quality of expert-provided labels, in terms of hit, miss and false alarm, compared to the submetering ground truth is shown in Tables 5 and 6. Hit is defined as the case when the expert-provided label is overlapping with the submetering label (equivalent to TP); Miss as the case when the submetering label has an activation, but the expert-provided label does not (equivalent to FN); and False alarm as the case when the submetering label does not have an activation, but the expert-provided label does (equivalent to FP). In cases when there is an activation both in submetering and expert-provided label, but they do not overlap, the label falls under the Miss & False alarm category. A histogram of expert confidence levels is given next to the number of labels belonging to each of the four categories, where red denotes low confidence, yellow middle, and green high confidence levels.

For kettle from REFIT house 5 in Fig. 12(a), using confidence levels did not improve the results — the labels are already of high quality, the number of misses and false alarms is very low compared to the number of hits, which is expected since the kettle has a single state, easily recognisable signature. Moreover, the expert assigned to most of the labels high confidence, as in the no-confidence level benchmark. However, a couple of mistakes have high confidence levels, which probably caused the confidence level curve to be slightly worse than no confidence level in Fig. 12(a). The same situation is observed in UK-DALE house 1 in Fig. 13(a). For microwave (Fig. 12(b)), which is a challenging appliance to label since activations are sparse and fluctuating, the power/watt level is lower compared to kettle, and there are different modes of running the appliance, expert-provided labels contain a significant number of mistakes. However, those mistakes are tagged with low confidence levels, so utilising user confidence levels did improve the results compared to the benchmark. For washing machine REFIT house 5, Fig. 12(c), which was labelled by another expert, there is a larger percentage of labelling mistakes, some of which have high confidence levels. However, there are low and mid-confidence levels among wrongly labelled samples, which was enough to lead to performance improvement compared to no confidence level case. For washing machine in UK-DALE house 1, Fig. 13(b), a vast majority of samples are correctly labelled, and tagged with high confidence. This causes weights to be very similar as in the case when confidence levels are not accounted for. Even though in Fig. 13(b) it looks like there is a significant gap between the two curves, note that the difference is at most 0.001 in $F_1$-score, so performance is practically the same. For dishwasher from REFIT house 5, Fig. 12(d), labelled by another expert, the provided labels are of higher quality since they have a more distinct signature than microwave. In addition, the correct labels mostly have high confidence values, which increased the contribution of confidence exploitation, and led to minor differences between the two curves. In UK-DALE house 1, Fig. 13(c), there are very few activations among queried samples before the active learning process stopped, and therefore there is almost no difference between the two curves — there are many correctly labelled negative examples tagged with high confidence levels.

The main challenges encountered in this experiment are the cases when an expert assigns the same confidence value to almost all samples — then the proposed weighing of samples based on expert's confidence approaches the case when no confidence is accounted for (the vast majority of samples get the same weight). This is not the problem in datasets with low noise levels, when labels are of very high quality (for example, washing machine in UK-DALE house 1, see Table 6) - the most of high-confidence samples are correctly labelled; but this is a problem in very noisy datasets where there are both correct and wrong labels, but the expert is either over-confident (many wrong labels tagged by high confidence) or under-confident (many correct labels tagged with low confidence). Therefore, skill level of experts poses a limitation to this approach to some extent.

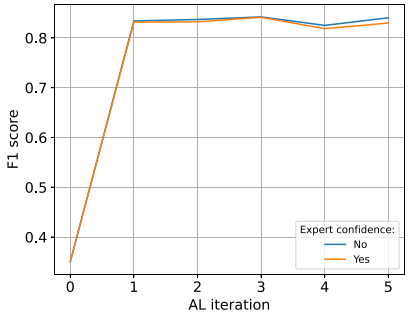



(a) Kettle (b) Microwave

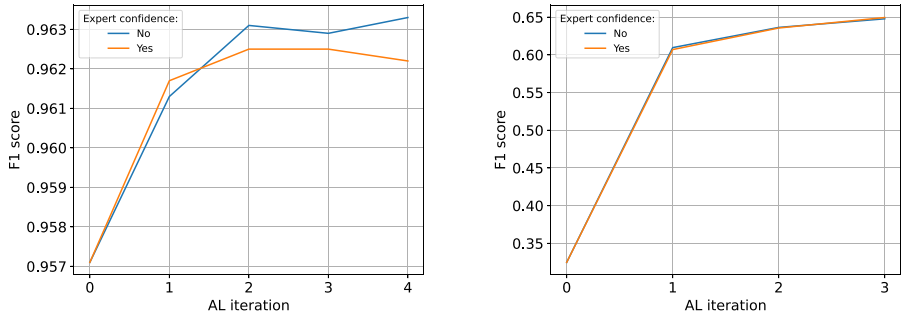



(c) Washing Machine (d) Dishwasher

**Fig. 12.** Experiment 2, REFIT house 5: Three experts asked to provide labels, where each expert labels one or two appliances. The performance curves are shown with and without expert confidence taken into account.

