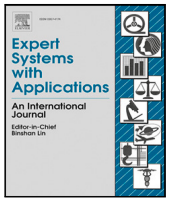
Expert Systems With Applications 237 (2024) 121427

Contents lists available at ScienceDirect

Expert Systems With Applications

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



A federated semi-supervised automatic sleep staging method based on relationship knowledge sharing

Bian Ma [a,b,c], Lijuan Duan [a,b,c,*], Zhaoyang Huang [d,e,**], Yuanhua Qiao [f], Bei Gong [a,b]

[a] Faculty of Information Technology, Beijing University of Technology, Beijing, 100124, China
[b] Beijing Key Laboratory of Trusted Computing, Beijing, 100124, China
[c] China National Engineering Laboratory for Critical Technologies of Information Security Classified Protection, Beijing, 100124, China
[d] Department of Neurology, Xuanwu Hospital, Capital Medical University, Beijing, 100053, China
[e] Beijing Key Laboratory of Neuromodulation, Beijing, 100053, China
[f] College of Applied Science, Beijing University of Technology, Beijing, 100124, China

ARTICLE INFO

Keywords:
Automatic sleep staging
Federated semi-supervised learning
Relationship knowledge
Prototype-contrastive learning
Pseudo-labeling optimization

ABSTRACT

Sleep staging is essential in assessing sleep quality and diagnosing sleep-related disorders, but the lack of labeled data impedes the development of automatic sleep staging models. Generally, institutions rely on semi-supervised approaches to enhance the utilization of their own unlabeled data. However, the task knowledge obtained from a limited amount of labeled data is often insufficient to guide the learning based on large amounts of unlabeled data, which may even lead to catastrophic forgetting and further degrade the performance of most existing methods. In this paper, we propose a novel strategy of building secure collaboration among multiple institutions, to achieve the implicit augmentation of labeled data and expansion of task knowledge for each participating institution by acquiring external knowledge from others. We adopt the Federated Learning (FL) to facilitate secure collaboration and propose a federated semi-supervised sleep staging method based on knowledge sharing, which enables the automatic scoring of sleep stages using only single-channel EEG data. The task knowledge in our method is contained in relationships, which exist naturally among sleep stages and can be extracted from both local labeled and unlabeled data. Furthermore, the knowledge sharing among participating institutions can be achieved by aligning the local relationships to the aggregated global relationships. Additionally, we employ prototype-contrastive learning to enhance the clarity of relationships extracted from labeled data, and propose pseudo-labeling optimization to generate reliable pseudo-labels for subsequent relationship extraction from unlabeled data. Our method is shown to be effective and outperforms compared methods in extensive experiments conducted on two publicly available datasets.

1. Introduction

Sleep is essential to human beings, and plays a pivotal role in their physical recuperation and emotional regulation (Seifpour et al., 2018). Numerous studies have demonstrated that inadequate sleep and sleep disorders can induce certain physical diseases (Grimaldi et al., 2016; Liew & Aung, 2021), and some psychological related diseases (Becker et al., 2018; Chellappa & Aeschbach, 2022). The number of people suffering from sleep disorders has increased dramatically in recent years, and sleep disorders have increasingly become a prominent public health problem (Benjafield et al., 2019; Hepsomali & Groeger, 2021; Phan et al., 2021). Sleep staging, revealing the physiological characteristics of related diseases, is an essential step in the process of sleep quality assessment and diagnosis of sleep-related diseases (Pan et al., 2022), and has become a focus of recent research. As mentioned in Assefa et al. (2015) that "'sleep' is not a single 'state of being'", the sleep can be divided into five stages: wake (W), non-rapid eye movement stage 1 (N1), N2, N3 and rapid eye movement (REM), according to the American Academy of Sleep Medicine (AASM) manual (Iber et al., 2007). In other words, sleep staging is a process to classify the polysomnography (PSG) or other types of recording signals into different sleep stages (Radhakrishnan et al., 2022).

