Multilabel Appliance Classification With Weakly Labeled Data for Non-Intrusive Load Monitoring

Multi-Label Appliance Classification with Weakly Labeled Data for Non-Intrusive Load Monitoring

Giulia Tanoni, Emanuele Principi, *Member, IEEE,* and Stefano Squartini, *Senior Member, IEEE*

***Abstract*—Non-Intrusive Load Monitoring consists in estimating the power consumption or the states of the appliances using electrical parameters acquired from a single metering point. State-of-the-art approaches are based on deep neural networks, and for training, they require a significant amount of data annotated at the sample level, defined as *strong* labels.**

This paper presents an appliance classification method based on a Convolutional Recurrent Neural Network trained with weak supervision. Learning is formulated as a Multiple-Instance Learning problem, and the network is trained on labels provided for an entire segment of the aggregate power, defined as *weak* labels. Weak labels are coarser annotations that are intrinsically less costly to obtain compared to *strong* labels. An extensive experimental evaluation has been conducted on the UK-DALE and REFIT datasets comparing the proposed approach to three benchmark methods. The results obtained for different amounts of strongly and weakly labeled data and mixing UK-DALE and REFIT confirm the effectiveness of weak labels compared to fully supervised and semi-supervised benchmarks methods.

***Index Terms*—Non-intrusive load monitoring, Appliance classification, Weak labels, Multiple-instance learning, Multi-label classification.**

I. Introduction

CLIMATE change represents one of the biggest challenges of this century, and to keep the increase in global average temperatures below 1.5 °C, reducing electrical energy consumption is a necessary step [1]. As highlighted in the recent "Electricity and heat statistics" report by Eurostat, in 2019, the Industry sector was responsible for 36.5% of the yearly electricity consumption, followed by Services with 28.5%, and Households with 27.6% [2]. Residential users, thus, play a key role in this process, and as several research studies have shown, making them aware of how they use energy can provide savings of up to 15% [3].

However, granular monitoring of households' energy consumption is costly and invasive since it requires multiple dedicated metering devices. As a consequence, the research community developed more efficient techniques for reaching the same objective by using a single metering point that go under the name of Non-Intrusive Load Monitoring (NILM). NILM was firstly proposed by Hart in 1980 [4], and it consists in extracting information about appliances' operation by measuring electrical parameters only at the mains. In energy disaggregation, the task consists in the direct estimation of individual active powers of the appliances [5], while classification consists in determining their states [6] and then estimating active power by using average power values associated to each state.

In the last decade, the research community proposed several approaches for NILM that can be divided into two main categories: signal processing-based techniques, such as Graph Signal Processing [7] and Principal Component Analysis [8], single channel source separation techniques such as Non-negative Tensor Factorization [9], Matrix Factorization [10] and Sparse Coding and Dictionary Learning [11], [12] and machine learning approaches, such as Hidden Markov Models [13], [14], Support Vector Machines [15] and Deep Neural Networks (DNN) [5], [16]–[28]. Several works demonstrated that DNNs outperforms other approaches, and nowadays they represent the state-of-the-art. However, high performance is achieved at the cost of needing a significant amount of annotated data on a sample-by-sample basis (i.e., with *strong* labels) for training. This requires a significantly time-consuming data labeling phase, especially when several appliances are considered. Moreover, in target environments, transfer learning techniques may be necessary to achieve the desired performance level, and this requires acquiring and labeling data on-site [29]. Weak supervision is a learning strategy that lightens the requirements of strong labels by using coarse annotations, or unlabeled data [5]. Although the latter approach has been used in previous works [25], [26], it still represents an open research field.

In the following, we present a brief overview of DNN-based NILM methods and describe the contribution of this paper.

A. Related Works and Contribution

Following the work of Kelly et al. [16] that firstly approached NILM with DNN, several alternatives have been proposed [30]. Based on the strategy adopted for training the networks, they can be divided into two groups: the first comprises methods based on strongly supervised learning and is the most numerous group [5], [16]–[24], and the second methods based on semi-supervised learning [25], [26]. With "strongly supervised learning" and "semi-supervised learning" we refer to the definitions reported in [31].

Among strongly supervised approaches for energy disaggregation, Kelly et al. [16] proposed three different architectures to estimate the appliance power consumption from sequences of aggregate samples. The architectures were based on a denoising Autoencoder (dAE), a Recurrent Neural Network (RNN), and the so-called Regress Start Time, End Time & Power network composed of convolutional and fully connected layers. Similarly, in [5], two Convolutional Neural

The authors are with the Department of Information Engineering, Università Politecnica delle Marche, 60131 Ancona, Italy (e-mail: g.tanoni@pm.univpm.it; e.principi@univpm.it; s.squartini@univpm.it).

Manuscript received April 19, 2005; revised August 26, 2015.

Networks (CNN) were trained with Root Mean Squared Error (RMSE) loss function, one using the sequence-to-point approach and the other the sequence-to-sequence approach. Kaselimi et al. [18] proposed a CNN-based recurrent architecture composed of two convolutional multi-channel modules. A Dilated-Residual Network has been proposed in [20] to reduce the vanishing gradient problem and training degradation. Langevin and colleagues [21] used a Variational Auto-Encoder (VAE) to improve the disaggregation of multi-state appliances' power consumption and generalization performance. In [22], a method based on Generative Adversarial Networks (GANs) has been presented, where a dAE has been trained by using an adversarial training strategy, and a recurrent CNN has been employed as discriminator. A Conditional-GAN approach was proposed in [23], where the problem was modeled as a sequence-to-subsequence estimation task.

Strongly supervised approaches for appliances' states classification have been presented in [19], [28]. In [19], the method is based on a CNN, and temporal pooling is used to aggregate features of different time resolutions. In [28], a Long-Short Term Memory (LSTM) autoencoder has been implemented to perform a multi-label classification and model the temporal variability of the power time series.

Some works proposed multi-task architectures, generally double-branched, to perform both classification and disaggregation and improve the performance of each task. Murray et al. [24] proposed two architectures, one CNN-based and the other based on Gated-Recurrent Units (GRU). Both networks were composed of two branches, one for classification and the other for disaggregation. Piccialli and Sudoso [17] also proposed a dual tasks architecture where the regression subnetwork was improved with an attention layer, and the regression output was combined with the related classification prediction.

A drawback of strongly supervised methods is that they require large amounts of labeled data for training the networks. Semi-supervised approaches have been proposed to address this aspect. Unlike strongly supervised methods, they are able to exploit unlabeled data, thus they require fewer annotations to achieve state-of-the-art performance [25], [26]. Yang and colleagues [25] proposed a teacher-student architecture based on a Temporal Convolutional Network (TCN) for multi-label appliance classification. In [26], Virtual Adversarial Training (VAT) was used for energy disaggregation to train a sequence-to-point network. Learning was based on a regularization term calculated as the average of local distributional smoothness (LDS), and superior performance was obtained compared to fully supervised learning.

An alternative to these approaches is represented by learning with *inexact supervision*, a form of weakly supervised learning where labels are provided at a coarser level compared to strongly supervised methods [31]. In this way, the annotation effort is significantly reduced compared to strongly supervised methods, but since annotations are still provided, they can improve the performance compared to semi-supervised approaches. This supervision strategy has been used in several application domains, such as computer vision [32], sound event detection [33], [34], and text processing [35], however, up to the authors' knowledge, it has never been applied to NILM for multi-label appliance classification.

Based on these considerations, this work explores this approach proposing a multi-label appliance classification method for low-frequency data based on Convolutional Recurrent Neural Networks (CRNNs) [36] and weakly labeled data. Differently from the examined literature, the proposed method relaxes the requirement of supervised methods for a large amount of strongly labeled data for training and improves the performance compared to semi-supervised approaches. Previous works applied CRNNs and weak labels to other application domains [33], but up to our knowledge this has never been performed in the context of NILM, particularly for multi-label appliance classification. Here, we model this task as a Multiple-Instance Learning (MIL) problem [37], and we modified the CRNN architecture to exploit both strong and weak annotations of the training data. In this way, different levels of information are used in the training phase, thus reducing labeling costs and improving generalization ability. Moreover, Clip Smoothing [38] has been integrated in the network for dealing with false activations. The UK-DALE [39] and REFIT [40] datasets have been used for performance evaluation, and the effectiveness of the method has been assessed in two experiments by varying the amount of strongly and weakly labeled data available for training. The obtained results demonstrate the increased generalization ability of the proposed method compared to a strongly supervised strategy while reducing the labeling effort. In a third experiment, we evaluated if mixing strongly and weakly labeled data of the two datasets provides an advantage, and the results confirmed that weak labels improve the performance on both UK-DALE and REFIT test sets. Up to our knowledge, this is the first work in which multi-label appliance classification has been addressed by using weakly labeled data.

In summary, the contributions of this work are the following:

- We propose a weakly supervised approach to multi-label appliance classification based on a CRNN, and to exploit weakly labeled data, we formulate the task as a MIL problem.
- We demonstrate that the proposed method is able to obtain superior performance compared to supervised and semi-supervised methods, particularly when the number of weak labels exceeds that of strongly annotated bags.
- We demonstrate that our method is able to reduce the quantity of strongly annotated data compared to supervised methods, while achieving comparative performance.

The outline of the paper is the following. Section II defines the multi-label appliance classification problem; Section III explains in details the proposed method; Section IV describes the experimental setup; Section V presents and discusses the obtained results, and finally, Section VI concludes the paper and presents future works.