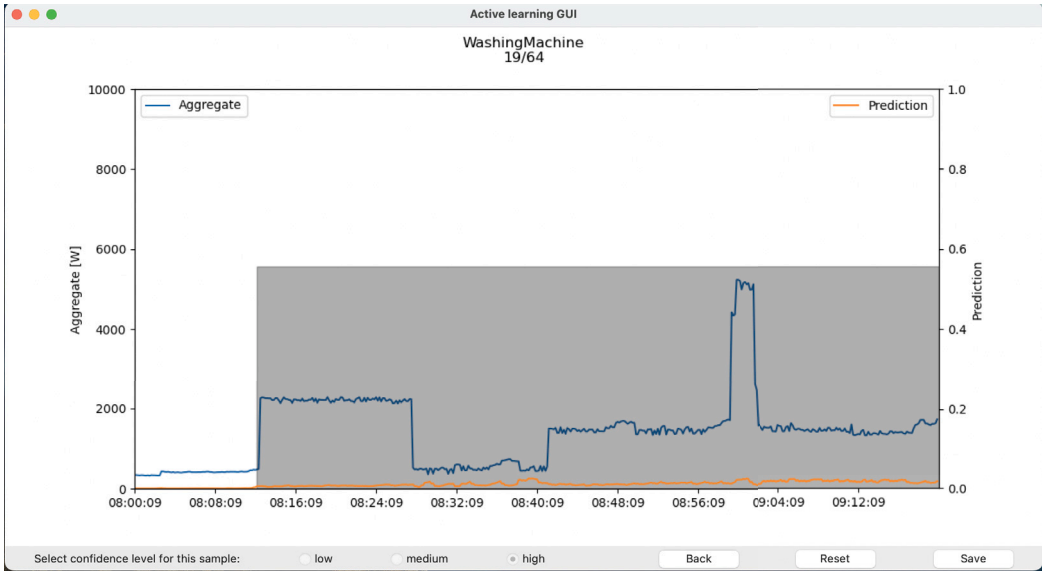



(a) Kettle

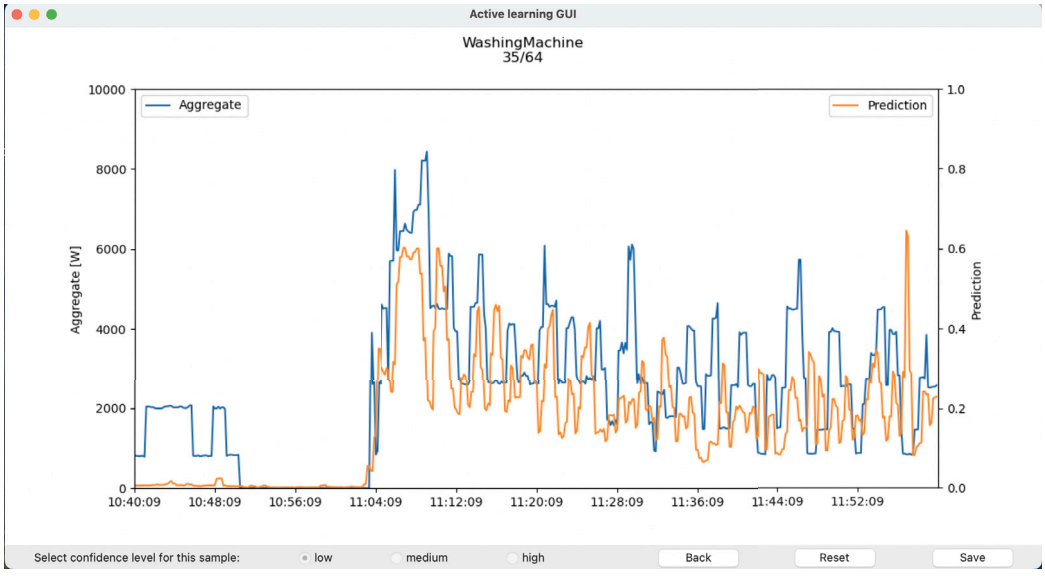



(b) Washing Machine (c) Dishwasher

**Fig. 13.** Experiment 2, UK-DALE house 1: Experts asked to provide labels, where each expert labels one or two appliances. The performance curves are shown with and without expert confidence taken into account.