At the outset, sleep staging was completed manually by sleep technicians who visually inspected the data and mapped them to particular sleep stages (Radhakrishnan et al., 2022; Thorey et al., 2019; Zhang

* Correspondence to: Beijing University of Technology, No.100, Pingleyuan, Chaoyang District, Beijing, China.
** Corresponding author.
E-mail addresses: mabian@emails.bjut.edu.cn (B. Ma), ljduan@bjut.edu.cn (L. Duan), drhuangzy@gmail.com (Z. Huang), qiaoyuanhua@bjut.edu.cn (Y. Qiao), gongbei@bjut.edu.cn (B. Gong).

https://doi.org/10.1016/j.eswa.2023.121427
Received 2 April 2023; Received in revised form 25 August 2023; Accepted 1 September 2023
Available online 6 September 2023

et al., 2019). However, it is time-consuming and labor-intensive, as well as being vulnerable to human error (Abdulla et al., 2019; Banluesombatkul et al., 2021; Stephansen et al., 2018). In an effort to free physicians from laborious manual processes, research into automatic sleep staging has been in full swing. The development of traditional machine learning methods has provided theoretical guidance and technical support for pristine automatic sleep staging research. Moreover, the classic machine learning algorithms such as decision tree learning (Hanaoka et al., 2001), K-means clustering and K-nearest neighbors (Güneş et al., 2010) and so on have been widely used to establish automatic sleep staging models. In addition, the framework of Hidden Markov models (HMMs) (Chen et al., 2015) and Extreme Gradient Boosting (XGB) classifier (Le-Dong et al., 2021) have also been adopted in automatic sleep staging. However, most traditional machine learning models rely heavily on manual feature extraction and lack generalization capabilities. Over the past few years, the improvement in available computing power, the increase of publicly accessible sleep datasets and the recent advances in artificial intelligence have greatly facilitated the research in deep sleep staging models, which can get knowledge directly from raw sleep data without extracting features manually in advance (Fiorillo et al., 2019; Yoo et al., 2022). Convolutional neural networks (CNN), one of the representative algorithms of deep learning, has made great achievements in sleep staging research and can complete automatic sleep staging without relying on the prior knowledge (Goshtasbi et al., 2022; Phan et al., 2019a; Sors et al., 2018; Supratak et al., 2017; Vaquerizo-Villar et al., 2021). The expertise of Recurrent Neural Network (RNN) in processing sequential data has contributed to fully exploiting the information contained in the sleep data (Phan et al., 2018, 2019b, 2021). With the attention mechanism having yielded brilliant results in the fields of natural language processing and computer vision, some attention-based models aimed at sleep staging have been put forward (Eldele et al., 2021; Phyo et al., 2022; Qu et al., 2020; Sun et al., 2020). Furthermore, emerging methods such as meta learning and transfer learning show great potential for improving automatic sleep staging (Banluesombatkul et al., 2021; Guillot & Thorey, 2021).

However, most of automatic sleep staging methods mentioned above are data-driven, requiring a great quantity of labeled data to acquire the relevant task knowledge. In practice, a significant challenge is the scarcity of available labeled data, as annotation work is time-consuming and labor-intensive (Banluesombatkul et al., 2021; Stephansen et al., 2018). Faced with the predicament, independent institutions have widely utilized semi-supervised learning (SSL) methods, which aim to simultaneously exploit labeled and unlabeled data to complete the classification task (Chapelle et al., 2006; Yang et al., 2021; Zhu & Goldberg, 2009). Unfortunately, the available labeled data within a single institution is often very limited, and the task knowledge extracted from it may be insufficient to guide the training process using a large number of unlabeled data. Worse still, a great quantity of unlabeled data within single institution may result in the problem of catastrophic forgetting, which further degrades the performance of the existing methods.