II. Problem Statement

Denoting with $y(t)$ the total active power consumed in a building, with $x_n(t)$ the active power of the n-th appliance,

and with $s_n(t) \in \{0, 1\}$ its state, we can write their relationship as follows:

$$y(t) = \sum_{n=1}^{N} s_n(t)x_n(t) + \epsilon(t), \tag{1}$$

where $\epsilon(t)$ is the measurement noise, and

$$s_n(t) = \begin{cases} 0, & \text{if appliance } n \text{ is OFF at the time index } t, \\ 1, & \text{if appliance } n \text{ is ON at the time index } t. \end{cases} \tag{2}$$

Typically, the interest is in monitoring a subset of appliances while the remaining contribute to the noise term [16]. Without loss of generality, we can suppose that the monitored subset is composed of the first K appliances. Equation (1), thus, can be rewritten as follows:

$$y(t) = \sum_{k=1}^{K} s_k(t)x_k(t) + v(t), \tag{3}$$

where the first term is the power of appliances of interest, and the $v(t)$ the cumulative noise term given by:

$$v(t) = \sum_{m=K+1}^{N} s_m(t)x_m(t) + \epsilon(t). \tag{4}$$

In the case where the objective is the direct estimation of the individual active power signals $x_k(t)$, NILM is a *regression* problem, and it has been treated as a denoising task [16] or as a blind source separation task [9]. In the case where the objective is the estimation of appliances states $s_k(t)$, NILM represents a *multi-label classification* problem [41]. In both cases, the algorithm exploits only the knowledge of the aggregate signal $y(t)$.

In this work, we are concerned with multi-label appliance classification, thus the objective is the estimation of state variables $s_k(t)$ of the K appliances of interest.

III. Proposed Method

In this work, we model multi-label appliance classification as a MIL problem, and we employ a deep neural network trained both on strongly and weakly labeled data. This solution has the dual consequence of improving generalization capability compared to fully supervised approaches and reducing labeling costs. Indeed, in principle, appliance classification does not necessarily require active power signals of individual appliances for training since annotation can be performed manually and the metering infrastructure can be simplified. On the other hand, manual annotation with strong labels requires a significant human effort that would not be easy to afford. On the contrary, with weak labels, annotations are provided on a wide temporal window, thus, it is sufficient to indicate if an appliance was active or not within that segment by using only a single weak label. In this sense, the method can deal with the inexactness that may originate from mislabeling by manual annotators.

MIL is a variant of supervised learning and a particular form of weak supervision [42]. In MIL, learning examples are represented by *bags* composed of multiple *instances* (e.g., feature vectors, raw samples), and labels are provided only at the bag level. During prediction, the objective can be to classify bags, individual instances, or both [43]. MIL can be applied to single-label classification tasks, where bags and instances are assigned only one label, or to multi-label classification tasks, where labels are multiple (multi-instance multi-label learning, MIML) [44]. Labels assigned to bags depend on the labels of individual instances inside them. In binary classification tasks, the *standard multiple instance assumption* states that the necessary and sufficient condition for a bag to be assigned a positive label is that one of its instances is positive, but later works have proposed other alternatives [45]. The same criterion can be easily extended to multi-class problems.

In the proposed method, instances are represented by the raw samples of the aggregate signal $y(t)$, and the related labels are represented by one-hot vectors $\mathbf{s}(t) \in \mathbb{R}^{K\times 1}$ defined as:

$$\mathbf{s}(t) = [s_1(t), s_2(t), \ldots, s_K(t)]^T. \tag{5}$$

A bag is a segment of $y(t)$ with length L. Supposing that $y(t)$ is divided into disjointed segments, the j-th bag is represented by the following vector:

$$\mathbf{y}_j = [y(jL), \ldots, y(jL + L - 1)]^T \in \mathbb{R}^{L\times 1}. \tag{6}$$

The related label is again encoded as a one-hot vector $\mathbf{w}_j \in \mathbb{R}^{K\times 1}$. As aforementioned, $\mathbf{w}_j$ depends on the instance labels inside it. Denoting with $\mathbf{S}_j = [\mathbf{s}(jL), \mathbf{s}(jL + 1), \ldots, \mathbf{s}(jL + L - 1)] \in \mathbb{R}^{K\times L}$ the set of instance labels related to segment j, the relationship can be represented by a pooling function $\mathbf{b} : \mathbb{R}^{K\times L} \to \mathbb{R}^K$ such that

$$\mathbf{w}_j = \mathbf{b}(\mathbf{S}_j). \tag{7}$$

Several pooling functions have been proposed in the literature, each having different characteristics [33]. The pooling function used in this work will be defined in the following section, along with the neural network architecture. The *bag* level, thus, contains information on the presence of one or more appliances in a time window, while, at the *instance* level, this information is provided at sample resolution. Bag labels are noisy, coarse, and inexact, thus they are commonly referred to as *weak labels*, while instance labels are referred to as *strong labels*.

In this work, the objective is to identify if an appliance is active or not at the sample level, thus the goal is to learn a function $\mathbf{f} : \mathbb{R}^L \to \mathbb{R}^{K\times L}$ such that:

$$\hat{\mathbf{S}} = \mathbf{f}(\mathbf{y}), \tag{8}$$

where $\mathbf{y}$ is an unknown aggregate segment, and $\hat{\mathbf{S}}$ contains the estimated instance-level probabilities for each class. The bag index j has been omitted for simplicity.

A. Neural Network Architecture

The function $\mathbf{f}(\cdot)$ in (8) is represented by a CRNN, and the related block scheme is depicted in Fig. 1. As mentioned in Section I, CRNNs have been previously applied to other application domains [32]–[35], but never for multi-label appliance classification, to the best of our knowledge. For each segment $\mathbf{y}_j$, the network produces the related instance-level

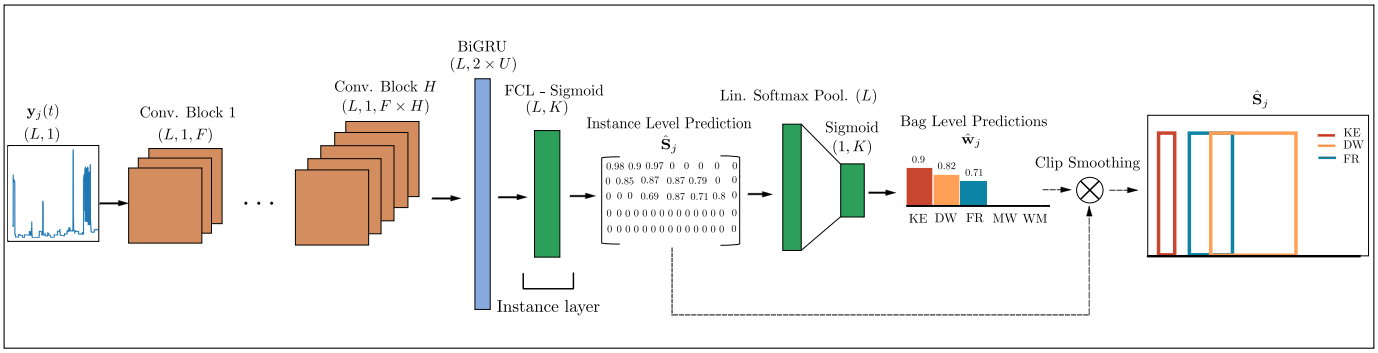


Fig. 1. Block scheme of the proposed approach. The network takes as input the aggregate power related to the j-th window and produces two outputs, one from the instance layer and one from the bag layer. The multiplication is related to clip smoothing. T is length of input and output window, FCL stands for Fully Connected Layer, K is number of appliances, and KE, MW, FR, WM, and DW stand respectively for Kettle, Microwave, Fridge, Washing Machine, Dishwasher.

estimate $\hat{\mathbf{S}}_j$ and the bag-level estimate $\hat{\mathbf{w}}_j$ for each appliance in interest.

The CRNN is formed of two main parts. The first comprises H blocks with a 2D convolutional layer with F filters and kernel of size K_e, a batch normalization layer, an activation layer with Rectified Linear Unit (ReLU) function, and dropout for regularization. The output of the last convolutional block feeds the recurrent part of the network, which is formed by a bidirectional layer of U Gated Recurrent Units [46]. The final layer comprises a fully-connected layer with a sigmoid activation function. This layer, denoted as *instance* layer, produces the instance-level estimate $\hat{\mathbf{S}}_j$.

After the instance layer, a pooling layer followed by a sigmoid activation function produces the bag-level prediction $\hat{\mathbf{w}}_j$. The pooling layer implements the pooling function $\mathbf{b}(\cdot)$ defined in the previous section. The proposed network, thus, has both an instance-level output and a bag-level output. In this way, it is possible to conjugate MIL with the supervised learning strategy based on strong labels.

As aforementioned, several alternatives exist for the pooling function. Based on the analysis conducted in [33], we chose *linear softmax* since it achieved the highest localization performance. The linear softmax pooling function calculates the k-th element of $\hat{\mathbf{w}}_j$ of the bag-level prediction as:

$$\hat{w}_{k,j} = \frac{\sum_{t=jL}^{L(j+1)-1} \hat{s}_k^2(t)}{\sum_{t=jL}^{L(j+1)-1} \hat{s}_k(t)}. \tag{9}$$

In this way, the bag-level prediction is calculated as weighted average of the instance-level predictions, with the larger ones receiving a larger weight [33].

B. Learning

Learning from bags raises important challenges that are unique to MIL formulation [43]. As aforementioned, from a single weak label, multiple combinations of instances exist that can produce the same bag label, thus it is expected that a learning algorithm trained only on weakly annotated data achieves inferior results than training on strongly annotated data. The availability of several datasets for NILM with strong annotations [47] motivated us to train the CRNN by using both weak and strong labels.

More in detail, denoting with $\mathcal{T}_w = \{(\mathbf{y}_1, \mathbf{w}_1), \ldots, (\mathbf{y}_{M_w}, \mathbf{w}_{M_w})\}$ the set of training bags annotated with weak labels and with $\mathcal{T}_s = \{(\mathbf{y}_1, \mathbf{w}_1, \mathbf{S}_1), \ldots, (\mathbf{y}_{M_s}, \mathbf{w}_{M_s}, \mathbf{S}_{M_s})\}$ the set of training bags annotated with both strong and weak labels, learning is performed on the training set $\mathcal{T} = \mathcal{T}_w \bigcup \mathcal{T}_s$.

