Journal of Ambient Intelligence and Humanized Computing (2023) 14:13049–13065
https://doi.org/10.1007/s12652-022-03768-2

ORIGINAL RESEARCH

An active semi-supervised deep learning model for human activity recognition

Haixia Bi[1,2] · **Miquel Perello-Nieto**[2] · **Raul Santos-Rodriguez**[2] · **Peter Flach**[2] · **Ian Craddock**[3]

Received: 3 August 2021 / Accepted: 9 February 2022 / Published online: 20 March 2022

Abstract

Human activity recognition (HAR), which aims at inferring the behavioral patterns of people, is a fundamental research problem in digital health and ambient intelligence. The application of machine learning methods in HAR has been investigated vigorously in recent years. However, there are still a number of challenges confronting the task, where one significant barrier lies in the longstanding shortage of annotations. To address this issue, we establish a new paradigm for HAR, which integrates active learning and semi-supervised learning into one framework. The main idea is to reduce the annotation cost by actively selecting the most informative samples for annotation, as well as leveraging the unlabelled instances in a semi-supervised way. In particular, we propose to utilize the massive unlabelled data via temporal ensembling of convolutional neural networks (CNN), which yields robust consensus predictions by aggregating the outputs of the training networks on different epochs. We conducted extensive experiments on three public benchmark datasets. The proposed method achieves Macro F1 values of 0.76, 0.45 and 0.91 in a low annotation scenario on PAMAP2, USCHAD and UCIHAR datasets respectively, outperforming a multitude of state-of-the-art deep models. The ablation study proves the effectiveness of the two components of the framework, i.e., active learning-based sample selection and semi-supervised model training with temporal ensembling, in alleviating the issue of insufficient labels. Cross-validation and statistical significance experiments further demonstrate the robustness and generalization ability of the proposed method. The source codes are available at https://github.com/HaixiaBi1982/ActSemiCNNAct.

Keywords Active learning · Convolutional neural network · Human activity recognition · Semi-supervised learning · Temporal ensembling

List of symbols

$\mathbf{X}$	All samples in the training set
N	Number of instances in $\mathbf{X}$
L	Annotation budget
S_s	Labelled sample set
l	Number of labelled samples
C	Number of of activity classes
Θ	Deep network parameters
$\mathcal{U}$	Candidate sample pool
$Loss$	Total loss function
$Loss_s$	Supervised loss term
$Loss_u$	Unsupervised loss term
n	Sample number chosen in active learning iteration

Abbreviations

HAR	Human activity recognition
CNN	Convolutional neural networks
SSL	Semi-supervised learning
RNN	Recurrent neural networks

✉ Peter Flach
peter.flach@bristol.ac.uk

Haixia Bi
haixia.bi@xjtu.edu.cn

Miquel Perello-Nieto
miquel.perellonieto@bristol.ac.uk

Raul Santos-Rodriguez
enrsr@bristol.ac.uk

Ian Craddock
ian.craddock@bristol.ac.uk

[1] School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an 710049, China

[2] Intelligent Systems Laboratory, University of Bristol, Bristol BS8 1UB, UK

[3] Department of Electrical and Electronic Engineering, University of Bristol, Bristol, UK

LSTM	Long short-term memory
BvSB	Best-versus-second-best
SGD	Stochastic gradient descent
CV	Cross-validation
LOSO-CV	Leave-one-subject-out CV
ActCNN	Active CNN
ActSemiCNN	Active semi-supervised CNN
DeepConvLSTM	Deep convolutional LSTM
CoTrCNN	Co-training-based CNN
ConvAE	Convolutional auto-encoder
MLP	Multi-layer perceptron
MTSelfCNN	Multi-task self-supervised CNN
SemiConvAttn	Attention-based SSL Convolution
MeanTeacher	Mean teacher
DAL	Dynamic active learning

1 Introduction

Many countries are currently experiencing a rapidly ageing population, leading to an immense pressure for healthcare resources (Chen et al. 2011). An appealing solution to this issue is to expand the role of continuous healthcare monitoring to private homes, complementing and potentially reducing the need for hospital inpatient care and face-to-face interaction with health professionals (Noor et al. 2020). In this context, smart home technology emerged as a feasible approach to attain this goal (Amiribesheli et al. 2015). Especially, the ubiquity of wearable devices and smartphones makes monitoring data easily accessible (Bi et al. 2021). This creates an opportunity to leverage the massive sensor data to extract clinically relevant information. Automatic human activity recognition (HAR) is therefore an enabling step towards inferring the behavior of people, further facilitating decision making and necessary interventions from carers and health-care professionals (Diethe et al. 2017).

HAR has drawn extensive attention in recent years, and a corpus of machine learning approaches has been explored to address this problem, such as decision trees (Xu et al. 2019), support vector machines (Khan et al. 2014), and naive Bayesian networks (Gomes et al. 2012). In particular, the current state-of-the-art performance in HAR consists in using deep learning architectures like Convolutional neural networks (CNN) (Wan et al. 2020; Gao et al. 2021) and long short-term memory (LSTM) (Ullah et al. 2019; Singh et al. 2021). The main benefit of these methods is that they can automatically extract discriminative and data-driven features from the raw input data.

Despite this remarkable progress, a significant challenge confronting this task lies in the lack of annotations, which the aforementioned methods heavily rely on (Bi et al. 2021). In real-world activity recognition systems, annotating HAR data is not only labour-intensive and time-consuming, but also demands domain-specific knowledge and skills (Bi et al. 2021). Furthermore, the annotation of HAR data in real life (through unscripted experiments) has some privacy and ethical concerns that may limit the annotation process (Zeng et al. 2017). These factors result in a noticeable scarcity of labelled data. Under these circumstances, *how to achieve a favourable performance with limited annotations becomes a challenging task*.

Semi-supervised learning (SSL) and active learning are two compelling solutions to tackle this issue. Semi-supervised learning improves the model's performance and generalization ability by leveraging unlabelled data, while active learning enables to reduce the amount of annotations necessary by strategically choosing samples with maximal information and highest training utility (Bi et al. 2021). Owing to their lower dependence on annotations, the application of SSL and active learning on HAR has evoked increasing interests recently. Semi-supervision paradigms, such as self-training (Bota et al. 2019), co-training (Chen et al. 2020; Lv et al. 2018) and graph-based models (Han et al. 2019) have been successfully applied in HAR. Active learning effectively boosts the performance of HAR as well, where different active sample selection strategies are exploited during the active learning process, including uncertainty (Bi et al. 2021), diversity (Saito et al. 2015) and representativeness (Lughofer 2012) based schemes.

Given that both paradigms are effective in overcoming the hurdle of label scarcity, yet solve this problem from different perspectives, we propose to explore the combination of active learning and SSL for HAR task, in the hope of enhancing the labelling efficiency with the former and taking advantage of unlabelled data with the latter. Active semi-supervised learning has been studied in a few image related tasks (Rottmann et al. 2018; Zhang et al. 2019). However, to our knowledge, the integration of active learning in semi-supervised learning has never been investigated in HAR.

With the booming development of deep learning, applying deep models in SSL has aroused growing attention and shown remarkable improvements in performance (Zeng et al. 2017; Chen et al. 2020). Most of the current SSL methods work in an iterative manner. They usually employ the network of the last epoch in each iteration to predict the labels of unlabelled samples and use these predictions as training targets for the following iteration. However, if the predictions of the latest epoch are unreliable—which normally happens due to various degrees of randomness in deep neural networks—they will mislead the consequent model training, further worsening the final performance. Based on the findings that an ensemble of multiple neural networks generally generates more robust predictions than a single network (Srivastava et al. 2014a), we propose to incorporate a temporal ensemble (Laine and Aila 2016) to CNN via aggregating the history outputs of the network with dropout regularization during training for HAR.

Overall, this paper proposes a deep active semi-supervised approach for human activity recognition, in order to promote the recognition performance with reduced annotation cost. The main novelties and contributions of the proposed approach are threefold.

1. We design a novel deep HAR model incorporating active learning and semi-supervised learning into one framework, which improves the model's performance in a low annotation regime by selecting the most informative samples to be annotated and taking advantage of the information of massive unlabelled instances. To the best of our knowledge, this is the first work to combine these two techniques into one framework for HAR.
2. A novel unsupervised loss term is introduced for employing the temporal ensemble of the deep model subject to consistency regularization, effectively enabling semi-supervised learning in combination with the supervised loss component. This unsupervised loss term reduces the impact of prediction uncertainty, producing more accurate and stable predictions for activities.
3. We evaluate our proposed method, which we call *ActSemiCNN*, on three real benchmark datasets for activity recognition, i.e., PAMAP2, USCHAD and UCIHAR datasets. Extensive experiments were conducted to assess the proposed approach. The results demonstrate that *ActSemiCNN* achieves state-of-the-art recognition performance with significantly reduced annotation cost, and exhibits strong robustness and generalization ability.

The remainder of this paper is organized as follows. Section 2 briefly recalls related paradigms and methodologies. We detail the proposed deep active semi-supervised method in Sect. 3. Section 4 presents a comparative study applied to real benchmark datasets. We discuss the proposed method in Sect. 5. Finally, Sect. 6 concludes the paper.

2 Related work

In this section we review related works in human activity recognition, where the deep learning-based methods are described in Sect. 2.1, and the active learning and SSL-based methods are introduced in Sect. 2.2.