To overcome the deficiency of task knowledge resulting from the limited labeled data within a single institution, we propose to learn relevant external task knowledge by collaborating with other institutions to effectively expand knowledge base of each participating institution. Considering the high privacy and sensitivity of sleep data, it is impractical to realize the collaboration by sharing raw data directly, and we adopt the Federated Learning (FL) (Kairouz et al., 2019; McMahan et al., 2017; Yang et al., 2019) to build the secure collaboration among multiple institutions. FL is a privacy-preserving solution that allows participating institutions maintain control over their data and ensure that it is not being used without their consent. In addition, the increasing popularity of portable sleep monitoring devices has made sleep staging studies based on single-channel EEG signals trendy (Balaji et al., 2023; Neng et al., 2021; Zhang et al., 2023), and we also conduct our research based on single-channel EEG signals. To sum up, we adopt the FL to tackle the dearth of the labeled data within a single institution, and propose a federated semi-supervised sleep staging method to automatically and accurately score sleep stages using a small amount of labeled data and a large amount of unlabeled data from single-channel EEG. Inspired by the research on inherent relationships existing among different categories of diseases (Mathur & Dinakarpandian, 2012; Oerton et al., 2018), we infer that there are similar relationships among different sleep stages, which can reflect the structural task knowledge and are independent of individuals and other external factors. In our work, we extract the relationships among sleep stages from both labeled data and unlabeled data in each participating institution, and aggregate them to obtain the global relationships. By aligning its local relationships to the global relationships, each single institution acquires external task knowledge, which is then used to constrain local supervised training based on labeled data and guide unsupervised learning based on unlabeled data. The main contributions of our work are summarized as follows.

- In this paper, we propose a new strategy to cope with the problem of insufficient labeled sleep data within a single institution. The strategy utilizes the secure collaboration among multiple institutions to achieve the implicit expansion of labeled data and compensate for the inadequacy of task knowledge in each participating institution.
- In view of the insufficient information provided by the labeled single-channel EEG signal and the inherent confusion between sleep stages, we conduct the prototype-contrastive learning to obtain the more distinct relationships among sleep stages from labeled data.
- Given the infeasibility of extracting relationships directly from unlabeled data, we take both local supervised model and global model into account and propose the pseudo-labeling optimization to generate reliable pseudo-labels. The proposed optimization can effectively avoid both local overfitting and the problem of ignoring local features.

2. Related work

In this section, we focus on the research of semi-supervised automatic sleep staging and give an overview of federated learning (FL).

2.1. Semi-supervised automatic sleep staging

With the rapid acceleration of digitization, vast amounts of unlabeled data are generated daily due to the constant influx of patients. As mentioned earlier, sleep annotation is time-consuming and labor-intensive (Banluesombatkul et al., 2021; Stephansen et al., 2018), making it impossible for doctors to label all collected sleep data in a timely manner. As a result, each medical institution accumulates some labeled data, but a greater amount of unlabeled sleep data. Semi-supervised Learning (SSL) provides an idea to overcome the shortage of labeled data by exploiting the information contained in unlabeled data (Chapelle et al., 2006; Yang et al., 2021; Zhu & Goldberg, 2009). At present, research on automatic sleep staging based on semi-supervised methods is ongoing. In Zhang et al. (2022), SSL was introduced to explore the impact of varying degrees of OSA severity on sleep staging and CMS2-Net was proposed. However, the research was based on multi-channel and multi-modal data, and its practical implementation in scenarios that only involve single-channel information is bound to lead to a decline in model performance.

Pseudo-labeling, a representative method of SSL, has been employed in sleep staging. For instance, Zhang et al. proposed a model named SHNN and utilized the pseudo-labeling to implement semi-supervised automatic sleep staging (Zhang et al., 2023). The combination of pseudo-labeling approach with existing automatic sleep staging models such as CNN (Sors et al., 2018), DeepSleepNet (Supratak et al., 2017) and CCRRSleepNet (Neng et al., 2021) has been proven to be effective in accomplishing semi-supervised sleep staging (Zhang et al., 2023).