The loss function is composed of the weighted sum of the binary cross-entropy losses calculated on strong and weak labels:

$$\mathcal{L} = \mathcal{L}_s + \lambda \mathcal{L}_w, \tag{10}$$

where $\mathcal{L}_s$ and $\mathcal{L}_w$ are respectively the loss related to strongly and weakly labeled data, and the weight λ balances their contribution.

The two loss terms $\mathcal{L}_s$ and $\mathcal{L}_w$ are the binary cross-entropy for each class an they are given by:

$$\mathcal{L}_s = -\frac{1}{K}\frac{1}{L}\sum_{k=1}^{K}\sum_{t=1}^{L}[s_k(t)\log(\hat{s}_k(t)) + (1 - s_k(t))\log(1 - \hat{s}_k(t))], \tag{11}$$

$$\mathcal{L}_w = -\frac{1}{K}\sum_{k=1}^{K}[w_k \log(\hat{w}_k) + (1 - w_k)\log(1 - \hat{w}_k)], \tag{12}$$

where the segment index j has been omitted for simplicity of notation. Training has been performed by using the Adam algorithm [48].

C. Post-Processing

The bag-level and instance-level outputs of the CRNN are class probabilities estimates in the range $[0, 1]$. These values are then transformed into 0 or 1 by using a threshold to determine whether an appliance is active or not.

This procedure, however, is prone to producing outputs where few isolated instances are 0 or 1. A popular solution is to apply a median filter to the outputs to reduce such spurious values. The median filter operates by calculating the median value within a segment of a certain length and replacing contiguous instances with a duration shorter than half the segment length with the median value.

Additionally to median filtering, here we explore a recent technique presented in [38] and named *clip smoothing*.

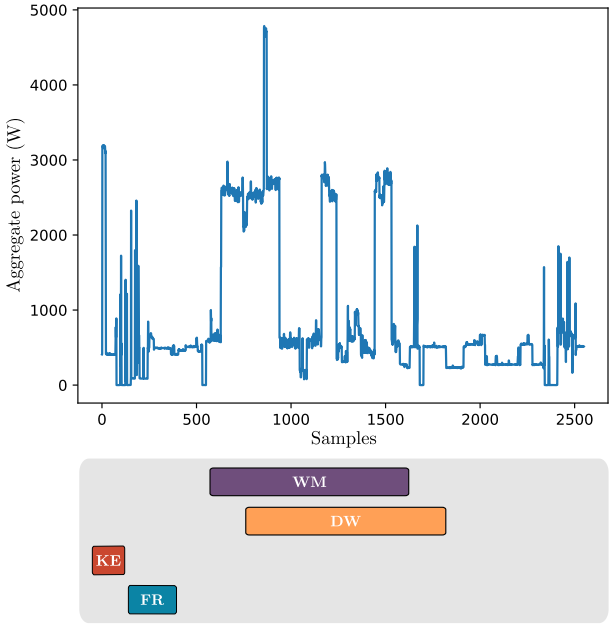


Fig. 2. An example of aggregate segment from house 2 of REFIT with the related labels.

Clip smoothing operates before thresholding and consists in multiplying the instance-level prediction with the bag-level prediction (Fig. 1). The rationale of clip smoothing is that instance and bag level predictions should be coherent: if a bag prediction is close to 0, instance-level predictions should be all close to 0, and vice versa. Multiplying the two predictions enforces this relationship. An advantage over median filtering is that clip smoothing is a learnable procedure intrinsic to the network. Note that the use of clip smoothing is only possible when the network outputs weak and strong predictions, thus it represents an additional advantage over strongly supervised methods.

IV. Experimental Setup

The proposed method has been implemented in Python using Tensorflow 2.4 and Keras. The source code is available here[1].

A. Datasets

The experiments have been conducted on two datasets, UK-DALE (UK Domestic Appliance-Level Electricity) [39], and REFIT [40]. The monitored appliances have been selected based on the recent literature [23], [29], [49], and they are the following: Kettle, Microwave, Fridge, Washing Machine, and Dishwasher. Each dataset has been processed to create two sets of bags, one for UK-DALE and one for REFIT, then used for training and testing the proposed method. The procedure for creating these sets is described in the following.

1) General procedure: The first step consisted in extracting the activations of the monitored appliances from the datasets. This has been performed by using NILMTK [50] using the parameters in [16] for UK-DALE, and the ones in [40] for REFIT.

The second step consisted in combining the extracted activations randomly to create bags with one to four concurrent appliances. In each dataset, the maximum length of an activation is about 1500, so we decided to set the bag length L to 2550. In this way, activations can be properly placed within the segment. The location of the activation inside the bag is determined randomly. Generally, the bag length can have a role in performance; however, in the following experiments, it is important that the same value is used in all the methods considered to evaluate only the influence of weak labels.

The third step consisted in the extraction of the noise contribution, i.e., the term $v(t)$ in (3). This has been obtained by selecting a random aggregate power segment of length L and then subtracting the monitored appliances' activations from it. The extracted noise term has been then summed to bags created in step two. This procedure is repeated for each bag, so noise terms are all different. Moreover, noise terms and activations of the monitored appliances always belong to the same building.

For each appliance, strong labels in a bag are set to 1 if a sample belongs to an activation (i.e., the appliance is in the ON state), and 0 otherwise. Weak labels are set to 1 if the activation of an appliance is present in the bag.

An example of aggregate segment related to house 2 of the REFIT dataset is shown in Fig. 2.

2) UK-DALE: The dataset contains data from 5 houses, with aggregate power readings sampled every 1 s and appliance-level measurements sampled every 6 s. The following dates were considered:

- house 1: 06/01/2016-31/08/2016;
- house 2: 01/06/2013-31/08/2013;
- house 3, 4: 16/03/2013-05/04/2013;
- house 5: 06/29/2014-09/05/2014.

We downsampled the aggregate active power readings from 1 s to 6 s, and we aligned the mains to the appliance readings using NILMTK [50]. All the houses were included, but only the Kettle and the Fridge were considered for houses 3 and 4. For training and validation, we used data from houses 1, 3, 4, and 5, while house 2 was kept out for testing on unseen data. The training, validation, and test set characteristics are reported in Table I: "Strongly and weakly annotated set" refers to bags with both strong and weak labels, while "Weakly annotated set" refers to bags annotated only with weak labels. For each appliance, the table reports the number of strong labels, i.e., the total number of samples, and the number of weak annotations, i.e., the total number of bags where it is present.

3) REFIT: The dataset contains measurements from 21 houses. Data were downsampled uniformly to 8 s. Each house contains different appliance-level power readings with a maximum of 4 devices. We used the same houses reported in [29], a part from house 20 since it contains only two Kettle activations. Houses 4, 9, and 15 have been used to test on unseen data, while the remaining for training. As in [49], we considered the following date intervals:

- houses 9, 12, 18: 07/12/2013-08/07/2015;

[1]https://github.com/GiuTan/Weak-NILM

TABLE I
UK-DALE DATASET CHARACTERISTICS. NUMBERS ARE IN THOUSANDS.

	Strongly and weakly annotated set						Weakly annotated set	Average power consumption in a activation (W)
	Training (k)		Validation (k)		Test (k)		Training (k)	
Appliances	Strong	Weak	Strong	Weak	Strong	Weak	Weak	
Kettle	996.6	31.4	196.3	6.9	91.4	2.2	11.7	1996
Microwave	849.7	31.0	157.2	7.0	83.8	2.4	11.9	1107
Fridge	1221.9	4.8	709.4	2.9	130.3	0.6	31.2	91
Washing Machine	837.7	1.2	881.4	1.2	102.5	0.2	30.9	487
Dishwasher	554.5	0.6	790.1	0.9	87.5	0.2	31.3	723
Nr. of bags	41.720		10.428		3.271		58.213	

TABLE II
REFIT DATASET CHARACTERISTICS. NUMBERS ARE IN THOUSANDS.

	Strongly and weakly annotated set						Weakly annotated set	Average power consumption in a activation (W)
	Training (k)		Validation (k)		Test (k)		Training (k)	
Appliances	Strong	Weak	Strong	Weak	Strong	Weak	Weak	
Kettle	2917.3	62.2	619.2	15.5	623.9	20.9	3.0	2048
Microwave	1858	40	455.6	9.9	467.7	12.0	20.0	893
Fridge	6030	10	1635.5	3.0	1396.1	1.4	55.0	90
Washing Machine	2402.2	6.1	2062.9	5.7	228.3	0.5	55.0	513
Dishwasher	2263.2	2.9	2822.5	4.4	472.0	0.5	53.0	881
Nr. of bags	97.385		24.297		22.425		102.078	

- houses 10, 17: 20/11/2013-30/06/2015;
- houses 2, 5, 7, 16: 17/09/2013-08/07/2015;
- house 13: 26/09/2013-08/07/2015;
- houses 3, 4, 6, 8, 11, 15, 19: 26/09/2013-08/07/2015.

Details on the training, validation and test sets are reported in Table II.

4) Pre-processing: Aggregate data were normalized with mean and standard deviation values computed from the training set.