2.1 Deep learning-based activity recognition

In the past decade, most wearable sensor-based HAR methods involve feature engineering based on domain expertise (Bi et al. 2020). However, these methods are relatively limited as they rely on human creativity to come up with novel features and lack the power to capture underlying explanatory factors in low-level sensory inputs (Saeed et al. 2019; Merritt et al. 2018).

Most recently, the explosion of deep learning techniques has paved a new way for a broad spectrum of problems as they can automatically extract representative features. Kumar et al. (2021) studied deep models, diverse embedding representations and ensembling technique on co-morbidity recognition from clinical records. The research of deep learning frameworks in HAR have been studied in a number of works (Chen and Xue 2015; Ordóñez and Roggen 2016; Yao et al. 2017; Bianchi et al. 2019; Haresamudram et al. 2019; Wan et al. 2020). Chen and Xue (2015) firstly proposed a CNN-based HAR method to classify activities collected with acceleration sensors. A CNN was utilized to automatically learn discriminative features from the signal sequences of accelerometers and gyroscopes in Bianchi et al. (2019). Haresamudram et al. (2019) leveraged unsupervised convolutional auto-encoder to firstly extract feature representation and then used multi-layer perceptron (MLP) to tune the network. To reduce the cost of hardware facilities, Wan et al. (2020) designed a real-time CNN-based HAR method for local feature extraction from smartphone accelerometer data. In Gao et al. (2021), a new multi-branch CNN was introduced, which performs kernel selection among multiple branches by means of attention mechanism.

In addition, recurrent neural networks (RNN) show competitive results when applied to HAR tasks. RNN and their extensions, such as gated recurrent unit and long short-term memory (LSTM), have been applied for HAR in several recent publications. Ordóñez and Roggen (2016) combined LSTM and CNN to explicitly model the temporal dynamics of sequential data, achieving prominent performance in HAR from sensor data. Murad and Pyun (2017) proposed to use RNN for building recognition models that are capable of capturing long-range dependencies in variable-length input sequences. Convolutional and recurrent neural networks were integrated to exploit local interactions among similar mobile sensors and extract temporal relationships to model signal dynamics in Yao et al. (2017). Ullah et al. (2019) developed an end-to-end deep model which consists of a single layer neural network for data pre-processing and a stacked multi-layer LSTM network. Attention mechanism was further incorporated with LSTM in Singh et al. (2021), which not only captures the spatio-temporal features but also learns important time points.

The emerging formulation of self-supervised learning was applied in HAR in recent two years. Saeed et al. (2019) designed an auxiliary task of recognizing diverse transformations performed on the raw input features, which is implemented by training a multi-task CNN, yielding features with high generalization ability. Haresamudram et al. (2020) introduced masked reconstruction to HAR task as a viable self-supervised pre-training objective, and demonstrated improved performance over state-of-the-art semi-supervised learning methods. A contrastive predictive coding framework was developed (Haresamudram et al. 2021) to capture the underlying temporal structure of HAR data, leading to significantly improved recognition performance.

2.2 SSL and active learning-based HAR methods

Semi-supervised learning has been extensively used for sensor-based activity recognition. Semi-supervision based HAR methods can be categorized from different perspectives. Considering how the unlabelled samples are utilized, they can be classified into self-training, co-training and graph-based methods (Zhu et al. 2018). Lopes et al. (2012) is a self-training-based approach, which uses an ensemble of classifiers to alternatively select the unlabelled samples with the most confident predictions and assume their predicted labels are correct in order to assist further training of the classifiers. Co-training (Chen et al. 2020) selects confident samples from independent feature spaces for two classifiers first and then the selected samples with the estimated labels are added to the training set. Liu et al. (2021) developed an HAR approach based on graph convolutional networks, which encodes arbitrary graphs by automatically updating the structure information under manifold regularization.

According to how features are extracted and utilized, semi-supervised HAR methods can be classified into feature engineering based methods (Subramanya et al. 2012) and feature learning based methods. One typical feature engineering based method was designed in Subramanya et al. (2012), which introduces boosted decision stumps based discriminative method to select features. The advent of deep learning enables the boosting of deep SSL-based HAR methods. Zeng et al. (2017) utilized unlabelled data in both feature learning and model learning stages using CNN-Ladder architecture. An adversarial network with an auto-encoder and two discriminator networks represented by fully connected layers was developed in Balabka (2019), which relieves the heavy reliance on large labelled dataset. Chen et al. (2020) designed a co-training HAR framework integrating attention mechanism with recurrent convolutional models. The state-of-the-art mean teacher semi-supervised model was introduced to HAR in Narasimman et al. (2021), which averages model weights over training steps to produce more robust results.

Active learning is another promising solution to address the annotation scarcity issue. It aims to reduce the labelling cost by selecting the most informative samples for annotation. Uncertainty sampling is the most extensively used strategy in active learning. Entropy (Hossain et al. 2017), marginal sampling (Bi et al. 2021) and least confident (Alemdar et al. 2011) measures have been adopted to measure the informativeness of the instances in HAR. In Shahmohammadi et al. (2017), Query-by-committee was employed to identify the samples that are worth annotating.

Table 1 tabulates the most representative works in deep learning, SSL and active learning-based human activity recognition. Analyzing the above methods, we can find that: (1) The application of active learning and SSL in HAR has been intensively investigated, and demonstrated effective in reducing labelling costs. However, none of the above methods focus on activity recognition by jointly combining active learning and SSL. (2) The above analyzed deep SSL models mostly employ the predictions of the network on the unlabelled samples in every iteration, however ignore the model's uncertainty, which may mislead the model training with erroneous estimated labels if the current predictions are unreliable. For this application, our proposed method differs from the aforementioned HAR methods in two aspects. Firstly, rather than using active learning or SSL separately, our proposed method integrates the two components into a unified framework. Secondly, we combine temporal ensembling with CNNs, incorporating the history information of networks during training, which is expected to generate more reliable predictions on unlabelled samples.

3 Methodology

We first formulate the problem and overview the pipeline of the proposed method in Sect. 3.1. Next, we introduce the two components, i.e., active learning based sample selection in Sect. 3.2 and semi-supervised model training with temporal ensembling in Sect. 3.3.

3.1 Problem formulation and overview

The HAR task is formulated as a classification problem where data samples are a set of sensor data sequences gathered over a time interval and classes correspond to activities. We will first introduce the application scenario and processing flow of the proposed method, and then give an overview of the active semi-supervised deep model.

The proposed method is designed for a common scenario where data from a number of participants is available, however labels are scarce due to a limited annotation budget. Figure 1 illustrates the processing pipeline. Given the sequential data from a number of participants as input, we firstly split the whole dataset into three independent subsets—training, validation and test set—in a subject-wise fashion (one subject is always in only one of the partitions). Next, we randomly select a very small number of samples (less than the total annotation budget) from the training set and manually label them. Based on the labelled training set and validation set we then determine the optimal hyperparameters of

Table 1 Representative related works with different degrees of supervision on human activity recognition

Reference	Supervision	Architecture	Advantages	Potential future enhancements
Ordóñez and Roggen (2016)	Supervised	*DeepConvLSTM*: a deep model with 4 convolutional layers and 2 dense LSTM layers	Do not require expert knowledge in extracting features; Explicitly model temporal dynamics	To reduce the requirements on large amount of annotations
Haresamudram et al. (2019)	Supervised	*ConvAE*: Convolutional encoder and decoder with a bottleneck layer in between	Use unsupervised auto-encoder technique to learn feature representation	Transfer learning can be integrated to improve the performance and adaptive capability
Wan et al. (2020))	Supervised	CNN with 3 convolutional layers, 3 pooling layers and a fully connected layer	Smartphone inertial accelerometer based CNN for HAR	To validate the real-time character of the proposed method
Singh et al. (2021)	Supervised	*AttnConvLSTM*: An embedding layer with convolutional filters, an encoder with LSTM layers and a self-attention layer	Not only capture the spatio-temporal features but also learn important time points with attention mechanism	To try out different attention mechanisms including the global and local attention
Zeng et al. (2017)	Semi-supervised	*SemiCNN-Encoder*: A deep auto encoder trained on unlabelled data and a neural network on labelled data	The first work to leverage unlabelled data in deep neural networks in HAR applications	To extend the current framework with more effective techniques to utilize unlabelled data
Balabka (2019)	Semi-supervised	*SemiAAE*: Adversarial network with an auto-encoder and two discriminator networks represented by fully connected layers	Utilize adversarial technique to relieve the heavy reliance on large labelled dataset.	To solve the slow training issue of adversarial network and validate the method on more public datasets
Chen et al. (2020)	Semi-supervised	*SemiConvAttn*: Attention-based recurrent convolutional models under a co-training framework	Solve the sample imbalance issue with an attention mechanism to strive for a balance from less labeled data	Validation can be further conducted with less annotations
Narasimman et al. (2021)	Semi-supervised	*MeanTeacher*: Framework combining CNN and mean teacher learning scheme which consists of a student model and a teacher model	Average model weights over training steps to produce more robust and accurate mode	To combine different ways of consistency regularizations, such as mean teacher and virtual adversarial training
Bi et al. (2021)	Active learning	*DAL*: Support vector machine with dynamic active learning which is able to detect unknown activities	Reduce the annotation cost; Adaptive to scenarios with novel activities and patterns	To extend the active learning approach to deep models