(a) High confidence - Hit




(b) Low confidence - Miss

**Fig. 14.** Experiment 2: User interface showing examples of signal windows from REFIT house 5 with washing machine tagged with low and high confidence levels by expert #3.

**Table 5**
Quality of expert-provided labels compared to ground truth for REFIT house 5. Red denotes low confidence, yellow middle, and green colour high confidence levels.

| | Kettle | Microwave | Washing M. | Dishwasher |
|---|---|---|---|---|
| Expert | #1 | #3 | #3 | #2 |
| Hit | 113 | 26 | 46 | 87 |
| Miss | 36 | 28 | 25 | 26 |
| False alarm | 32 | 30 | 7 | 6 |
| Miss & False alarm | 5 | 2 | 0 | 0 |
| Total # of labels | 320 | 768 | 384 | 448 |

**Table 6**
Quality of expert-provided labels compared to ground truth for UK-DALE house 1. Red denotes low confidence, yellow middle, and green colour high confidence levels.

| | Kettle | Washing M. | Dishwasher |
|---|---|---|---|
| Expert | #1 | #3 | #3 |
| Hit | 99 | 78 | 17 |
| Miss | 12 | 4 | 6 |
| False alarm | 7 | 1 | 0 |
| Miss & False alarm | 3 | 0 | 0 |
| Total # of labels | 320 | 256 | 192 |

## 6. Conclusions

This paper proposes a human-in-the-loop active learning methodology for time series data, demonstrated and evaluated for the non-intrusive load monitoring problem. Novel contributions to enable the proposed overall active learning methodology comprise: design of an acquisition function based on maximum a posteriori hypothesis testing, accounting for both model uncertainty and balancing classes; a stopping criterion once optimal performance is achieved, to minimise resource-intensive labelling effort; mitigating the effect of wrong labels possibly provided by users throughout the process via two mechanisms by returning possibly wrongly labelled samples for re-labelling, and accounting for user's certainty level about provided labels, respectively.

Two experiments are conducted, applying novel AL-based approaches to the problem of time series classification of individual loads in aggregate smart meter measurements, leveraging on publicly available REFIT (Murray et al., 2017) and UK-DALE (Kelly and Knottenbelt, 2015) datasets, and transformer-based deep learning ELECTRIcity model (Sykiotis et al., 2022). The first set of experiments show that the proposed acquisition function achieves similar performance to state-of-the-art methods, but with smaller number of samples labelled due to balancing better classes and cleverly stopping when good performance is reached. Labelling effort is reduced by between 61% (in the case of dishwasher) and 88% (in the case of kettle) in REFIT house 5, and between 78% (in the case of washing machine) and 93% (in the case of kettle) in UK-DALE house 1. Furthermore, even with errors introduced throughout the labelling process, the proposed active learning method enhances the model to be generalised for various profiles for the same label. The proposed re-labelling mechanism is shown to be effective in detection of mistakes during the labelling process, and offers the possibility to improve the performance by providing new labels for uncertain data samples. Finally, including confidence levels of human experts, especially in cases where samples are noisy, is beneficial as it prevents a drop in performance caused by accumulation of wrong labels. The second experiment verifies the use of proposed active learning approaches in real-world scenarios, where despite unintentionally introduced errors, model performance is still boosted, especially with the use of the proposed methods for error effect mitigation.

The proposed active learning approach demonstrated improved performance when pre-trained NILM models are transferred to new, unseen homes. Even when the initial performance prior to active learning is poor, the proposed approach can largely improve performance by labelling a considerably small amount of data. The method can scale to many houses (hundreds, thousands) - algorithms are adjusted to each house separately — no data needs to be exported, and users (house owners) can help label their own data based on time when specific appliances are used, until the algorithms become well tuned and high performing. Considering recordings from a long period of time ensures heterogeneity of data and stability of the model. Even if circumstances in their house change (e.g., an appliance is replaced or a new high-consuming load is introduced), which impact the aggregate measurements and hence the NILM algorithm performance, the active learning process can adjust the model, ensuring performance stability.

The proposed approach is demonstrated to be applicable to sensor measurements where the data being measured is fluctuating, varies across houses (domains), is noisy, and labelling is challenging. Furthermore, the very challenging nature of the load disaggregation problem is akin to the broader single source separation problem arising often from environmental sensing and therefore the method's efficacy in NILM stretches to other application domains based on solving single source separation problem from noisy time-series reading.