B. Benchmark Methods

The proposed method has been compared to two benchmark methods that recently appeared in the literature. The first is the LSTM network presented in [16], that has been already used as benchmark method for classification in [25], [51]. As in [25], to perform multi-label classification, the last layer of the network has been replaced with a fully-connected layer composed of 5 neurons followed by a sigmoid activation function. The network is trained only on strongly labeled data using the loss defined in (11). The second benchmark method is the Semi-Supervised Multi-Label TCN (SSML-TCN) proposed in [25]. The network has been implemented and trained with the hyperparameters reported by the authors. The SSML-TCN network has been trained using both strongly and weakly labeled data, with the latter used as unlabeled data. The resulting loss function is the sum of the cross-entropy loss defined in (11) and the consistency loss computed on the student and teacher predictions as in [25].

The proposed solution has been evaluated also against a CRNN trained only on strongly annotated data as the LSTM network. Referring to Fig. 1, this means that this network outputs only instance-level predictions, and it does not comprise the linear softmax pooling layer and clip smoothing. This network will be denoted as S-CRNN in the following.

C. Evaluation Metrics

The performance of the algorithm has been assessed at the instance level, while the bag-level output has not been considered. The metrics used in the evaluation are the F_1-score (F_1) and the Total Energy Correctly Assigned (TECA) [52]. The F_1-score is used to evaluate the model prediction ability, balancing between the presence of accurate classification and false activations. F_1-score for appliance k is calculated as:

$$F_1^{(k)} = \frac{2 \cdot TP^{(k)}}{2 \cdot TP^{(k)} + FP^{(k)} + FN^{(k)}}, \tag{13}$$

where $TP^{(k)}$ is the number instances correctly assigned to appliance k (true positives), $FP^{(k)}$ is the number instances incorrectly assigned to appliance k (false positives), and $FN^{(k)}$ is the number instances incorrectly assigned to other appliances (false negatives). The average performance across appliances is calculated by using the micro-averaged F_1-score:

$$F_1\text{-micro} = \frac{2 \cdot \sum_{k=1}^{K} TP^{(k)}}{\sum_{k=1}^{K} \left(2 \cdot TP^{(k)} + FP^{(k)} + FN^{(k)}\right)}. \tag{14}$$

TECA has been introduced in [52] to evaluate the energy disaggregation error and is defined as:

$$\text{TECA} = 1 - \frac{\sum_k \sum_t |\hat{x}_k(t) - \bar{x}_k(t)|}{2 \sum_t \bar{y}(t)}, \tag{15}$$

where $\hat{x}_k(t)$ is the power of appliance k at the time instant t, $\bar{x}_k(t)$ the related ground-truth power, and $\bar{y}(t) = \sum_k \bar{x}_k(t)$. The estimated power $\hat{x}_k(t)$ is reconstructed by multiplying the estimated states $\hat{s}_k(t)$ and the average power in an activation of appliance k, while $\bar{x}_k(t)$ by considering the ground-truth states $s_k(t)$. Average powers are reported in Table I and Table II respectively for the UK-DALE and REFIT datasets.

Differently from F_1-micro, TECA is more influenced by high power appliances, thus, it may result in high values even when low-power appliances are classified poorly [52].

D. Experimental procedure

Referring to the strongly and weakly annotated training sets reported in Table I and Table II, we performed three experiments in different training conditions:

1) Experiment 1: all the weakly annotated training bags are used for training, while the number of strongly annotated bags is varied from 0% to 100% (step 20%);
2) Experiment 2: the amount of strongly annotated bags is fixed to 20%, while the number of weakly annotated bags is varied from 0% to 100% % (step 20%);
3) Experiment 3: we evaluate if mixing strongly labeled data of UK-DALE and weakly labeled data of REFIT improves the performance on the respective test sets compared to training only on strongly labeled data.

The objective of the first two experiments is to evaluate how weakly labeled data influence performance, particularly if they indeed provide an improvement when the amount of strongly labeled data is modest. In Experiment 1, we progressively decrease the amount of strongly labeled data, and we evaluate when the contribution of weakly labeled data is significant. In Experiment 2, we consider a certain amount of strongly labeled data for which weakly labeled data provide a performance improvement, and then we vary the amount of weakly labeled data. In this way, we study which amount of weakly labeled data provides a performance improvement. Note that in the first experiment, 0% of strongly annotated training data means that training is performed by using only weak supervision. In the third experiment, we consider the case where it is possible to acquire additional data on a target environment, but annotation is performed only with weak labels. For example, when end users perform annotation as a result of a prompt to label an aggregate power segment in which unknown loads are present. In this case, users annotate the entire segment with a weak label, thus indicating only whether an appliance was active or not. In this situation, we want to evaluate if mixing this additional data with strongly annotated data from a public dataset provides some benefits. To perform this evaluation, weakly labeled data and test data from REFIT have been resampled to 6 s as UK-DALE strongly labeled data.

A tuning procedure has been performed for each training condition to find the values of hyperparameters that achieve the highest performance on the validation set. The procedure has been conducted separately for the proposed method and the S-CRNN network. In this way, we reduce the possibility that the performance difference is due to a wrong or biased choice of the values of the hyperparameters.

TABLE III
TRAINING HYPERPARAMETERS NOT SUBJECT TO TUNING.

Parameters	Value
Batch size	64 (UK-DALE) 128 (REFIT)
Learning rate	0.002
Training epochs	1000
Patience	15
Stride	1
Padding	Same
Weights initializer	Glorot Uniform
Bias initializer	Zeros

TABLE IV
HYPERBAND PARAMETERS.

Parameters	[Range], Step	Distribution
Max epochs	20	-
Factor	2	-
U	[8, 16, 32, 64, 128, 256]	Random choice
H	[2, 6], 1	Uniform
K_e	[3, 7], 2	Uniform
p	[0.1, 0.5], 0.1	Uniform

TABLE V
HYPERPARAMETERS DETERMINED AFTER TUNING.

Dataset	% Weak	% Strong	H	U	K_e	p
UKDALE	100	0	4	16	5	0.1
		20-100	3	64	5	0.1
	0	20-100	3	64	5	0.1
	20-100	20	3	64	5	0.1
REFIT	0	20	4	256	5	0.3
		40	3	64	3	0.2
		60	3	128	3	0.2
		80	4	128	7	0.3
		100	5	64	3	0.2
	20, 40		4	256	5	0.3
	60, 80	20	3	64	3	0.1
	100		3	64	3	0.3
		40	4	64	5	0.1
	100	60	4	64	3	0.2
		80, 100	4	64	5	0.1
		0	3	32	3	0.1

Hyperband [53] has been used for searching the following hyperparameters: number of convolutional layers (H), number of units in the recurrent layers (U), the dropout rate (p), and kernel size (K_e). The number of filters F in each convolutional layer increases doubling layer by layer with an initial value of 32. Table III reports the values of the hyperparameters not subject to tuning, Table IV the hyperparameters of Hyperband, and Table V the values determined after tuning for the different training conditions. The value of λ has been initially set to 1. Then, we monitored the values assumed by the two losses $\mathcal{L}_s$ and $\mathcal{L}_w$, and we selected the final value of λ to make them of the same order of magnitude.

E. Post-processing

We selected whether to apply median filtering, clip smoothing, or none of the two by evaluating the results obtained on the validation set. Median filtering did not improve the classification performance, so it was not used.

The threshold for obtaining the final classification values from output probabilities has been selected on the validation set, based on the value that maximizes the F_1-score.

V. RESULTS AND DISCUSSION

This section firstly presents the results obtained with a fixed amount of weakly and strongly labeled data (Experiment 1 and 2), then the results obtained by mixing strongly labeled data of UK-DALE and weakly labeled data of REFIT (Experiment 3). As a first note, Table VI reports the maximum model size, and

TABLE VI
MAXIMUM MODEL SIZE, TRAINING AND TEST TIME OF ALL THE EVALUATED METHODS.

Method	Max Model Size	Training Time	Testing Time
LSTM	4.97 MB	172 ms/step	3.6 ms
SSML-TCN	6.18 MB	143 ms/step	4 ms
S-CRNN	4 MB	215 ms/step	0.3 ms
Proposed	1.39 MB	214 ms/step	0.3 ms

TABLE VII
RESULTS OBTAINED ON THE UK-DALE AND REFIT DATASETS BY USING WEAKLY LABELED DATA ONLY.

	0% Strong, 100% Weak						
	KE	MW	FR	WM	DW	F_1-micro	TECA
UKDALE	0.89	0.76	0.29	0.36	0.39	0.52	0.57
REFIT	0.21	0.17	0.01	0.09	0.17	0.11	0.06

the training and inference times for all the evaluated methods. Note that the network of the proposed approach is the smallest of the evaluated methods and, along with S-CRNN, requires the least amount of time for testing. On the other hand, it requires more time for training, as S-CRNN, compared to other methods. Training and test times have been obtained on a NVIDIA DGX Station A100 [54].

A. Experiment 1: Fixed amount of weakly labeled data

1) UK-DALE: The results related to this experiment are reported in Table VII, Table VIII, and Fig. 3. Table VII shows the results obtained by using only weak labels for training. Observing the results, Kettle and Microwave F_1-scores are above 0.75, with the former equal to 0.89. On the contrary, Fridge, Washing Machine, and Dishwasher scores are below 0.5, meaning that the absence of strong labels impacts their results more than other appliances.

Table VIII reports the results obtained when strongly labeled data are used concurrently with weak labels. In terms of F_1-micro, apart when 100% of strongly labeled data is used, the proposed method provides better performance with respect to benchmark approaches. In terms of TECA, the S-CRNN achieves the overall greatest value, but on average the proposed method achieves superior performance. In particular with 20%, 40% and 60% of strongly labeled data, i.e., when the number of strong labels is modest, the proposed method shows more accuracy. Considering the average across the different percentages of strongly labeled data (last line of Table VIII), the proposed method significantly improves the performance of all the appliances, with the only exception of Kettle. On average, the F_1-micro improvements compared to LSTM, SSML-TCN, and S-CRNN are respectively 16.22%, 36.51%, and 3.61%. Among benchmark methods, S-CRNN performs more accurately compared to LSTM and SSML-TCN.