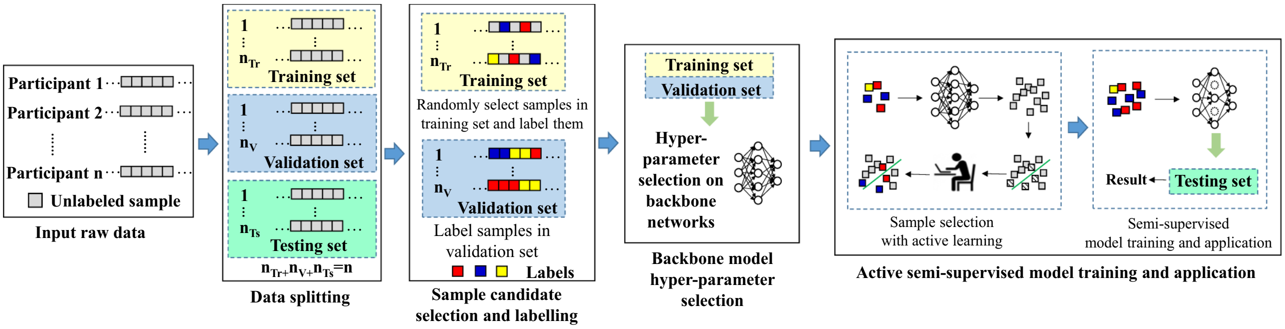

Participant 1
Participant 2
Participant n
Unlabeled sample
Input raw data
1
n_Tr
Training set
n_V
Validation set
n_Ts
Testing set
n_Tr+n_V+n_Ts=n
Data splitting
Randomly select samples in training set and label them
Label samples in validation set
Labels
Sample candidate selection and labelling
Training set
Validation set
Hyper-parameter selection on backbone networks
Backbone model hyper-parameter selection
Sample selection with active learning
Result
Testing set
Semi-supervised model training and application
Active semi-supervised model training and application

Fig. 1 Workflow of the proposed method. Given data from a number of participants, we firstly split the dataset into three independent subsets. Then samples from the validation set and randomly selected ones (a small number) from the training set are manually annotated. Next, hyperparameters of the backbone model are selected based on the labelled training and validation sets. In the active semi-supervised model training process, we first select informative samples via active learning for annotation, and then train the model with both labelled and unlabelled instances. Finally, the obtained model is applied on the test set

the backbone neural network. In the active semi-supervised model training process, we first select informative samples via active learning to request for annotations, and then train the semi-supervised model with both the labelled and unlabelled instances. Finally, the model is applied on the test set, examining the models performance.

Next, we describe the framework of the proposed active semi-supervised deep model. For a given HAR training set with N instances, we use $\mathbf{X}$ to denote all the training instances, L indicates the number of labels of the annotation budget. The labelled sample set is denoted as $S_s = \{(x_i, y_i), i = 1, \dots, l\}$ with $y_i \in \{1, \dots, C\}$, where C indicates the number of activity classes and $l < L$. Figure 2 illustrates the framework, which includes two steps, i.e., active learning based sample selection (*Step 1*) and semi-supervised model training with temporal ensembling (*Step 2*).

In *Step 1*, given the input raw data with labelled sample set S_s, we iteratively select samples to be annotated via active learning under the restriction of annotation budget L. In each iteration, we first train a supervised deep model with the existing labels and then make predictions on the unlabelled candidates, based on which the most informative samples are selected and annotated, yielding an updated set of labels for next iteration. It is noted that the initial labels S_s for the first active learning iteration are a small number of randomly selected and annotated samples. In *Step 2*, both the labelled and unlabelled data are fed into a temporal ensembling-based 1-dimensional (1-D) CNN networks for training with dropout as a regularization method. We apply an

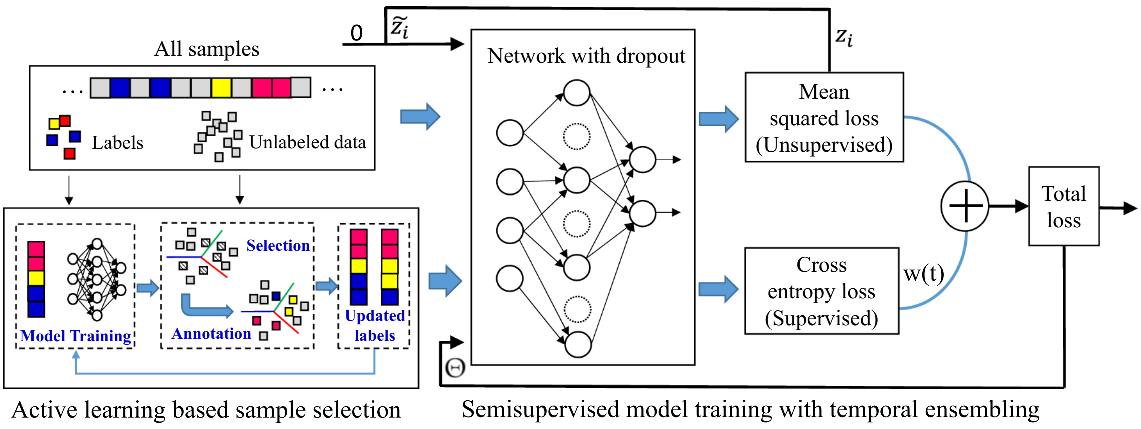

All samples
Labels
Unlabeled data
Model Training
Selection
Annotation
Updated labels
Active learning based sample selection
0
$\tilde{z}_i$
Network with dropout
z_i
Mean squared loss (Unsupervised)
Cross entropy loss (Supervised)
w(t)
Total loss
Θ
Semisupervised model training with temporal ensembling

Fig. 2 Model framework. *Step 1*: Active learning based sample selection. This is an iterative process, where in each iteration, we first train a supervised deep CNN model with existing labels and then make predictions on the unlabelled candidates, based on which the most informative samples are actively selected to request new annotations, yielding an updated set of labels. *Step 2*: Semi-supervised model training with temporal ensembling. Both the labelled and unlabelled samples are fed into the network with dropout regularization. An integrated loss function which consists of a supervised and an unsupervised loss term is then calculated based on the predictions of the current network. Next, we update network parameters Θ by optimizing the loss function via stochastic gradient descent (SGD) algorithm. The updated Θ and normalized ensemble prediction $\tilde{z}$ act as input for the network training for the next iteration

integrated semi-supervised loss function with two terms—i.e., supervised loss term and unsupervised loss term—to learn the weights of the networks. The optimization of the loss function is an iterative process, where the updated network parameters Θ and normalized ensemble prediction $\tilde{z}$ in current iteration act as input for the network training of the next iteration. The $\tilde{z}$ is set as 0 in the first iteration. We will describe the two steps with more details in Sects. 3.2 and 3.3 respectively.

In this work, we exploit a backbone CNN structure as illustrated in Fig. 3, which consists of nine convolutional layers, two max-pooling layers, two dropout layers, one average pool layer, one fully connected layer and a softmax layer connected to the output (Laine and Aila 2016). Batch normalization and Leaky rectified linear unit (LReLU) activation function (Maas et al. 2013) are sequentially applied following each of the convolutional layers. With regard to the network parameters, please refer to Fig. 3 for more details. This CNN structure is utilized in both the active learning based sample selection and semi-supervised model training processes. It should be pointed out that other network structures can be flexibly incorporated with the proposed framework as well. Yet in this work, we only report the results with the CNN structure as shown in Fig. 3.

3.2 Active learning based sample selection

Unlike supervised learning which trains classifiers with randomly chosen and annotated samples, we use active learning to tactfully select the most beneficial set of samples to be annotated. By doing so, the unnecessary annotation of samples carrying little information is circumvented, which effectively improves the labelling efficacy, thus reducing the labelling cost. Therefore, it is essential to define an effective criterion to assess the informativeness of candidate samples and then select proper candidates to be annotated.

Uncertainty-based active sampling scheme, which tends to select samples with highest uncertainty, is the most recognized and extensively employed sampling criterion. Entropy is a typical indicator for measuring the uncertainty of a probabilistic distribution. Higher values of entropy imply more uncertainty in the distribution (Settles 2009). However, several works in diverse applications (Cao et al. 2020; Bi et al. 2019, 2021) have experimentally shown that, although the performance of entropy-based active learning is generally superior to passive random selection, the performance improvement of the strategy decreases when high entropies are caused by small class probabilities of nonsignificant classes. This issue becomes more severe when a large label set is present for multi-class classification tasks. Previous results (Cao et al. 2020; Bi et al. 2021) show that best-versus-second-best (BvSB)-based active selection can effectively overcome the shortcoming of entropy-based sampling scheme by measuring the difference in class probabilities between the first and second most probable classes. For this reason, we apply BvSB as the sample selection scheme in this work. Let Θ represent the learnable parameters of the deep model, $P(y_{\mathrm{B}}|x_i, \Theta)$ and $P(y_{\mathrm{SB}}|x_i, \Theta)$ denote the two highest estimated class probabilities of sample x_i output from the classifier, the sampling criterion can be described as:

$$x_i^{BvSB} = \underset{x_i, i \in \mathcal{U}}{\arg\min} \left(P\left(y_{\mathrm{B}}|x_i, \Theta\right) - P\left(y_{\mathrm{SB}}|x_i, \Theta\right) \right). \tag{1}$$

With this sampling scheme, the instances close to the decision boundaries are preferred to be selected. In BvSB-based active learning, we first compute the class probabilities of samples in the candidate pool $\mathcal{U}$. Samples meeting the BvSB