As some types of labels are very hard to be provided by users (for example, regression labels for the problem of load disaggregation, or strong labels for time-series windows in general), it would be worth exploring the use of Siamese networks in future work, that could be pre-trained for both regression and classification tasks at the same time, or with both strong and weak labels at the same time. Furthermore, user-provided confidence levels could be used to further train the model to learn its own confidence level. Moreover, along with model prediction, some explanation tools could be used to inform the expert of the reasoning behind the prediction to help labelling.

## CRediT authorship contribution statement

**Tamara Sobot:** Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. **Vladimir Stankovic:** Conceptualization, Funding acquisition, Project administration, Supervision, Validation, Writing – review & editing. **Lina Stankovic:** Writing – review & editing, Supervision, Funding acquisition, Project administration.

## Declaration of competing interest

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.

## Data availability

An open-access, publicly available dataset was used in this study. The dataset can be accessed at https://pureportal.strath.ac.uk/en/datasets/refit-electrical-load-measurements-cleaned, hosted at University of Strathclyde (Murray et al., 2017).

## Acknowledgements

We would like to thank Apostolos Vavouris, Djordje Batic and Stavros Sykiotis for supporting this research by participating in the second experiment.

## Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955422.

## References

Angelis, G.F., Timplalexis, C., Krinidis, S., Ioannidis, D., Tzovaras, D., 2022. NILM applications: Literature review of learning approaches, recent developments and challenges. Energy Build. 261, 111951.

Anon, 2013. Smart Metering Equipment Technical Specifications Version 2. https://www.gov.uk/government/publications/smart-metering-implementation-programme-information-leaflet (Accessed 19 June 2023).

Anon, 2022. Delivering the European green deal. https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal/delivering-european-green-deal_en (Accessed 19 June 2023).

Ash, J.T., Zhang, C., Krishnamurthy, A., Langford, J., Agarwal, A., 2019. Deep batch active learning by diverse, uncertain gradient lower bounds. arXiv preprint arXiv:1906.03671.

Bloodgood, M., Vijay-Shanker, K., 2014. A method for stopping active learning based on stabilizing predictions and the need for user-adjustable stopping. arXiv preprint arXiv:1409.5165.

Budd, S., Robinson, E.C., Kainz, B., 2021. A survey on active learning and human-in-the-loop deep learning for medical image analysis. Med. Image Anal. 71, 102062. http://dx.doi.org/10.1016/j.media.2021.102062, URL https://www.sciencedirect.com/science/article/pii/S1361841521001080.

Cheplygina, V., Perez-Rovira, A., Kuo, W., Tiddens, H.A.W.M., de Bruijne, M., 2016. Early experiences with crowdsourcing airway annotations in chest CT. In: Carneiro, G., Mateus, D., Peter, L., Bradley, A., Tavares, J.M.R.S., Belagiannis, V., Papa, J.P., Nascimento, J.C., Loog, M., Lu, Z., Cardoso, J.S., Cornebise, J. (Eds.), Deep Learning and Data Labeling for Medical Applications. Springer International Publishing, Cham, pp. 209–218.

D'Incecco, M., Squartini, S., Zhong, M., 2020. Transfer learning for non-intrusive load monitoring. IEEE Trans. Smart Grid 11 (2), 1419–1429. http://dx.doi.org/10.1109/TSG.2019.2938068.

European Commission, Directorate-General for Communications Networks, Content and Technology, 2019. Ethics guidelines for trustworthy AI. Publications Office, http://dx.doi.org/10.2759/346720.

Fatouh, A.M., Nasr, O.A., Eissa, M.M., 2018. New semi-supervised and active learning combination technique for non-intrusive load monitoring. In: 2018 IEEE International Conference on Smart Energy Grid Engineering. SEGE, pp. 181–185. http://dx.doi.org/10.1109/SEGE.2018.8499498.

Ghai, B., Liao, Q.V., Zhang, Y., Bellamy, R., Mueller, K., 2021. Explainable active learning (XAL): Toward AI explanations as interfaces for machine teachers. Proc. ACM Hum.-Comput. Interact. 4 (CSCW3), http://dx.doi.org/10.1145/3432934.

Gu, Q., Dai, Q., Yu, H., Ye, R., 2021. Integrating multi-source transfer learning, active learning and metric learning paradigms for time series prediction. Appl. Soft Comput. 109, 107583.

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. http://dx.doi.org/10.1109/ACCESS.2020.3003778.

Huber, P., Calatroni, A., Rumsch, A., Paice, A., 2021. Review on deep neural networks applied to low-frequency nilm. Energies 14 (9), 2390.

Kaselimi, M., Protopapadakis, E., Voulodimos, A., Doulamis, N., Doulamis, A., 2022. Towards trustworthy energy disaggregation: A review of challenges, methods, and perspectives for non-intrusive load monitoring. Sensors 22 (15), http://dx.doi.org/10.3390/s22155872, URL https://www.mdpi.com/1424-8220/22/15/5872.