TABLE VIII
RESULTS OBTAINED ON THE UK-DALE DATASET RELATED TO EXPERIMENT 1. BEST SCORES FOR EACH STRONG PERCENTAGE ARE HIGHLIGHTED IN BOLD. BEST SCORE AMONG ALL THE PERCENTAGE ARE UNDERLINED.

% Strong	Method	KE	MW	FR	WM	DW	F_1-micro	TECA
20	LSTM [16]	0.95	0.74	0.35	0.44	0.69	0.61	0.79
	SSML-TCN [25]	0.82	0.70	0.16	0.39	0.60	0.46	0.60
	S-CRNN	0.98	0.67	0.42	0.80	**0.81**	0.72	0.86
	Proposed	**0.99**	**0.92**	**0.58**	**0.87**	0.74	**0.81**	**0.91**
40	LSTM [16]	**0.99**	0.93	0.59	0.59	**0.88**	0.77	0.89
	SSML-TCN [25]	0.92	0.86	0.37	0.62	0.48	0.64	0.77
	S-CRNN	**0.99**	0.95	**0.70**	**0.88**	0.84	0.86	**0.94**
	Proposed	0.98	**0.96**	0.69	0.87	**0.88**	**0.87**	**0.94**
60	LSTM [16]	**0.99**	0.93	0.53	0.69	**0.84**	0.76	0.89
	SSML-TCN [25]	0.95	0.87	0.39	0.70	0.64	0.68	0.82
	S-CRNN	**0.99**	**0.96**	0.67	**0.90**	0.71	0.84	0.93
	Proposed	**0.99**	**0.96**	**0.70**	0.87	0.83	**0.86**	**0.94**
80	LSTM [16]	**0.99**	0.95	0.58	0.68	0.69	0.75	0.88
	SSML-TCN [25]	0.96	0.84	0.41	0.76	0.60	0.68	0.83
	S-CRNN	**0.99**	**0.96**	**0.70**	0.83	**0.89**	0.86	**0.94**
	Proposed	0.98	0.95	**0.70**	**0.89**	0.87	**0.87**	0.93
100	LSTM [16]	**0.99**	0.95	0.65	0.78	0.75	0.80	0.91
	SSML-TCN [25]	0.97	0.85	0.43	0.76	0.61	0.71	0.84
	S-CRNN	**0.99**	**0.96**	0.70	**0.89**	**0.91**	**0.88**	**0.95**
	Proposed	0.98	**0.96**	**0.74**	**0.89**	0.86	**0.88**	0.93
AVG.	LSTM [16]	0.98	0.90	0.54	0.64	0.77	0.74	0.87
	SSML-TCN [25]	0.92	0.82	0.35	0.65	0.59	0.63	0.77
	S-CRNN	**0.99**	0.90	0.64	0.86	0.83	0.83	0.92
	Proposed	0.98	**0.95**	**0.68**	**0.88**	**0.84**	**0.86**	**0.93**

Fig. 3 shows the difference between the F_1-scores of each appliance, the F_1-micro, and the TECA of the proposed method and S-CRNN for the different percentages of strongly labeled data. S-CRNN has been chosen among benchmark methods since it is the best performing among them. Moreover, it allows highlighting the contribution of weak labels since the architecture is very similar to the one of the proposed method. It is evident that the greatest improvement occurs when the percentage of strongly labeled data is 20%, i.e., when the difference between the amount of strongly and weakly labeled data is the largest, meaning that in this case weak labels influence more the learning phase. Apart from the Dishwasher, the improvement is consistent for all the appliances.

Above 20%, the improvement of the proposed method reduces, but it remains significant up to 100%. In this case, the F_1-micros are comparable, meaning that the contribution of weak labels is less important. Observing the performance of the individual appliances, weak labels influence to a lesser extent the performance of Kettle and Microwave.

The appliances that exhibit a less consistent behavior with weak labels are Dishwasher and Washing Machine. Regarding the former, with 40% and 60% of strongly labeled data, the proposed method improves the performance with respect to full supervision, while with 20%, 80%, and 100% the perfor-

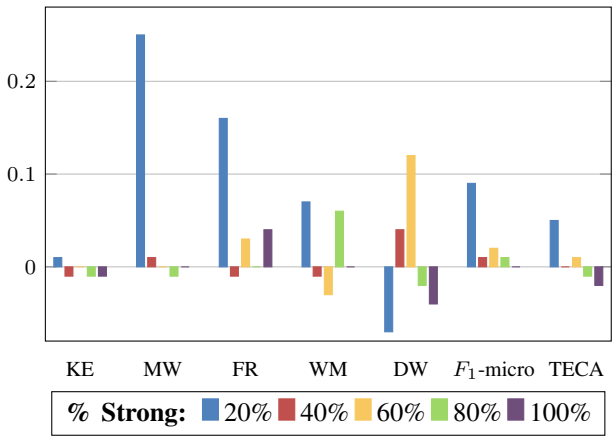
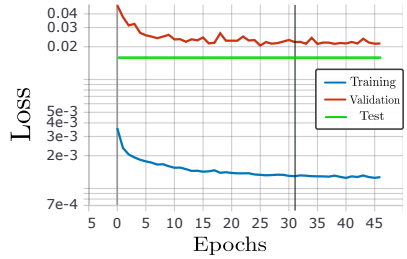
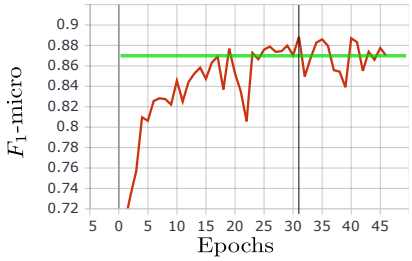


Fig. 3. Difference between F_1-scores of each appliance, F_1-micro, and TECA of the proposed method and S-CRNN for UK-DALE for the different percentages of strongly labeled data.



(a) Training, validation, and test losses.



(b) Validation and test F_1-micro.

Fig. 4. Training loss and validation loss and F_1-score for the experiment related to 40% strong data and 100% weak data for UK-DALE. Vertical bar indicates the early stopping epoch.

mance is lower. The same holds for Washing Machine where the performance improves with 20% and 80% of strongly labeled data, while in the other cases weak supervision does not improve the classification ability. A possible explanation for this behavior can be related to the shape of the activations of these appliances, which are more complex compared to the others, as also reported in the previous literature [29].

Observing the results of the individual appliances, for Kettle and Microwave weak labels allow to use a less amount of strong labels for obtaining the same F_1-score. For the Fridge, on the other hand, weak labels provide the overall best performance when 100% of strongly labeled data is used.

Fig. 4 shows an example of the loss trend for training, validation, and test, the F_1-micro trend during training, and the final value on the test set. Early stopping occurs on the 46th epoch.

2) REFIT: REFIT is a more challenging dataset than UK-DALE as it is significantly noisier [55]. Indeed, the results shown in Table VII obtained by using only weakly labeled data are lower compared to the ones obtained with UK-DALE leading to the conclusion that weakly labeled data only are not sufficient to achieve satisfactory performance.

Table IX reports the results with a fixed amount of weakly labeled data and varying percentages of strongly labeled data. Observing the F_1-micro for the different percentages and the average value, the proposed method achieves superior performance compared to benchmark methods, with the only exception of 60% of strongly labeled data where S-CRNN performs the same. The best F_1-micro is reached with the proposed method when the percentage of strongly labeled data is 40%. In terms of TECA, on average, S-CRNN and the proposed method achieve similar results, with the former obtaining a value 0.01 greater. The performance of the appliances with the highest average power consumption in an

TABLE IX
RESULTS OBTAINED ON THE REFIT DATASET RELATED TO EXPERIMENT 1. BEST SCORES FOR EACH STRONG PERCENTAGE ARE HIGHLIGHTED IN BOLD. BEST SCORE AMONG ALL THE PERCENTAGE ARE UNDERLINED.

% Strong	Method	KE	MW	FR	WM	DW	F_1-micro	TECA
20	LSTM [16]	**0.86**	0.53	0.23	0.46	0.67	0.51	**0.74**
	SSML-TCN [25]	0.72	0.71	0.12	0.59	0.51	0.42	0.58
	S-CRNN	0.68	0.40	0.29	**0.68**	**0.68**	0.44	0.65
	Proposed	0.68	**0.80**	**0.50**	0.54	0.58	**0.59**	0.65
40	LSTM [16]	0.84	0.77	0.25	0.27	0.70	0.54	0.76
	SSML-TCN [25]	0.81	0.71	0.08	0.52	0.47	0.39	0.65
	S-CRNN	**0.85**	**0.83**	0.28	**0.72**	0.76	0.62	**0.81**
	Proposed	0.74	0.80	**0.45**	0.59	**0.85**	**0.63**	0.76
60	LSTM [16]	0.67	0.80	0.31	0.48	0.46	0.51	0.71
	SSML-TCN [25]	0.74	0.70	0.10	0.61	0.48	0.36	0.63
	S-CRNN	**0.81**	**0.86**	0.35	**0.72**	0.63	**0.62**	**0.79**
	Proposed	0.78	0.82	**0.40**	0.54	**0.82**	**0.62**	0.77
80	LSTM [16]	0.70	0.77	**0.43**	0.59	0.69	0.58	0.73
	SSML-TCN [25]	**0.80**	0.74	0.08	0.63	0.55	0.43	0.70
	S-CRNN	0.76	0.75	0.32	**0.76**	0.81	0.60	**0.77**
	Proposed	0.58	**0.79**	0.41	**0.76**	**0.89**	**0.61**	0.74
100	LSTM [16]	0.57	0.80	**0.43**	0.53	0.31	0.51	0.67
	SSML-TCN [25]	**0.77**	0.73	0.11	0.60	0.52	0.42	0.68
	S-CRNN	0.64	0.80	0.33	0.65	0.65	0.55	0.71
	Proposed	0.73	**0.84**	0.36	**0.77**	**0.78**	**0.62**	**0.78**
AVG.	LSTM [16]	0.73	0.73	0.33	0.47	0.57	0.53	0.72
	SSML-TCN [25]	**0.77**	0.72	0.10	0.59	0.51	0.40	0.65
	S-CRNN	0.75	0.73	0.31	**0.71**	0.71	0.57	**0.75**
	Proposed	0.70	**0.81**	**0.42**	0.64	**0.78**	**0.61**	0.74

activation and the composition of the test set influence the behavior for the different percentages of strongly labeled data. As shown in Table II, the Kettle is the appliance with the highest power consumption, and with 20%, 40%, and 60% of strongly labeled data, the method with the greatest F_1-score on the Kettle also achieves the highest TECA. When the percentage of strongly labeled data is 80% and 100%, SSML-TCN achieves the highest F_1-score, but the overall F_1-micro is significantly lower than the proposed method and S-CRNN, and the value of TECA is consequently lower. However, it is worth remarking that the proposed method achieves an average TECA close to the one of the S-CRNN, while providing a higher F_1-micro..