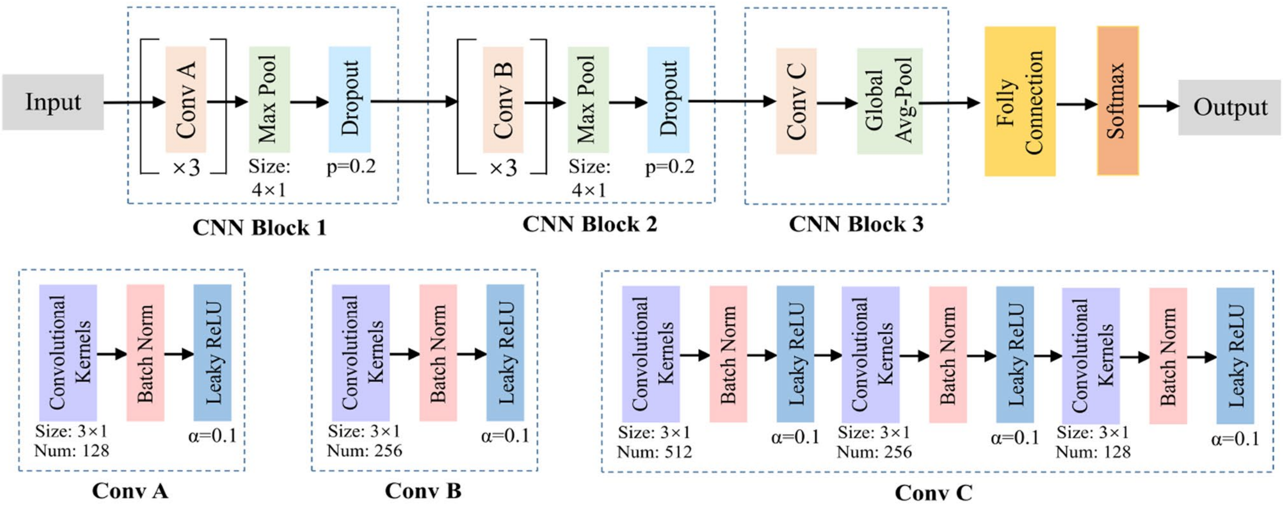


Fig. 3 Backbone network architecture

sampling criterion are then iteratively selected to be annotated and incorporated to the training set for the consequent classifier retraining.

It is noteworthy that this work is more concerned about the combination of active learning and SSL, and the application of the unified framework on HAR. BvSB is one promising solution for providing informative samples for the consequent model training. However, other active learning strategies can be applied as well provided that they are capable of selecting the most beneficial instances.

3.3 Semi-supervised CNN model with temporal ensembling

Given the updated labelled sample set S_s, we next introduce our proposed semi-supervised CNN model with temporal ensembling which aims to learn a deep model making use of both labelled and unlabelled samples. To this end, we define a loss function given by:

$$Loss\left(\Theta|\mathbf{X}, S_s\right) = Loss_s\left(\Theta|\mathbf{X}, S_s\right) + w(t) \times Loss_u(\Theta|\mathbf{X}), \quad (2)$$

where Θ denotes the combination of the CNN parameters to be optimized. $Loss_s(\Theta)$ stands for the *Supervised loss term*, and $Loss_u(\Theta)$ is the *Unsupervised loss term*. $w(t)$ is the unsupervised loss weighting function, which starts from zero, and ramps up along a Gaussian curve during the first 80 training epochs (Laine and Aila 2016).

3.3.1 Supervised loss term

The supervised loss term is designed to enforce the consistency between the CNN prediction on the labelled samples and ground truth labels, which follows the cross entropy loss form as,

$$\begin{aligned} & Loss_s(\Theta|\mathbf{X}, S_s) \\ & \quad = -\frac{1}{L}\sum_{i=1}^{L}\sum_{j=1}^{C} 1\{y_i = j\} \log P(y_i = j|x_i, \Theta), \end{aligned} \quad (3)$$

where $P(y_i = j|x_i, \Theta)$ indicates the probability of predicting x_i to have class label j.

3.3.2 Unsupervised loss term

Due to the fact that an ensemble of multiple neural networks generally yields better predictions than a single network (Srivastava et al. 2014a), we adopt a temporal ensembling strategy (Laine and Aila 2016) to promote the model's performance. In this scheme, the training is performed on a single network, however, the predictions are made on a number of pre-networks by accumulating the predictions of multiple frozen instances of the same network during training. Therefore the history-involved predictions correspond to an ensemble consensus of a large number of pre-networks from different epochs. Dropout approach (Srivastava et al. 2014b) is utilized as a regularizer, which has been shown effective to generalize and provide more certain predictions.

Based on the above analysis, we apply the temporal ensembling to the activity prediction of unlabelled samples, where the labels inferred in this way are exploited as training targets for the unlabelled instances. The unsupervised loss term measures the divergence between the current network outputs and the previous ensemble predictions, as given by,

$$Loss_u(\Theta|\mathbf{X}) = -\frac{1}{CN}\sum_{i=1}^{N} \|z_i - z_i\|^2, \quad (4)$$

where z_i indicates the current prediction, and $\tilde{z}_i$ denotes the ensemble output aggregated from the deep neural networks in previous epochs. After every training epoch, the network outputs z_i are accumulated into ensemble outputs $\mathbf{Z}_i$ by,

$$\mathbf{Z}_i = \alpha\mathbf{Z}_{i-1} + (1-\alpha)z_i, \quad (5)$$

where α is a momentum term that controls how far the ensemble reaches to previous training epochs. To generate training targets $\tilde{z}_i$, we perform bias correction on $\mathbf{Z}_i$ by,

$$\tilde{z}_i = \frac{\mathbf{Z}_i}{1-\alpha^t}, \quad (6)$$

where $\mathbf{Z}$ and $\tilde{z}$ are set to zero in the first training epoch.

From the above formula, we can see that at the start of training the network, the total loss is dominated by the supervised loss term. As the training evolves, the unsupervised loss term plays a more important role. The optimization of the loss function in Eq. 2 is conducted using the SGD algorithm. In the tth iteration, the model parameters Θ are updated with τ as the learning rate,

$$\Theta_{t+1} = \Theta_t - \tau\frac{\partial Loss(\Theta|\mathbf{X}, S_s)}{\partial\Theta}. \quad (7)$$

To conclude this section, Algorithm 1 and Fig. 4 illustrate the pseudo-code and flowchart of the proposed approach respectively.

Algorithm 1 Pseudo code for the proposed method.

```
Input: Raw temporal data sequences $\mathbf{X}$;
       Annotation budget $L$
       Ensembling momentum $\alpha$, $0 < \alpha < 1$
       Number of samples selected in each active learning iteration $n$
Output: Model parameter $\Theta$.
Randomly select $l$ samples ($l<L$) from $\mathbf{X}$ and query for annotations, getting labelled sample set $S_s$
while $num\_labels < L$ do
    Deep supervised model training with $S_s$
    Make prediction on unlabelled candidate pool $\mathscr{U}$
    Actively select $n$ samples from $\mathscr{U}$ and query for annotations
    Update $\mathscr{U}$ and $S_s$
end while
$\mathbf{Z}$ Initialization: $\mathbf{Z} \leftarrow 0$
$\tilde{\mathbf{z}}$ Initialization: $\tilde{z} \leftarrow 0$
Iteration $t$=1
while $t < num\_epochs$ do
    Calculate supervised loss $Loss_s$ based on current network with parameters $\Theta$
    Calculate unsupervised loss $Loss_u$ based on current network with parameters $\Theta$
    Update total loss function $Loss = Loss_s + w(t) \times Loss_u$
    Update $\Theta$ by optimizing the total loss function via SGD algorithm
    Accumulate ensemble predictions $\mathbf{Z} \leftarrow \alpha\mathbf{Z} + (1-\alpha)z$
    Construct target vectors by bias correction $\tilde{z} \leftarrow \frac{\mathbf{Z}}{1-\alpha^t}$,
    $t \leftarrow t+1$
end while
return $\Theta$
```

4 Experimental evaluation

In this section, we justify and analyze the performance of the proposed method on three real HAR datasets. We first introduce the datasets and experimental settings in Sect. 4.1. Next, we conduct ablation study in Sect. 4.2. Section 4.3 reports the comparative results with several competing approaches on benchmark object test, statistical significance test, and cross validation (CV) settings. The impact of parameters on the model performance is presented in Sect. 4.4.