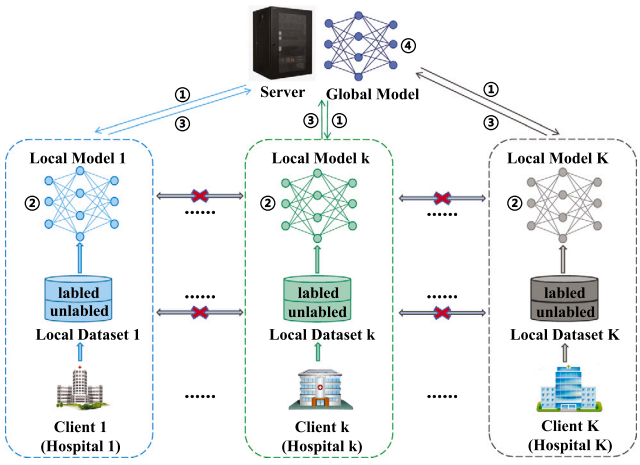


Fig. 1. The framework and pipelines of federated learning. Step ①: The server distributes the initial global model to each client; Step ②: Each client updates the received model using its private datasets that consist of labeled samples and unlabeled samples; Step ③: Each client sends its updated model to the server; Step ④: The server aggregates the local models from each client into the global model. The server is assumed to be trusted and the steps are repeated until the global model converges or reaches the max training rounds.

Unfortunately, most of the current methods that utilize the pseudo-labeling approach heavily depend on the task knowledge obtained from the labeled data. Consequently, when the amount of labeled data is small, or there is a substantial difference between labeled data and unlabeled data, catastrophic forgetting and degradation of model performance may occur.

2.2. Federated learning (FL)

Different from the traditional centralized learning approaches, Federated Learning (FL) (Kairouz et al., 2019; McMahan et al., 2017; Yang et al., 2019) enables model training to be conducted locally on client devices rather than on a central device. As a novel learning scheme, FL aims to build a federated model by enabling collaboration among different clients in a privacy-preserving manner. Specifically, the collaboration among clients can be accomplished by sharing the weights or gradients of the local models, without additional exchange of their sensitive raw data. Federated Averaging (FedAvg) algorithm (McMahan et al., 2017) is widely used in FL that averages the local model weights to obtain a global model for the subsequent round of training. We adopt the FedAvg method for aggregation in this paper.

There are two classic frameworks in FL, the server-client (CS) framework and the Peer-to-Peer (P2P) framework. In the CS framework, a central server communicates with all clients, while clients communicate with each other directly in the P2P framework. Undoubtedly, in the P2P framework, the client side bears a significant burden of computational and communication overheads. Therefore, we adopt the CS framework in our work, as shown in Fig. 1.

The study in Lou et al. (2021) demonstrates the feasibility of applying the FL framework to sleep staging studies. Moreover, FL has been widely studied and applied in various areas such as computer vision (Liu et al., 2021a), natural language processing (Deng et al., 2021), recommend systems (Yi et al., 2021) and so on. However, much of the work is premised on the unrealistic assumption that all the data from clients are labeled. In fact, relatively little attention has been devoted to FL in weakly supervised scenarios (Jin et al., 2020), and research on automatic sleep staging using federated semi-supervised methods is even rarer. In our work, we aim to utilize federated learning to address the issue of insufficient labeled sleep data in independent institutions, thus further broadening the research boundaries of federated semi-supervised learning in the field of automatic sleep staging.

3. Methodology

In this section, we focus on the proposed method. We introduce the problem formulation in Section 3.1 and provide an overview of the method in Section 3.2. We also delve into the details of local training on each client, and expound the local supervised training and local unsupervised learning in Section 3.3 and Section 3.4, respectively.

3.1. Problem formulation

The framework we propose consists of a credible single server and K independent clients. The server does not store any data, while each client holds its private dataset comprising of both labeled and unlabeled single-channel EEG signals. We denote the dataset owned by the k-th client as $D^k = D^{k,L} \cup D^{k,U}$, where $D^{k,L}$ and $D^{k,U}$ represent the labeled and unlabeled subsets, and $k = 1, \dots, K$. Specifically, $D^{k,L} = \{(x_i^L, y_i)\}_{i=1}^{N^{k,L}}$, where $N^{k,L}$ is the number of labeled data on client k, x_i^L is the i-th labeled instance and y_i is its corresponding label. And $y_i \in \mathbb{R}^C$, where C represents the total number of categories of sleep stages, and $C = 5$ in our study. The unlabeled data on client k is $D^{k,U} = \{x_i^U\}_{i=1}^{N^{k,U}}$, where x_i^U is the i-th unlabeled instance, $N^{k,U}$ is the number of unlabeled samples and $N^{k,L} \ll N^{k,U}$. Our objective is to utilize both labeled and unlabeled data from each client collaboratively to train a global federated model M^G, whose parameters are denoted by θ^G.