Regarding individual appliances, in terms of average F_1-scores, the proposed method achieves the greatest performance for Microwave, Fridge, and Dishwasher, while SSML-TCN for Kettle and S-CRNN for Washing Machine. The best F_1-scores across all the percentages (underlined results in Table IX) are obtained by using the proposed method for Washing Machine, Dishwasher, and Fridge, while with LSTM for the Kettle and S-CRNN for the Microwave.

Fig. 5 shows the difference between the F_1-scores of each appliance, the F_1-micro, and the TECA of the proposed method and S-CRNN. We focus on S-CRNN as with UK-DALE for the same reasons, i.e., since it is the best performing among benchmark methods, and it allows to highlight the contributions of weak labels. Microwave, Fridge and Dishwasher are the appliances that benefit most from weak labels during training, in particular when strong data are only 20%. On the other hand, Kettle and Washing Machine exhibit the greatest benefit from weak labels when the amount of strongly labeled data is large.

B. Experiment 2: Fixed amount of strongly labeled data

As aforementioned, in this experiment the amount of strongly labeled data is fixed and the amount of weakly labeled data varies. Both for UK-DALE and REFIT, the percentage

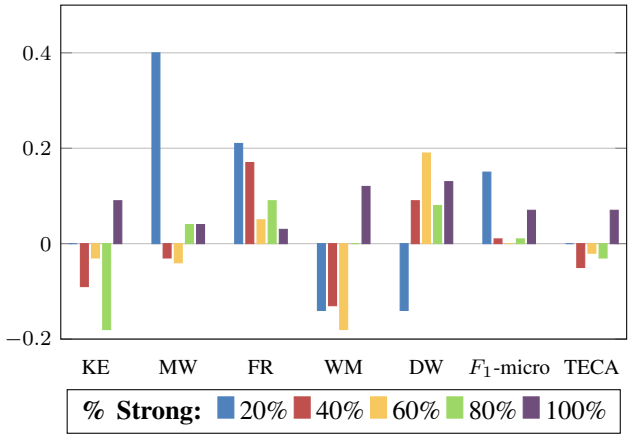


Fig. 5. Difference between F_1-scores of each appliance, F_1-micro, and TECA of the proposed method and S-CRNN for REFIT for the different percentages of strongly labeled data.

TABLE X
RESULTS OBTAINED ON THE UK-DALE DATASET RELATED TO EXPERIMENT 2. THE BEST RESULTS OBTAINED USING THE LEAST AMOUNT OF WEAKLY LABELED DATA ARE HIGHLIGHTED IN BOLD.

% Weak	Method	KE	MW	FR	WM	DW	F_1-micro	TECA
0	S-CRNN	0.98	0.67	0.42	0.80	0.81	0.72	0.86
20	SSML-TCN	0.85	0.71	0.19	0.46	0.65	0.54	0.68
	Proposed	0.98	**0.93**	**0.58**	0.79	**0.84**	**0.81**	**0.91**
40	SSML-TCN	0.94	0.64	0.21	0.54	0.66	0.56	0.76
	Proposed	**0.99**	0.92	0.51	0.81	0.69	0.77	0.89
60	SSML-TCN	0.89	0.66	0.21	0.51	0.68	0.56	0.73
	Proposed	0.99	0.92	0.58	0.82	0.82	0.81	0.91
80	SSML-TCN	0.83	0.73	0.18	0.39	0.63	0.50	0.63
	Proposed	0.98	0.92	0.53	0.83	0.81	0.80	0.91
100	SSML-TCN	0.82	0.70	0.16	0.39	0.60	0.46	0.60
	Proposed	0.99	0.92	0.58	**0.87**	0.74	0.81	0.91
AVG.	SSML-TCN	0.87	0.69	0.19	0.46	0.64	0.52	0.68
	Proposed	**0.99**	**0.92**	**0.56**	**0.82**	**0.78**	**0.80**	**0.91**

of strongly labeled data is fixed to 20%, the lowest value considered in Experiment 1. In this experiment, the objective is to evaluate to what extent weakly labeled data influence the performance when the amount of strongly labeled data is modest. For each percentage, we trained the proposed method and SSML-TCN since the other benchmark methods use only strongly labeled data for training and the training set does not change. For the sake of conciseness, in Table X and Table XI, we report only the results of the proposed method, SSML-TCN, and S-CRNN since it is the method that achieved the best average performance in Experiment 1.

1) UK-DALE: Table X presents the results related to the UK-DALE dataset. Observing the results, in terms of F_1-micro, introducing 20% of weak labels allows achieving the highest performance. Indeed, introducing more weak data does not provide significant improvements in that sense. In terms of TECA, the greatest value is obtained by using 20%, 60%, 80%, and 100% of weakly labeled data. Compared to S-CRNN and SSML-TCN, the proposed method always achieves greater F_1-micro and TECA.

Regarding individual appliances, the greatest average F_1-micro is always achieved by using the proposed method. The highest F_1-scores for most appliances are obtained with the lower percentages of weak data (20% and 40%). The F_1-score of Kettle and Microwave is almost independent of the number of weak labels since it changes only by 0.01. Instead, the F_1-score of the Washing Machine improves constantly with the increase of weakly labeled data used. The Dishwasher exhibits a significant improvement by using 20% of weak data, then the behavior is less consistent. A possible explanation is that the performance is more influenced by the composition of the weak dataset and the related unbalance of the classes.

In fact, the Dishwasher is significantly unbalanced considering weak annotations with a presence of 0.89%, with respect to the total presences of all the appliances in the dataset when weakly annotated data considered are 40%. In fact, for 20% the presence is about 1.4%, for 60% is 3.4%, for 80% is 9.6% and for 100% is 16.9%.

TABLE XI
RESULTS OBTAINED ON THE REFIT DATASET RELATED TO EXPERIMENT 2. THE BEST RESULTS OBTAINED USING THE LEAST AMOUNT OF WEAKLY LABELED DATA ARE HIGHLIGHTED IN BOLD.

% Weak	Method	KE	MW	FR	WM	DW	F_1-micro	TECA
0	S-CRNN	0.68	0.40	0.29	0.68	0.68	0.44	0.65
20	SSML-TCN	**0.74**	0.74	0.06	0.54	0.32	0.36	0.54
	Proposed	0.72	0.70	0.38	**0.77**	0.69	0.58	**0.73**
40	SSML-TCN	0.73	0.72	0.10	0.60	0.30	0.37	0.49
	Proposed	0.66	**0.85**	0.36	0.77	0.66	**0.59**	0.73
60	SSML-TCN	0.72	0.69	0.09	0.53	0.30	0.34	0.49
	Proposed	0.67	0.74	0.28	0.68	**0.74**	0.54	0.70
80	SSML-TCN	0.72	0.71	0.07	0.49	0.45	0.37	0.58
	Proposed	0.60	0.81	0.39	0.69	0.70	0.58	0.68
100	SSML-TCN	0.72	0.71	0.12	0.59	0.51	0.42	0.58
	Proposed	0.68	0.80	**0.50**	0.54	0.58	0.59	0.65
AVG.	SSML-TCN	**0.73**	0.71	0.09	0.55	0.38	0.37	0.54
	Proposed	0.67	**0.78**	**0.38**	**0.69**	**0.67**	**0.58**	**0.70**

2) REFIT: Table XI reports the results related to the REFIT dataset. Generally, the F_1-micro related to the proposed method for different percentages of weakly labeled data does not change significantly, apart for 60%. Regardless the percentage, the proposed method always outperforms SSML-TCN and S-CRNN in terms of F_1-micro and the highest value is obtained for 40% of weakly labeled data. In terms of TECA, the proposed method outperforms both S-CRNN and SSML-TCN, achieving the overall greatest value with 20% and 40% of weakly labeled data.

Regarding individual appliances, on average, the highest F_1-scores are achieved by using the proposed method with the only exception of the Kettle. For the different weakly labeled data percentages, the F_1-scores behaves differently depending on the appliance, but generally highest scores occur for lower percentages (20%-40%). This applies to the Kettle, Washing Machine and Microwave, while for the Dishwasher the best F_1-score is obtained when the percentage is 60% and for the Fridge when it is 100%. For the Microwave, the F_1-score is always higher than the one of the S-CRNN method. SSML-TCN achieves the highest F_1-score for the Kettle. However, the proposed method classifies the Kettle better than the S-CRNN when the weak data are modest (20%).

C. Experiment 3: Mixed training set

In this experiment, we evaluate whether mixing weakly labeled data of REFIT and strongly labeled data of UK-DALE during training improves the performance compared to S-CRNN on the test sets of both datasets. Among the

TABLE XII
RESULTS OBTAINED ON THE UK-DALE TEST SET WITH MIXED TRAINING SET. BEST SCORES ARE REPORTED IN BOLD.