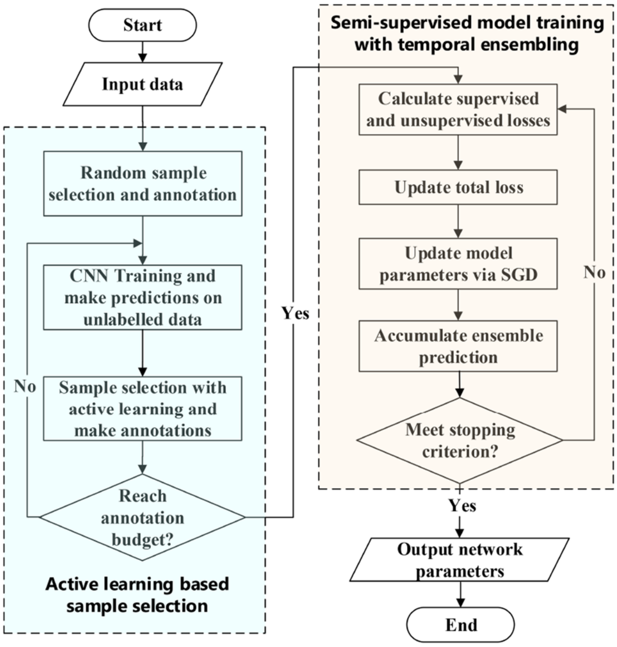

Start
Input data
Random sample selection and annotation
CNN Training and make predictions on unlabeled data
Sample selection with active learning and make annotations
Reach annotation budget?
No
Yes
Active learning based sample selection
Semi-supervised model training with temporal ensembling
Calculate supervised and unsupervised losses
Update total loss
Update model parameters via SGD
Accumulate ensemble prediction
Meet stopping criterion?
No
Yes
Output network parameters
End

Fig. 4 Flowchart of the proposed approach

4.1 Datasets and experimental setting

4.1.1 Datasets

We consider 3 public benchmark datasets for the evaluation of the proposed method, where the key information of them is summarized in Table 2.

The PAMAP2 dataset (Reiss and Stricker 2012) was collected on 9 participants wearing three inertial measurement units on the hand, chest, and ankle respectively, sampled at 100 Hz. The sensors include accelerometer, gyroscope, magnetometer, temperature and heart rate sensor. There are 52 raw features in total. This dataset contains 12 activities including *lying, sitting, standing, walking, running, cycling, Nordic walking, ascending stairs, descending stairs, cleaning, ironing, and rope jumping*. Following (Saeed et al. 2019; Haresamudram et al. 2020), the data from participant 106 is used for testing, the data from participant 105 for validation and the rest for training.

The USCHAD dataset (Zhang and Sawchuk 2012) was recorded on 14 volunteers using triaxial accelerometer and gyroscope which were attached to participant's front right hip, yielding 6 features in total. The sampling rate of sensor data is 100 Hz. The dataset includes 12 activities: *walking forward, walking left, walking right, walking upstairs,* *walking downstairs, running forward, jumping, sitting, standing, sleeping, elevator up* and *elevator down*. Following (Saeed et al. 2019; Haresamudram et al. 2020), data from participants 1 to 10 is used for training, 11 and 12 for validation, and 13 and 14 for testing.

The UCIHAR Dataset (Anguita et al. 2013) was collected with triaxial accelerometer and gyroscope from a Samsung Galaxy SII smart phone worn by 30 volunteers on their waist, providing 6 features. The activity set includes 6 basic human activities, *walking, walking upstairs, walking downstairs, sitting, standing* and *laying*. Following Chen et al. (2020), Khan and Ahmad (2021), the training set was created with 70% of the volunteers, whereas the remaining 30% were selected to generate the test test.

4.1.2 Experimental setting

Table 3 displays the experimental configurations of the proposed method. The learning rate follows the cosine annealing function (Loshchilov and Hutter 2016). We choose the minimum value of learning rate τ and weight decay by grid search on the baseline CNN with a limited training set, i.e., 200 randomly selected training samples, and the validation set. We performed experiments with τ over $\{0.00001, 0.0001, 0.001, 0.01\}$ and weight decay over $\{0.0001, 0.0005, 0.001, 0.005, 0.01\}$. The τ and weight decay with the best performance are selected as the hyperparemters used for model training. Based on the obtained results, the weight decay and τ are selected as 0.0005 and 0.0001 respectively. We empirically set batch size as 20 in low annotation setting with an annotation budget less than 500, and otherwise 100. We run 50 epochs for network training. All the CNN based methods use the same hyperparameters as exhibited in Table 3.

We studied the impact of the dropout rate and ensembling momentum on the semi-supervised temporal ensembling, where the analysis can be found in Sect. 4.4. Based on the comparative results, dropout rate and ensembling momentum are set as 0.2 and 0.6 respectively. For active learning, we randomly select 100 samples in the first iteration for the supervised model training, and actively select 100 samples to be annotated in each active learning iteration. All the experiments were executed on a workstation with GeFroce RTX 3090 GPU and 64GB RAM, whist the codes are implemented with Pytorch library.

Table 3 Experimental configurations

Parameter	Value	Parameter	Value
Min. learning rate	0.0001	Dropout rate	0.2
Weight decay	0.0005	Ensembling momentum	0.6
Batch size	20	Samples selected	100

4.1.3 Evaluation metric

As is common in HAR research, we use Macro F1-Score as the evaluation metric, defined in the usual way:

$$\begin{aligned} MacroF1 &= \frac{1}{C}\sum_{i=1}^{C}\frac{2\cdot Precision_i\cdot Recall_i}{Precision_i+Recall_i},\\ Precision_i &= \frac{TP_i}{TP_i+FP_i},\\ Recall_i &= \frac{TP_i}{TP_i+FN_i}, \end{aligned} \tag{8}$$

where for a given class i, TP_i and FP_i denote the number of true positives and false positives respectively, and FN_i represents the number of false negatives. Macro F1-Score is denoted as *MacroF1* in following sections.

4.2 Ablation study

In this set of experiments, taking the PAMAP2 data set as an example, we conduct an ablation study to examine the contributions of the key components of the method, i.e., active learning based sample selection and semi-supervised learning using temporal ensembling, to the prediction performance. To verify their effectiveness, apart from the baseline CNN, we establish two comparison methods which successively add each of the two components. The compared methods include:

1. CNN: This is the baseline CNN model with the architecture depicted in Fig. 3.
2. ActCNN (short for active convolutional neural networks): the combination of CNN and active learning.
3. *ActSemiCNN* (short for active semi-supervised convoluational neural networks): the combination of CNN, active learning and SSL, i.e., the proposed method.

Table 2 Summary of the dataset

Dataset	# of activities	Test subject	# of users	# of features	# of samples	Sensors
PAMAP2	12	106	9	52	38,857	A, G, M, T, H
USCHAD	12	13, 14	14	6	56,228	A, G
UCIHAR	6	30%	30	6	10,299	A, G

A accelerometer, *G* gyroscope, *M* magnetometer, *T* temperature, *H* heart rate

To analyze the sensitivity of the three compared methods to the number of labels, we performed experiments with different annotation budgets. To get more reliable results, we performed 5 rounds of independent repetitions for all three methods, and the average *MacroF1* values and standard deviations are reported. It is noted that for the 100 label annotation budget setting, 50 samples are randomly selected in the first iteration and 50 samples are further chosen in the active learning process. Figure 5 presents the *MacroF1* value curves as a function of the annotation budgets, and the numerical results are listed in Table 4. Figure 6 gives the confusion matrices of the inner steps in one repetition of *ActSemiCNN* with an annotation budget of 200. These results lead to the below observations:

1. *ActSemiCNN* constantly yields the highest *MacroF1* value with any number of labels compared. The comparisons between ActCNN and CNN, and *ActSemiCNN* and ActCNN respectively reflect the benefits brought by active learning and SSL. In this experiment, when only 200 samples can be annotated, ActCNN outperforms CNN by 0.05 *MacroF1* units, while *ActSemiCNN* outperforms ActCNN by 0.04.
2. Inspecting Fig. 6a, we can notice obvious confusions between *walking* and *Nordic walking*, and among *ascending stairs*, *descending stairs*, *cleaning* classes. Yet, confusions among the 5 most confusing classes are greatly relieved in the results of ActCNN as shown in Fig. 6b, which can be explained as follows. Take the repetition in Fig. 6 for example, our active learning policy chooses more instances (81) from the 5 classes compared with random selection (44), and selects less samples from the classes that the current classifier is confident with (1 sample for *lying* and no sample from *sitting*). ActCNN surpasses CNN with random samples by a *MacroF1* value of 0.05, precisely justifying the effectiveness of active learning.
3. Fig. 6c reflects a further improvement on *MacroF1* value (0.04) compared with Fig. 6b, which proves that utilizing unlabelled data is particularly effective in improving the recognition performance.
4. The *MacroF1* value (0.76) of *ActSemiCNN* with 200 labels is even higher than the one achieved by the CNN with 500 labels (0.74), showing an obvious reduction of the annotation cost.