3.2. The overview of the proposed method

In this paper, we propose an efficient federated semi-supervised sleep staging method, as shown in Fig. 2, to enable secure collaboration among independent institutions and fully utilize information from the unlabeled data on each client. By aligning the local task knowledge of clients to the global knowledge, where the knowledge is mainly contained in the relationships among sleep stages, our method facilitates knowledge sharing among clients. The global knowledge is further used to constrain the supervised learning and guide the unsupervised learning on each client.

In the proposed framework, the server is responsible for distributing and aggregating the global model M^G, global relationship matrix R^G and global prototype P^G in each federated round. The clients focus on their local training, and update their respective local elements (M^k, R^k, and P^k, where $k = 1, \dots, K$), and then send them to the server for global aggregation.

- **Server: distributing and aggregating.** At the start of a new federated learning round (e.g., round t), the server takes the global model M_{t-1}^G, global relationship matrix R_{t-1}^G and global prototype P_{t-1}^G aggregated at the end of last round (i.e. round $t-1$) as the initial elements, and sends them to each client simultaneously. The clients conduct local training using their private data and correspondingly obtain updated elements M_t^k, R_t^k and P_t^k. And then these updated elements are sent to the server for aggregation. At the end of round t, the classic FedAvg (McMahan et al., 2017) approach is adopted to aggregate the local models to obtain the updated global model, i.e. $M_t^G = \frac{1}{K}\sum_{k=1}^{K} M_t^k$, where $k = 1, \dots, K$. The global relation matrix R^G and global prototype P^G are obtained in the same way. It follows that M^G, R^G and P^G necessarily contain the knowledge extracted from all clients. The updated M_t^G, R_t^G and P_t^G are then sent to clients again in the next round of training. The iterative process continues until either the federated model converges or reaches the maximum federated round.
- **Clients: conducting local training.** At the start of the t-th federated round, each client receives the aggregated global federated model M_{t-1}^G from the server, and uses it as the initial local model for conducting local training with its private data for E epochs. Along with M_{t-1}^G, clients also receive the global relation matrix R_{t-1}^G and the global