	KE	MW	FR	WM	DW	F_1-micro	TECA
S-CRNN	**0.98**	0.67	**0.42**	**0.80**	0.81	0.72	0.86
Proposed (Mixed)	0.96	**0.75**	0.34	0.79	**0.88**	**0.75**	**0.88**

TABLE XIII
RESULTS OBTAINED ON REFIT TEST SET WITH MIXED TRAINING SET. BEST SCORES ARE REPORTED IN BOLD.

	KE	MW	FR	WM	DW	F_1-micro	TECA
S-CRNN	0.68	0.40	0.29	**0.68**	0.68	0.44	0.65
Proposed (Mixed)	**0.78**	**0.45**	0.21	0.43	**0.74**	**0.47**	**0.68**

benchmark approaches, we chose S-CRNN since it is the best performing, and it allows us to highlight the contribution of weakly labeled data since its architecture is similar to that of the proposed method. The percentage of UK-DALE strongly labeled training set is 20%.

As shown in Table XII for the UK-DALE dataset, the proposed network trained on mixed datasets improves both F_1-micro and TECA with respect to supervised learning. In particular, for Microwave and Dishwasher, the improvement is consistent, while for Kettle, Fridge, and Washing Machine, the performance slightly deteriorates.

On the REFIT test set, F_1-micro improves by 6.8% when the mixed training set is used compared to when training is performed only on strongly labeled REFIT data (Table XIII). TECA is also higher for the proposed method, with a 4.6% improvement over S-CRNN. Note, however, that the F_1-score of all appliances increases, and the only exceptions are the Fridge and the Washing Machine. This result is coherent to what was reported in [29], where Washing Machine was the only appliance with lower performance when training and testing were performed on different datasets. Moreover, consider also that in our case, we used only the UK-DALE validation set (Table I) for early stopping and hyperparameters optimization, and Washing Machine is the appliance having the largest quantity of strong labels compared to the others. Nonetheless, this result evidences how a modest quantity of strong data with weak annotations can positively enhance classification on unseen data for most appliances.

D. Discussion

The results evidenced that the proposed method provides an overall positive contribution compared to benchmark methods and full supervision alone. The first two experiments highlighted that the highest average scores have been obtained with weakly labeled data and that for certain appliances, it is possible to obtain the same performance with a lower amount of strongly labeled data (Kettle and Fridge for UK-DALE and Fridge for REFIT). In particular, the appliances that mostly benefit from weak labels are Microwave, Fridge and Dishwasher for both datasets. Washing Machine scores improve with weak data for UK-DALE but not for REFIT. The same holds for Kettle. Moreover, the benefits of the proposed approach are most evident when the percentage of strong annotations is modest while, depending on the dataset composition, the number of weakly labeled data can influence differently the classification. In the view of a practical application, the results obtained in Experiment 3 evidenced that mixing two datasets improves the performance both when the test set is from the same domain of strongly labeled data of the training set (UK-DALE) and when the test set is from the same domain of weakly labeled data of the training set (REFIT). The latter result is particularly significant since it implies that acquiring weakly labeled data from a target environment and mixing it with strongly labeled data from a public dataset provides a significant performance improvement.

The results also evidenced that depending on the appliances and the dataset the behavior is not always consistent, and the contribution of weak labels varies for the different percentages of strong and weak annotations. It is worth noting that such behavior affects also benchmark methods, suggesting that it may be a critical aspect of neural networks-based multi-label appliance classification methods. We consider this as an open problem that requires further studies and specific works.

In summary, the obtained results evidenced that generally the proposed method is particularly advantageous when the number of weak labels exceeds that of strongly annotated bags. On the other hand, the advantage reduces when the two amounts are comparable. Thus, when using the presented architecture, a possible strategy is to augment a dataset annotated only with strong labels with a large amount of weakly labeled data, as it is easier to collect and yield better performance.

VI. CONCLUSION

This work presented a multi-label appliance classification method based on a deep neural network and weakly labeled data. The task has been formulated as a MIL problem, and a CRNN with the related learning strategy for exploiting both weak and strong labels has been presented.

In the experiments, we evaluated if weak information is indeed able to provide a performance advantage while requiring less labeling effort. The experiments on the UK-DALE dataset conducted in different training conditions showed that weakly labeled data improve the performance in terms of F_1-micro and TECA, particularly when the amount of strongly labeled data is modest. Moreover, combining the strongly labeled UK-DALE training set and the weakly labeled REFIT training set proved advantageous in the respective test sets, demonstrating the effectiveness of adding weakly labeled data from a different dataset to the training set.

Future works will extend the potentiality of weak supervision to transfer learning methods and other tasks related to NILM. More in detail, the proposed method can be employed for estimating active power profiles of individual appliances by using real-valued weak labels instead of categorical annotations. Although using weak labels can prevent annotation errors, an investigation on wrongly annotated data in presence of weak labels can be useful in a practical real-world scenario. Critical aspects that emerged here, such as the performance for the different appliances and training set compositions will be

further investigated. Moreover, aspects specific to the neural network operation and learning, such as the pooling function and the contribution of the weak and loss functions, will be further studied. Finally, the deployment of the proposed method on a embedded platform and the related evaluation on a real application scenario will be considered.

ACKNOWLEDGMENT

This work was supported by the Marche Region in implementation of the financial programme POR MARCHE FESR 2014-2020, project "Miracle" (Marche Innovation and Research fAcilities for Connected and sustainable Living Environments), CUP B28I19000330007.

REFERENCES

[1] J. Rogelj, M. Den Elzen, N. Höhne, T. Fransen, H. Fekete, H. Winkler, R. Schaeffer, F. Sha, K. Riahi, and M. Meinshausen, "Paris agreement climate proposals need a boost to keep warming well below 2 °C," *Nature*, vol. 534, no. 7609, pp. 631–639, 2016.

[2] Eurostat, "Electricity and heat statistics: Consumption of electricity and derived heat," 2021. [Online]. Available: https://ec.europa.eu/eurostat/statistics-explained

[3] K. C. Armel, A. Gupta, G. Shrimali, and A. Albert, "Is disaggregation the holy grail of energy efficiency? The case of electricity," *Energy Policy*, vol. 52, pp. 213–234, 2013.

[4] G. Hart, "Nonintrusive appliance load monitoring," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870–1891, 1992.

[5] C. Zhang, M. Zhong, Z. Wang, N. Goddard, and C. Sutton, "Sequence-to-point learning with neural networks for non-intrusive load monitoring," in *Proc. of the 32nd AAAI Conf. on Artificial Intelligence*, New Orleans, LA, USA, Feb. 2-7 2018, pp. 2604–2611.

[6] S. M. Tabatabaei, S. Dick, and W. Xu, "Toward non-intrusive load monitoring via multi-label classification," *IEEE Trans. on Smart Grid*, vol. 8, no. 1, pp. 26–40, 2017.

[7] B. Zhao, K. He, L. Stankovic, and V. Stankovic, "Improving event-based non-intrusive load monitoring using graph signal processing," *IEEE Access*, vol. 6, pp. 53 944–53 959, 2018.

[8] A. Moradzadeh, O. Sadeghian, K. Pourhossein, B. Mohammadi-Ivatloo, and A. Anvari-Moghaddam, "Improving residential load disaggregation for sustainable development of energy via principal component analysis," *Sustainability*, vol. 12, no. 8, 2020.

[9] M. Figueiredo, B. Ribeiro, and A. de Almeida, "Electrical signal source separation via nonnegative tensor factorization using on site measurements in a smart home," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 2, pp. 364–373, 2014.

[10] A. Rahimpour, H. Qi, D. Fugate, and T. Kuruganti, "Non-intrusive energy disaggregation using non-negative matrix factorization with sum-to-k constraint," *IEEE Trans. Power Syst.*, vol. 32, no. 6, p. 4430–4441, Nov 2017.

[11] J. Kolter, S. Batra, and A. Ng, "Energy disaggregation via discriminative sparse coding," in *Advances in Neural Information Processing Systems*, J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds., vol. 23. Curran Associates, Inc., 2010. [Online]. Available: https://proceedings.neurips.cc/paper/2010/file/7810ccd41bf26faaa2c4e1f20db70a71-Paper.pdf

[12] S. Singh and A. Majumdar, "Deep sparse coding for non–intrusive load monitoring," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 4669–4678, 2018.

[13] M. Zhong, N. Goddard, and C. Sutton, "Signal aggregate constraints in additive factorial hmms, with application to energy disaggregation," in *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[14] R. Bonfigli, E. Principi, M. Fagiani, M. Severini, S. Squartini, and F. Piazza, "Non-intrusive load monitoring by using active and reactive power in additive factorial hidden markov models," *Applied Energy*, vol. 208, pp. 1590–1607, 2017.

[15] K. T. Chui, M. D. Lytras, and A. Visvizi, "Energy sustainability in smart cities: Artificial intelligence, smart monitoring, and optimization of energy consumption," *Energies*, vol. 11, no. 11, 2018.

[16] J. Kelly and W. Knottenbelt, "Neural NILM: Deep Neural Networks Applied to Energy Disaggregation," in *Proc. of the 2nd ACM Int. Conf. on Embedded Systems for Energy-Efficient Built Environments*, New York, USA, Nov. 4–5 2015, pp. 55–64.

[17] V. Piccialli and A. M. Sudoso, "Improving Non-Intrusive Load Disaggregation through an Attention-Based Deep Neural Network," *Energies*, vol. 14, no. 4, p. 847, 2021.