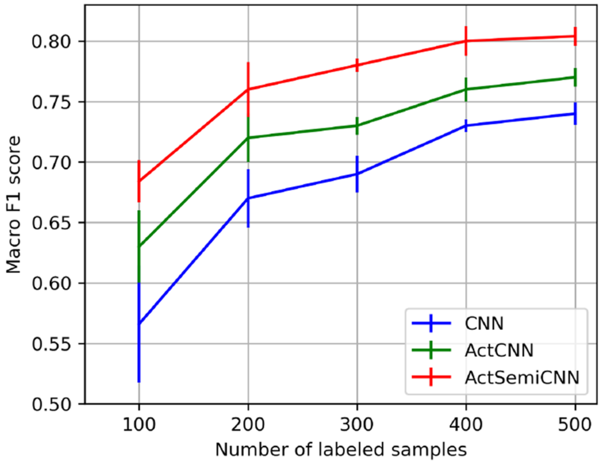


Fig. 5 *MacroF1* values as a function of the number of labels on the PAMAP2 dataset

4.3 Results and comparison

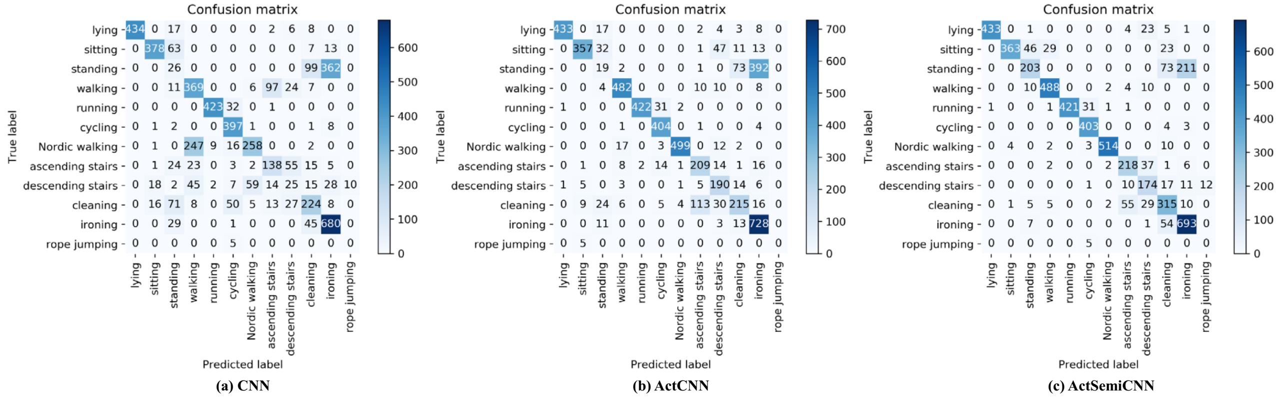
In this section, we conduct an extensive evaluation of our approach on the PAMAP2, USCHAD and UCIHAR datasets. The comparison results with 7 competitors on benchmark test objects, statistical significance test and CV settings are presented in Sects. 4.3.1–4.3.3 respectively. The compared methods include:

1. DeepConvLSTM (Ordóñez and Roggen 2016): DeepConvLSTM is a supervised deep architecture based on the combination of convolutional and LSTM recurrent layers to recognize activities. DeepConvLSTMs reached state-of-the-art in distinguishing complex human activities (Slaton et al. 2020; Mahmud et al. 2020).
2. CoTrCNN: This is a semi-supervised method using a co-training pipeline (Stikic et al. 2008) while with a CNN model (Laine and Aila 2016) as shown in Fig. 3.
3. ConvAE (Haresamudram et al. 2019): This is a state-of-the-art deep architecture (Mahmud et al. 2020; Liu

Table 4 *MacroF1* values with different combinations of methods on PAMAP2 dataset with different training sample numbers

Method	100	200	300	400	500
CNN	0.57 ± 0.05	0.67 ± 0.02	0.69 ± 0.02	0.73 ± 0.01	0.74 ± 0.01
ActCNN	0.63 ± 0.03	0.72 ± 0.02	0.73 ± 0.01	0.76 ± 0.01	0.77 ± 0.01
ActSemiCNN	**0.68** ± 0.02	**0.76** ± 0.02	**0.78** ± 0.01	**0.80** ± 0.01	**0.80** ± 0.01

Bold values represent the highest numerical values and the method with best performance



(a) CNN

(b) ActCNN

(c) ActSemiCNN

Fig. 6 Confusion matrices. **a** CNN with 100 initial samples. **b** ActCNN with 100 initial samples and 100 actively selected samples. **c** ***ActSemiCNN*** with 100 initial samples and 100 actively selected samples

et al. 2020) which includes a convolutional encoder and decoder with a bottleneck layer in between. The feature representation from the bottleneck layer is used by an MLP for the classification.

4. MTSelfCNN (Haresamudram et al. 2020): This is a state-of-the-art self-supervised method. The accelerometer representations are learnt by training a multi-task CNN to recognize eight transformations applied to the raw input signal. The original paper reports the results on UCIHAR dataset in a semi-supervised setting, which is used for comparison in this paper.
5. SemiConvAttn (Chen et al. 2020): This is a semi-supervised deep model based on co-training framework, where attention-based recurrent convolutional models are introduced to handle multi-modality data.
6. MeanTeacher (Narasimman et al. 2021): This is a semi-supervised HAR model combining state-of-the-art mean teacher learning scheme and CNN.
7. DAL (Bi et al. 2021): This is a state-of-the-art active learning-based model which selects samples via marginal sampling scheme coupled with temporal-frequency features.

Among the above 7 approaches, we self-implemented DeepConvLSTM, DAL, CoTrCNN and MeanTeacher. The numerical results used for comparison with other 3 methods are directly extracted from the original papers (Haresamudram et al. 2019, 2020; Chen et al. 2020).

4.3.1 Performance on benchmark test object

The quantitative results on benchmark test objects are summarized in Table 5, wherein the *MacroF1* values, used labels and time consumption are reported, with the best results highlighted in bold. The standard deviations of the 5 self-implementing methods are obtained by averaging the results of 5 independent repetitions.

Table 5 suggests that our proposed method yields the best performance compared with other competitors on all three datasets. With the same number of used labels, *ActSemiCNN* outperforms DeepConvLSTM, DAL, CoTrCNN and MeanTeacher by 0.14, 0.09, 0.10 and 0.09 on PAMAP2 dataset, 0.05, 0.03, 0.10 and 0.04 on USCHAD dataset, 0.29, 0.16, 0.10, and 0.06 on UCIHAR dataset respectively. *ActSemiCNN* even outperforms ConvAE with 20% labels and SemiConvAttn with 1000 labels which demonstrates that our proposed method substantially reduces the annotation cost without loss of predictive performance. The superiority of *ActSemiCNN* over DAL again demonstrates the advantage of utilizing unlabelled data during model training.

We can see that the overall *MacroF1* performance on the USCHAD dataset is lower than on the PAMAP2 dataset, which is because USCHAD is a challenging dataset. Firstly, the sensor data is collected from the motion node attached to the hip, which provides less information than the multi-position case as PAMAP2 dataset. Secondly, the activities involve orientation such as elevator up or down which are difficult to discriminate (Mahmud et al. 2020). Especially, CoTrCNN achieves the lowest *MacroF1* value of 0.35, which is due to two reasons. First, CoTrCNN directly assigns pseudo labels to unlabelled samples and utilizes them in the model retraining, which is unreliable when the model's performance is unsatisfactory. Secondly, CoTrCNN splits the features into two views, which limits the performance of the model with incomplete views compared with the complete view scenario.

As revealed in Table 5, *ActSemiCNN* and MeanTeacher consume more time than others, which is owing to their intrinsic mechanism that the massive unlabelled samples are engaged in training throughout the whole process.

Table 5 *MacroF1* values with different methods on benchmark test objects

Dataset	Method	Supervision	Used labels	MacroF1	Running time
PAMAP2	DeepConvLSTM (2016)	Supervised	200	0.62 ± 0.02	3 m 21 s
	ConvAE (2019)	Supervised	20% (5681)	0.72	–
	DAL (2021)	Active learning	200	0.67 ± 0.03	33 m 16 s
	CoTrCNN (2016)	Semi-supervised	200	0.64 ± 0.02	8 m 22 s
	SemiConvAttn (2020)	Semi-supervised	1000	0.73	–
	MeanTeacher (2021)	Semi-supervised	200	0.67 ± 0.03	31m19s
	ActSemiCNN	Semi-supervised	200	**0.76** ± 0.02	35m15s
USCHAD	DeepConvLSTM (2016)	Supervised	200	0.40 ± 0.02	12 m 41 s
	ConvAE (2019)	Supervised	20% (7251)	0.43	–
	DAL (2021)	Active learning	200	0.42 ± 0.01	5 m 46 s
	CoTrCNN (2016)	Semi-supervised	200	0.35 ± 0.03	6 m 42 s
	MeanTeacher (2021)	Semi-supervised	200	0.41 ± 0.02	37 m 11 s
	ActSemiCNN	Semi-supervised	200	**0.45** ± 0.02	34 m 43 s
UCIHAR	DeepConvLSTM (2016)	Supervised	300	0.62 ± 0.02	3 m 55 s
	DAL (2021)	Active learning	300	0.75 ± 0.01	1 m 21 s
	CoTrCNN (2016)	Semi-supervised	300	0.81 ± 0.01	4 m 58 s
	MTSelfCNN (2020)	Semi-supervised	300	0.87	–
	SemiConvAttn (2020)	Semi-supervised	1000	0.73	–
	MeanTeacher (2021)	Semi-supervised	300	0.85 ± 0.01	8 m 11 s
	ActSemiCNN	Semi-supervised	300	**0.91** ± 0.01	7 m 34 s

Bold values represent the highest numerical values and the method with best performance

4.3.2 Statistical significance comparison

Apart from the numerical evaluation, we further compare the statistical significance by carrying out the variance-based hypothetical F-test. Given that the original papers of ConvAE, MTSelfCNN and SemiConvAttn only provide the mean experimental results yet without variance information, the significance comparison was conducted on five self-implementing methods. It is hypothesized that the *MacroF1* values of multiple repetitions of each compared method follow the Normal distribution. With this assumption, we calculated the *p* values to represent the level of evidence against null hypothesis which suggests that no statistical difference exists between two sets of observations. It is supposed that there is a significant difference between two groups of results when the obtained *p* value is less than a commonly employed threshold of 0.05.