Kelly, J., Knottenbelt, W., 2015. The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes. Sci. Data 2 (1), 1–14.

Kirsch, A., van Amersfoort, J., Gal, Y., 2019. Batchbald: Efficient and diverse batch acquisition for deep Bayesian active learning. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., Garnett, R. (Eds.), In: Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc., URL https://proceedings.neurips.cc/paper_files/paper/2019/file/95323660ed2124450caaac2c46b5ed90-Paper.pdf.

Kothandaraman, D., Shekhar, S., Sancheti, A., Ghuhan, M., Shukla, T., Manocha, D., 2023. SALAD: Source-free active label-agnostic domain adaptation for classification, segmentation and detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 382–391.

Li, D., Li, J., Zeng, X., Stankovic, V., Stankovic, L., Xiao, C., Shi, Q., 2023. Transfer learning for multi-objective non-intrusive load monitoring in smart building. Appl. Energy 329, 120223. http://dx.doi.org/10.1016/j.apenergy.2022.120223, URL https://www.sciencedirect.com/science/article/pii/S0306261922014805.

Liebgott, F., Yang, B., 2017. Active learning with cross-dataset validation in event-based non-intrusive load monitoring. In: 2017 25th European Signal Processing Conference. EUSIPCO, pp. 296–300. http://dx.doi.org/10.23919/EUSIPCO.2017.8081216.

Martins, V.E., Cano, A., Junior, S.B., 2023. Meta-learning for dynamic tuning of active learning on stream classification. Pattern Recognit. 138, 109359.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., Fernández-Leal, Á., 2023. Human-in-the-loop machine learning: A state of the art. Artif. Intell. Rev. 56 (4), 3005–3054.

Murray, D., Stankovic, L., Stankovic, V., 2017. An electrical load measurements dataset of United Kingdom households from a two-year longitudinal study. Sci. Data 4, 1–12. http://dx.doi.org/10.1038/sdata.2016.122.

Murray, D., Stankovic, L., Stankovic, V., Lulic, S., Sladojevic, S., 2019. Transferability of neural network approaches for low-rate energy disaggregation. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 8330–8334. http://dx.doi.org/10.1109/ICASSP.2019.8682486.

Prabhu, V., Chandrasekaran, A., Saenko, K., Hoffman, J., 2021. Active domain adaptation via clustering uncertainty-weighted embeddings. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8505–8514.

Ren, P., Xiao, Y., Chang, X., Huang, P.Y., Li, Z., Gupta, B.B., Chen, X., Wang, X., 2021. A survey of deep active learning. ACM Comput. Surv. 54 (9), http://dx.doi.org/10.1145/3472291.

Sener, O., Savarese, S., 2017. Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489.

Settles, B., 2009. Active learning literature survey.

Sykiotis, S., Kaselimi, M., Doulamis, A., Doulamis, N., 2022. Electricity: An efficient transformer for non-intrusive load monitoring. Sensors 22 (8), http://dx.doi.org/10.3390/s22082926, URL https://www.mdpi.com/1424-8220/22/8/2926.

Tinati, R., Luczak-Roesch, M., Simperl, E., Hall, W., 2017. An investigation of player motivations in eyewire, a gamified citizen science project. Comput. Hum. Behav. 73, 527–540. http://dx.doi.org/10.1016/j.chb.2016.12.074, URL https://www.sciencedirect.com/science/article/pii/S0747563216309037.

Todic, T., Stankovic, V., Stankovic, L., 2023. An active learning framework for the low-frequency non-intrusive load monitoring problem. Appl. Energy 341, 121078. http://dx.doi.org/10.1016/j.apenergy.2023.121078, URL https://www.sciencedirect.com/science/article/pii/S0306261923004427.

Ueno, T., Ishibashi, H., Hino, H., Ono, K., 2021. Automated stopping criterion for spectral measurements with active learning. NPJ Comput. Mater. 7 (1), 139.

Wang, W., Chen, P., Xu, Y., He, Z., 2022. Active-MTSAD: multivariate time series anomaly detection with active learning. In: 2022 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks. DSN, IEEE, pp. 263–274.

Zhang, Z., Strubell, E., Hovy, E., 2022. A survey of active learning for natural language processing. arXiv preprint arXiv:2210.10109.

Zhu, J., Wang, H., Hovy, E., Ma, M., 2010. Confidence-based stopping criteria for active learning for data annotation. ACM Trans. Speech Lang. Process. (TSLP) 6 (3), 1–24.