[18] M. Kaselimi, E. Protopapadakis, A. Voulodimos, N. D. Doulamis, and A. D. Doulamis, "Multi-channel recurrent convolutional neural networks for energy disaggregation," *IEEE Access*, vol. 7, pp. 81 047–81 056, 2019.

[19] L. Massidda, M. Marrocu, and S. Manca, "Non-intrusive load disaggregation by convolutional neural network and multilabel classification," *Applied Sciences*, vol. 10, no. 4, 2020.

[20] M. Xia, K. Wang, X. Zhang, Y. Xu *et al.*, "Non-intrusive load disaggregation based on deep dilated residual network," *Electric Power Systems Research*, vol. 170, pp. 277–285, 2019.

[21] A. Langevin, M.-A. Carbonneau, M. Cheriet, and G. Gagnon, "Energy disaggregation using variational autoencoders," *Energy and Buildings*, vol. 254, p. 111623, 2022.

[22] M. Kaselimi, N. Doulamis, A. Voulodimos, A. Doulamis, and E. Protopapadakis, "EnerGAN++: A Generative Adversarial Gated Recurrent Network for Robust Energy Disaggregation," *IEEE Open Journal of Signal Processing*, vol. 2, p. 1, 2021.

[23] Y. Pan, K. Liu, Z. Shen, X. Cai, and Z. Jia, "Sequence-to-subsequence learning with conditional GAN for power disaggregation," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, Barcellona, Spain, May 4-8 2020, pp. 3202–3206.

[24] D. Murray, L. Stankovic, V. Stankovic, S. Lulic, and S. Sladojevic, "Transferability of Neural Network Approaches for Low-rate Energy Disaggregation," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Process.*, Brighton, UK, May 12-17 2019, pp. 8330–8334.

[25] Y. Yang, J. Zhong, W. Li, T. A. Gulliver, and S. Li, "Semisupervised Multilabel Deep Learning Based Nonintrusive Load Monitoring in Smart Grids," *IEEE Trans. Ind. Informat.*, vol. 16, no. 11, pp. 6892–6902, 2020.

[26] N. Miao, S. Zhao, Q. Shi, and R. Zhang, "Non-Intrusive Load Disaggregation Using Semi-Supervised Learning Method," in *2019 Int. Conf. on Security, Pattern Analysis, and Cybernetics (SPAC)*, Guangzhou Shi, China, Dec. 20-23 2019, pp. 17–22.

[27] A. Faustine, L. Pereira, H. Bousbiat, and S. Kulkarni, "Unet-nilm: A deep neural network for multi-tasks appliances state detection and power estimation in nilm," in *Proc. of the 5th Int. Workshop on Non-Intrusive Load Monitoring*, New York, USA, Nov. 18 2020, p. 84–88.

[28] S. Verma, S. Singh, and A. Majumdar, "Multi-label lstm autoencoder for non-intrusive appliance load monitoring," *Electric Power Systems Research*, vol. 199, p. 107414, 2021.

[29] M. D'Incecco, S. Squartini, and M. Zhong, "Transfer learning for non-intrusive load monitoring," *IEEE Trans. on Smart Grid*, vol. 11, pp. 1419–1429, 2019.

[30] P. Huber, A. Calatroni, A. Rumsch, and A. Paice, "Review on Deep Neural Networks Applied to Low-Frequency NILM," *Energies*, vol. 14, no. 9, 2021.

[31] Z. H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, 2018.

[32] D. Pathak, P. Krahenbuhl, and T. Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Santiago, Chile, Dec. 11-28 2015, pp. 1796–1804.

[33] Y. Wang, J. Li, and F. Metze, "A Comparison of Five Multiple Instance Learning Pooling Functions for Sound Event Detection with Weak Labeling," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Brighton, UK, May 12-17 2019, pp. 31–35.

[34] A. Kumar and B. Raj, "Audio event detection using weakly labeled data," in *Proc. of the 24th ACM Int. Conf. on Multimedia*, Amsterdam, Netherlands, Oct. 15-19 2016.

[35] N. Pappas and A. Popescu-Belis, "Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis," in *Proc. of the Int. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, Oct. 25-29 2014, pp. 455–466.

[36] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection," *IEEE/ACM Trans. on Audio Speech and Lang. Process.*, vol. 25, no. 6, pp. 1291–1303, 2017.

[37] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1, pp. 31–71, 1997.

[38] H. Dinkel, X. Cai, Z. Yan, Y. Wang, J. Zhang, and Y. Wang, "The smallrice submission to the dcase2021 task 4 challenge: A lightweight approach for semi-supervised sound event detection with unsupervised data augmentation," DCASE2021 Challenge, Tech. Rep., June 2021.

[39] J. Kelly and W. Knottenbelt, "The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes," *Scientific Data*, vol. 2, no. 150007, 2015.
[40] D. Murray, L. Stankovic, and V. Stankovic, "An electrical load measurements dataset of United Kingdom households from a two-year longitudinal study," *Scientific Data*, vol. 4, no. 1, p. 160122, 2017.
[41] A. Faustine and L. Pereira, "Multi-label learning for appliance recognition in nilm using fryze-current decomposition and convolutional neural network," *Energies*, vol. 13, no. 16, 2020.
[42] O. Maron and T. Lozano-Pérez, "A Framework for Multiple-Instance Learning," in *Advances in Neural Information Processing Systems*, vol. 10, 1998.
[43] M. A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329–353, 2018.
[44] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li, "Multi-instance multi-label learning," *Artificial Intelligence*, vol. 176, no. 1, pp. 2291–2320, 2012.
[45] J. R. Foulds and E. Frank, "A review of multi-instance learning assumptions," *The Knowledge Engineering Review*, vol. 25, pp. 1 – 25, 2010.
[46] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1724–1734.
[47] H. K. Iqbal, F. H. Malik, A. Muhammad, M. A. Qureshi, M. N. Abbasi, and A. R. Chishti, "A critical review of state-of-the-art non-intrusive load monitoring datasets," *Electric Power Systems Research*, vol. 192, p. 106921, 2021.
[48] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. of the 3rd Int. Conf. on Learning Representations (ICLR)*, San Diego, CA, USA, May 7-9 2015, pp. 2449–2457.
[49] L. Wang, S. Mao, B. Wilamowski, and R. M. Nelms, "Pre-trained Models for Non-intrusive Appliance Load Monitoring," *IEEE Transactions on Green Communications and Networking*, vol. 2400, pp. 1–1, 2021.
[50] N. Batra, R. Kukunuri, A. Pandey, R. Malakar, R. Kumar, O. Krystalakos, M. Zhong, P. Meira, and O. Parson, "Towards reproducible state-of-the-art energy disaggregation," *Proc. of the 6th ACM Int. Conf. on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pp. 193–202, Nov. 13-14 2019.
[51] L. Massidda, M. Marrocu, and S. Manca, "Non-intrusive load disaggregation by convolutional neural network and multilabel classification," *Applied Sciences*, vol. 10, no. 4, 2020.
[52] J. Z. Kolter and M. J. Johnson, "REDD: A public data set for energy disaggregation research," in *Proc. of Workshop on Data Mining Applications in Sustainability (SIGKDD)*, San Diego, CA, USA, 2011, pp. 59–62.
[53] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A novel bandit-based approach to hyperparameter optimization," *J. Mach. Learn. Res.*, vol. 18, no. 1, p. 6765–6816, 2017.
[54] "NVIDIA DGX Station A100," nvidia.com, (accessed Apr. 15, 2022). [Online]. Available: https://www.nvidia.com/en-us/data-center/dgx-station-a100/
[55] S. Makonin and F. Popowich, "Nonintrusive load monitoring (NILM) performance evaluation: A unified approach for accuracy reporting," *Energy Efficiency*, vol. 8, no. 4, pp. 809–814, 2015.

Giulia Tanoni was born in Recanati (MC), Italy, in 1994. She received the bachelor's degree in biomedical engineering and the M.S. degree in biomedical engineering with honors from Università Politecnica delle Marche, Ancona, in 2018 and 2020, respectively. She is currently working towards the Ph.D. degree in edge-centric computing, focusing on deep learning algorithms for smart living, with the Department of Information Engineering (DII) of Università Politecnica delle Marche.

Emanuele Principi (Member, IEEE) was born in Senigallia (AN), Italy, in 1978. He got the Italian Laurea with honors in electronic engineering from Università Politecnica delle Marche, Italy, in 2004, and he obtained his Ph.D. at the same university in 2009. He is now a tenure track Assistant Professor of Electrical Engineering from December 2019 at the Department of Information Engineering of Università Politecnica delle Marche. His current research interests are in the area of digital signal processing and computational intelligence, with a special focus on smart grids, and audio processing. He is author and co-author of many international scientific peer-reviewed articles, and he is an Associate Editor of Neural Computing and Applications and Artificial Intelligence Review both edited by Springer from 2017. Dr. Principi joined the organizing and technical committees of several international conferences. Dr. Principi is member and secretary of the Adriatic section of the Italian Association of Electrotechnics, Electronics, Automation, Computer Science and Telecommunications. As of January 2021, Dr. Principi is the chair of the IEEE CIS Task Force on Computational Audio Processing.

Stefano Squartini (Senior Member, IEEE) was born in Ancona, Italy, on March 1976. He received the Italian Laurea with honors in electronic engineering and the Ph.D. degree in electronics and telecommunications from the Polytechnic University of Marche (UnivPM), Ancona, Italy, in 2002 and 2005, respectively. He joined the Department of Information Engineering as an Assistant Professor in circuit theory in 2007. He has been a Full Professor with UnivPM since 2020. He is author or coauthor of more than 210 international scientific papers. His current research interests include the area of computational intelligence and digital signal processing, with special focus on audio processing and energy management. Prof. Stefano is an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CYBERNETICS, and IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE. He joined the Organizing and the Technical Program Committees of more than 80 International Conferences and Workshops.