Tables 6, 7 and 8 tabulate the statistical significance results on three experimental datasets. S1–S5 in the tables indicate the DeepConvLSTM, DAL, CoTrCNN, MeanTeacher and *ActSemiCNN* correspondingly. To specify, the symbols ‘0’, ‘+1’, ‘−1’ respectively mean that the method in the column is significantly equivalent, better and worse than the method in the row. The symbol ‘-’ denotes that the method in the column is the same with the one in the row.

Inspecting the 3 tables, we can draw below conclusions:

1. Our proposed **ActSemiCNN** is statistically superior to other competitors on all three datasets.
2. The two semi-supervised methods which involve all the unlabelled instances during training, i.e., MeanTeacher and *ActSemiCNN*, consistently yield better at least comparable results compared to other 3 methods.

4.3.3 Performance on cross-validation experiment

To demonstrate the robustness of the proposed method with regard to sensitivity to specific test subjects, we further conduct a Leave-one-subject-out CV (LOSO-CV) experiment. During the experiment, we repeatedly hold the data from one of the subjects out of the training set and use it only for testing purposes. This is done until all the subjects have been used in the test set, and the average values of *MacroF1*

Table 6 Significance comparison on PAMAP2 dataset

	S1	S2	S3	S4	**S5**
S1	–	−1	−1	−1	−1
S2	+1	–	0	0	-1
S3	+1	0	–	-1	−1
S4	+1	0	+1	–	−1
S5	+1	+1	+1	+1	–

Bold value represents the highest numerical values and the method with best performance

Table 7 Significance comparison on USCHAD dataset

	S1	S2	S3	S4	**S5**
S1	–	0	+1	0	−1
S2	0	–	+1	+1	−1
S3	−1	−1	–	−1	−1
S4	0	−1	+1	–	−1
S5	+1	+1	+1	+1	–

Bold value represents the highest numerical values and the method with best performance

Table 8 Significance comparison on UCIHAR dataset

	S1	S2	S3	S4	**S5**
S1	–	−1	−1	−1	−1
S2	+1	–	−1	−1	−1
S3	+1	+1	–	−1	−1
S4	+1	+1	+1	–	−1
S5	+1	+1	+1	+1	–

Bold value represents the highest numerical values and the method with best performance

scores over all subjects are computed. In this experiment, we only focus in the low annotation scenario with 200 labels available. It should be noted that the correspondence between the samples and objects in UCIHAR dataset is unavailable, thereby the results reported here were obtained with a 5-fold CV instead.

Table 9 presents the results of CV experiments. As can be observed from the table, the *MacroF1* scores of the proposed method are constantly higher than the competing methods for CV experiments. *ActSemiCNN* outperforms DeepConvLSTM, DAL, CoTrCNN and MeanTeacher by 0.12, 0.13, 0.08 and 0.10 on PAMAP2 dataset, 0.13, 0.08, 0.15 and 0.03 on USCHAD dataset, 0.27, 0.14, 0.15 and 0.03 on UCIHAR dataset respectively, which suggests that *ActSemiCNN* is robust to inter-subject variability.

4.4 The impact of parameters

This section investigates the impact of the key parameters of *ActSemiCNN*–dropout rate and ensembling momentum–on the recognition performance. Taking PAMAP2 dataset as an example, we conducted comparative experiments on the above 2 parameters with varying values, where the results are tabulated in Tables 10 and 11.

The dropout rate specifies the proportion of nodes randomly dropped out during training. Validation experiments were performed with dropout rates ranging [0, 0.5] with a step of 0.1, where the value 0 means that we do not drop any nodes in the network. From Table 10, we can find that the highest *MacroF1* score is achieved with a dropout rate of 0.2. Based on these results, we have set the dropout rate to 0.2 in all other experiments.

Table 9 *MacroF1* values of CV experiments with different methods on two datasets

Method	Used labels	PAMAP2	USCHAD	UCIHAR
DeepConvLSTM	200	0.54	0.49	0.58
DAL	200	0.53	0.54	0.71
CoTrCNN	200	0.58	0.47	0.70
MeanTeacher	200	0.56	0.59	0.82
ActSemiCNN	200	**0.66**	**0.62**	**0.85**

Bold values represent the highest numerical values and the method with best performance

Ensembling momentum controls how far the aggregation reaches to the training history, which has a value range of [0, 1). Value 1 indicates that the current loss is totally originated from the history predictions, while value 0 means that no history information is included in the current loss. We conducted experiments with ensembling momentum ranging [0.4, 0.8] with a step of 0.1. We can discover from Table 11 that the best result is yielded with a value of 0.6, which is employed as the ensembling momentum configuration in all experiments.

5 Discussions

The limited availability of annotations is arguably the most critical barrier that hinders the development of HAR. Thereby, it is crucial to develop approaches which can achieve favorable activity recognition performance while easing the burden of annotations. This is precisely the motivation of this work. The novelties of this paper mainly lie in two aspects: (i) the initial integration of active learning and semi-supervised learning into one HAR framework; (ii) the exploitation of temporal ensembling with consistency regularization in the deep CNN optimization. In what follows, we will: (1) offer explanations of *ActSemiCNN*, (2) provide insights on potential future research, (3) analyze the limitations of the proposed model.

5.1 Explanations

The comparative experiments in Sect. 4 suggest that the proposed model considerably boosts the recognition performance in low annotation regime, and exhibits strong robustness and generalization ability. The superiority of *ActSemiCNN* is attributed to the effectiveness of its two components. (1) The applied active learning selects samples which are most difficult to classify, therefore the inclusion of the

Table 10 *MacroF1* values of the proposed method on PAMAP2 dataset with different dropout rates

Dropout Rate	0	0.1	0.2	0.3	0.4	0.5
MacroF1	0.57 ± 0.03	0.62 ± 0.03	**0.63** ± 0.02	0.62 ± 0.03	0.61 ± 0.02	0.58 ± 0.03

Table 11 *MacroF1* values of the proposed method on PAMAP2 dataset with different ensembling momentum values

Momentum	0.4	0.5	0.6	0.7	0.8
MacroF1	0.58 ± 0.03	0.61 ± 0.02	**0.63** ± 0.02	0.60 ± 0.03	0.59 ± 0.03

annotations of these samples in training effectively helps the classifier refine the decision boundaries. (2) Leveraging the unlabelled data during training gives rise to more accurate results and the consensus ensemble prediction further alleviates the prediction uncertainty.

5.2 Insights

By analysis, we can conclude that the integration of diverse weak supervision paradigms ends up with additive benefits on the activity prediction, which offers insights for potential future researches. (1) Besides the label scarcity issue, the HAR task is confronted with another challenge, i.e., individual diversity. The hybrid of active learning and transfer learning is believed to be a promising solution for transferable personalized activity analysis, yet with sparse annotations. (2) Activity data are usually collected with multiple sensors, which brings challenges on the data fusion. The synthetic integration of self-supervised learning and semi-supervised learning is a compelling venue to address this issue, where the former enables the mutual and complementary interaction via proxy task between data of different views, and the latter is capable of inferring useful information from the vast amount of unlabelled data.

5.3 Limitations

Although prominent improvements are revealed by this study, *ActSemiCNN* still suffers from two limitations. (1) Involving the unlabelled data throughout the whole training process inevitably increases the time consumption. More computation resources are needed if there are requirements on the processing timeliness. (2) Although active learning greatly reduces the number of samples to be annotated, it poses higher standards on the labelling quality. This is because in active learning paradigm, the queries are usually raised on samples that are most difficult to classify, which requires that the experts have the capability to assign correct labels out of two or several options. This means that the oracles need to be cautiously selected to guarantee the smooth and effective proceeding of active learning.

6 Conclusion

Recognizing human activities from wearable sensors has been a challenging task, especially when annotations are scarce. The prime purpose of this paper is to explore the effect of (1) the combination of different paradigms of weakly supervised learning, and (2) the ensemble consensus, on the accuracy and robustness of human activity recognition. To this end, we presented a novel activity recognition approach called *ActSemiCNN* which integrates active learning and semi-supervised learning benefiting from temporal ensembling into one framework. We draw below conclusions from extensive comparative experiments: (1) The integration of active learning and semi-supervised learning leads to state-of-the-art performance, and the annotation cost is greatly reduced, which is attributed to the active sample selection and the utilization of massive unlabelled data. (2) The statistical significance and cross validation tests highlight the effectiveness of ensemble consensus in enhancing the robustness of HAR models.

In future work we plan to investigate few-shot learning for adaptive and personalized human activity recognition. We are also interested in exploring the interplay between different views of multi-modality activity data.

Acknowledgements This work was supported by SPHERE Next Steps Project funded by the U.K. Engineering, and Physical Sciences Research Council (EPSRC) under Grant EP/R005273/1.

References

Alemdar H, van Kasteren TL, Ersoy C (2011) Using active learning to allow activity recognition on a large scale. In: International joint conference on ambient intelligence. Springer, pp 105–114

Amiribesheli M, Benmansour A, Bouchachia A (2015) A review of smart homes in healthcare. J Ambient Intell Humaniz Comput 6(4):495–517

Anguita D, Ghio A, Oneto L, Parra X, Reyes-Ortiz JL (2013) A public domain dataset for human activity recognition using smartphones. In: ESANN

Balabka D (2019) Semi-supervised learning for human activity recognition using adversarial autoencoders. In: Adjunct proceedings of the 2019 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2019 ACM international symposium on wearable computers, pp 685–688

Bi H, Xu F, Wei Z, Xue Y, Xu Z (2019) An active deep learning approach for minimally supervised PolSAR image classification. IEEE Trans Geosci Remote Sens 57(11):9378–9395

Bi H, Xu L, Cao X, Xue Y, Xu Z (2020) Polarimetric SAR image semantic segmentation with 3D discrete wavelet transform and Markov random field. IEEE Trans Image Process 29:6601–6614

Bi H, Perello-Nieto M, Santos-Rodriguez R, Flach P (2021) Human activity recognition based on dynamic active learning. IEEE J Biomed Health Inform 25(4):922–934

Bianchi V, Bassoli M, Lombardo G, Fornacciari P, Mordonini M, De Munari I (2019) IoT wearable sensor and deep learning: an integrated approach for personalized human activity recognition in a smart home environment. IEEE Internet Things J 6(5):8553–8562

Bota P, Silva J, Folgado D, Gamboa H (2019) A semi-automatic annotation approach for human activity recognition. Sensors 19(3):501

Cao X, Yao J, Xu Z, Meng D (2020) Hyperspectral image classification with convolutional neural network and active learning. IEEE Trans Geosci Remote Sens 58(7):4604–4616

Chen L, Nugent CD, Wang H (2011) A knowledge-driven approach to activity recognition in smart homes. IEEE Trans Knowl Data Eng 24(6):961–974

Chen K, Yao L, Zhang D, Wang X, Chang X, Nie F (2020) A semisupervised recurrent convolutional attention model for human activity recognition. IEEE Trans Neural Netw Learn Syst 31(5):1747–1756

Chen Y, Xue Y (2015) A deep learning approach to human activity recognition based on single accelerometer. In: 2015 IEEE international conference on systems, man, and cybernetics. IEEE, pp 1488–1492

Diethe T, Twomey N, Kull M, Flach P, Craddock I (2017) Probabilistic sensor fusion for ambient assisted living. arXiv:1702.01209

Gao W, Zhang L, Huang W, Min F, He J, Song A (2021) Deep neural networks for sensor-based human activity recognition using selective kernel convolution. IEEE Trans Instrum Meas 70:1–13

Gomes JB, Krishnaswamy S, Gaber MM, Sousa PA, Menasalvas E (2012) Mars: a personalised mobile activity recognition system. In: 2012 IEEE 13th international conference on mobile data management. IEEE, pp 316–319

Han J, He Y, Liu J, Zhang Q, Jing X (2019) Graphconvlstm: spatiotemporal learning for activity recognition with wearable sensors. In: 2019 IEEE global communications conference (GLOBECOM). IEEE, pp 1–6

Haresamudram H, Essa I, Plötz T (2021) Contrastive predictive coding for human activity recognition. Proc ACM Interact Mobile Wearable Ubiquitous Technol 5(2):1–26

Haresamudram H, Anderson DV, Plötz T (2019) On the role of features in human activity recognition. In: Proceedings of the 23rd international symposium on wearable computers, pp 78–88

Haresamudram H, Beedu A, Agrawal V, Grady PL, Essa I, Hoffman J, Plötz T (2020) Masked reconstruction based self-supervision for human activity recognition. In: Proceedings of the 2020 international symposium on wearable computers, pp 45–49

Hossain HS, Khan MAAH, Roy N (2017) Active learning enabled activity recognition. Pervasive Mob Comput 38:312–330

Khan ZN, Ahmad J (2021) Attention induced multi-head convolutional neural network for human activity recognition. Appl Soft Comput 110:107671

Khan AM, Tufail A, Khattak AM, Laine TH (2014) Activity recognition on smartphones via sensor-fusion and KDA-based SVMs. Int J Distrib Sens Netw 10(5):503291

Kumar V, Recupero DR, Riboni D, Helaoui R (2021) Ensembling classical machine learning and deep learning approaches for morbidity identification from clinical notes. IEEE Access 9:7107–7126

Laine S, Aila T (2016) Temporal ensembling for semi-supervised learning. arXiv:1610.02242

Liu W, Fu S, Zhou Y, Zha ZJ, Nie L (2021) Human activity recognition by manifold regularization based dynamic graph convolutional networks. Neurocomputing 444:217–225

Liu Z, Yao L, Bai L, Wang X, Wang C (2020) Spectrum-guided adversarial disparity learning. In: Proceedings of ACM SIGKDD, pp 114–124

Lopes A, Mendes-Moreira J, Gama J (2012) Semi-supervised learning: predicting activities in android environment. In: Workshop on ubiquitous data mining. Citeseer, vol 38

Loshchilov I, Hutter F (2016) Sgdr: stochastic gradient descent with warm restarts. arXiv:1608.03983

Lughofer E (2012) Hybrid active learning for reducing the annotation effort of operators in classification systems. Pattern Recogn 45(2):884–896

Lv M, Chen L, Chen T, Chen G (2018) Bi-view semi-supervised learning based semantic human activity recognition using accelerometers. IEEE Trans Mob Comput 17(9):1991–2001

Maas AL, Hannun AY, Ng AY (2013) Rectifier nonlinearities improve neural network acoustic models. In: Proc. icml, vol 30, p 3

Mahmud S, Tonmoy M, Bhaumik KK, Rahman A, Amin MA, Shoyaib M, Khan MAH, Ali AA (2020) Human activity recognition from wearable sensor data using self-attention. arXiv:2003.09018

Merritt P, Bi H, Davis B, Windmill C, Xue Y (2018) Big earth data: a comprehensive analysis of visualization analytics issues. Big Earth Data 2(4):321–350

Murad A, Pyun JY (2017) Deep recurrent neural networks for human activity recognition. Sensors 17(11):2556

Narasimman G, Lu K, Raja A, Foo CS, Aly MS, Lin J, Chandrasekhar V (2021) A*HAR: a new benchmark towards semi-supervised learning for class-imbalanced human activity recognition. arXiv:2101.04859

Noor MHM, Salcic Z, Wang K (2020) Ontology-based sensor fusion activity recognition. J Ambient Intell Humaniz Comput 11(8):3073–3087

Ordóñez FJ, Roggen D (2016) Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. Sensors 16(1):115

Reiss A, Stricker D (2012) Introducing a new benchmarked dataset for activity monitoring. In: 2012 16th international symposium on wearable computers. IEEE, pp 108–109

Rottmann M, Kahl K, Gottschalk H (2018) Deep bayesian active semi-supervised learning. In: 2018 17th IEEE international conference on machine learning and applications (ICMLA). IEEE, pp 158–164

Saeed A, Ozcelebi T, Lukkien J (2019) Multi-task self-supervised learning for human activity detection. Proc ACM Interact Mobile Wearable Ubiquitous Technol 3(2):1–30

Saito PT, Suzuki CT, Gomes JF, de Rezende PJ, Falcao AX (2015) Robust active learning for the diagnosis of parasites. Pattern Recogn 48(11):3572–3583

Settles B (2009) Active learning literature survey. University of Wisconsin-Madison

Shahmohammadi F, Hosseini A, King CE, Sarrafzadeh M (2017) Smartwatch based activity recognition using active learning. In: Proceedings of CHASE. IEEE, pp 321–329

Singh SP, Sharma MK, Lay-Ekuakille A, Gangwar D, Gupta S (2021) Deep convlstm with self-attention for human activity decoding using wearable sensors. IEEE Sens J 21(6):8575–8582

Slaton T, Hernandez C, Akhavian R (2020) Construction activity recognition with convolutional recurrent networks. Autom Constr 113:103138

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(56):1929–1958

Stikic M, Van Laerhoven K, Schiele B (2008) Exploring semi-supervised and active learning for activity recognition. In: 2008 12th IEEE international symposium on wearable computers. IEEE, pp 81–88

Subramanya A, Raj A, Bilmes JA, Fox D (2012) Recognizing activities and spatial context using wearable sensors. arXiv:1206.6869

Ullah M, Ullah H, Khan SD, Cheikh FA (2019) Stacked lstm network for human activity recognition using smartphone data. In: 2019 8th European workshop on visual information processing (EUVIP). IEEE, pp 175–180

Wan S, Qi L, Xu X, Tong C, Gu Z (2020) Deep learning models for real-time human activity recognition with smartphones. Mobile Netw Appl 25(2):743–755

Xu H, Pan Y, Li J, Nie L, Xu X (2019) Activity recognition method for home-based elderly care service based on random forest and activity similarity. IEEE Access 7:16217–16225

Yao S, Hu S, Zhao Y, Zhang A, Abdelzaher T (2017) Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In: Proceedings of the 26th international conference on world wide web, pp 351–360

Zeng M, Yu T, Wang X, Nguyen LT, Mengshoel OJ, Lane I (2017) Semi-supervised convolutional neural networks for human activity recognition. In: 2017 IEEE international conference on big data (Big Data). IEEE, pp 522–529

Zhang XY, Shi H, Zhu X, Li P (2019) Active semi-supervised learning based on self-expressive correlation with generative adversarial networks. Neurocomputing 345:103–113

Zhang M, Sawchuk AA (2012) USC-HAD: a daily activity dataset for ubiquitous activity recognition using wearable sensors. In: Proceedings of the 2012 ACM conference on ubiquitous computing, pp 1036–1043

Zhu Q, Chen Z, Soh YC (2018) A novel semisupervised deep learning method for human activity recognition. IEEE Trans Ind Inf 15(7):3821–3830

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.