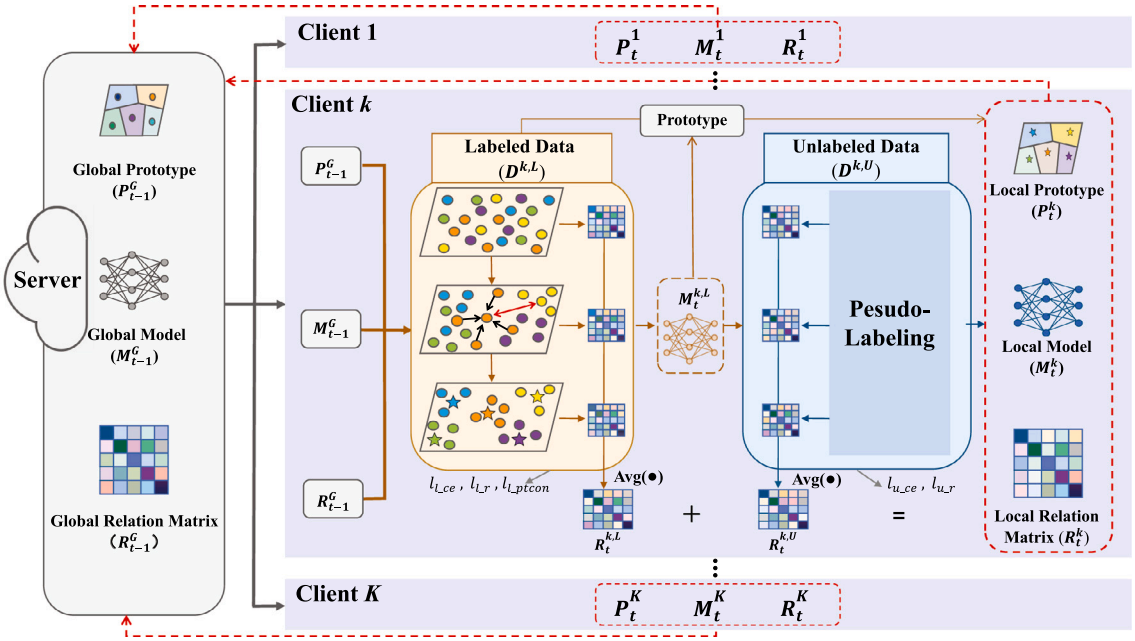


Fig. 2. The framework of the proposed method. **Server:** distributes the initial global model M_{t-1}^{G}, global relation matrix R_{t-1}^{G} and global prototype P_{t-1}^{G} at the start of a new round t, and aggregates the local models M_t^k, local relation matrices R_t^k and local prototype P_t^k at the end of t-th round, $k = 1, \ldots, K$. **Clients:** conduct the local training after receiving global materials. (1) Local supervised training: conducts supervised training and prototype-contrastive learning to obtain the more distinct relationships from labeled data, and obtains corresponding labeled relation matrix $R_t^{k,L}$, local prototype P_t^k and the updated supervised model $M_t^{k,L}$. (2) Local unsupervised learning: generates pseudo-labels for unlabeled data, and continues to train the supervised model to get the final updated local model M_t^k and obtain unlabeled relation matrix $R_t^{k,U}$ based on the pseudo-labels.

prototype P_{t-1}^{G}, enabling each client to access external knowledge from others for its local training.

The local training on clients can be divided into two phases: the supervised learning using labeled data and subsequent unsupervised learning using unlabeled data. Concretely, for client k, in the supervised learning phase, the initial model M_{t-1}^{G} is updated using the labeled data to obtain a supervised model denoted as $M_t^{k,L}$. In addition, the relation matrix $R_t^{k,L}$ extracted from labeled data and the local prototype P_t^k can also be obtained in this phase. In the subsequent unsupervised learning phase, both the local supervised model $M_t^{k,L}$ and the global model M_{t-1}^{G} are taken into account to generate reliable pseudo-labels for unlabeled data. The generated pseudo-labels are then used to update the local supervised model $M_t^{k,L}$ and extract the relationships, yielding the final updated local model M_t^k and unlabeled relation matrix $R_t^{k,U}$. Notably, the global knowledge contained in R_{t-1}^{G} not only constrains the supervised training, but also guides the unsupervised training. The following sub-sections provide details of above processes.

3.3. Relationship extraction from labeled data and prototype-contrastive learning

In this section, we focus on the local supervised learning based on labeled data, and elaborate the process of extracting relationships from labeled data and aligning them to the global relationships.

• **Relationships extraction and alignment:** We believe certain inherent relationships exist among sleep stages that are independent of external factors and can effectively reflect the structural knowledge. Following the approach in Liu et al. (2021b), we can obtain the stage relationships from the class ambiguity captured by the model. In the t-th round, the extracted relationships from labeled data on client k can be expressed in matrix form as $R_t^{k,L} = [s_{t(1)}^{k,L}, \ldots, s_{t(C)}^{k,L}]$, where $s_{t(c)}^{k,L}$ represents the soft label distribution for the c-th stage and effectively captures the relationship between the c-th stage and all other stages:

$$s_{t(c)}^{k,L} = \mathrm{softmax}(v_{t(c)}^{k,L}/\tau_1)$$
$$\text{with} \quad v_{t(c)}^{k,L} = \frac{1}{N_{(c)}^{k,L}} \sum_{i=1}^{N^{k,L}} \mathbb{1}_{[y_i=c]} \hat{f}(x_i^L), \tag{1}$$

where τ_1 is the temperature of the softened softmax function and $\tau_1 > 1$ (Dou et al., 2020), $v_{t(c)}^{k,L} \in \mathbb{R}^C$ is the per-category mean feature vectors and reflects the knowledge captured by model on sleep stage c. $N_{(c)}^{k,L}$ is the number of labeled samples belong to sleep stage c, $\mathbb{1}_{[\cdot]}$ is an indicator function to indicate whether the i-th labeled sample belongs to class c, and $\hat{f}$ represents the model without the softmax layer with $f = M_{t-1}^{G}$ here.

As mentioned before, the global relation matrix effectively integrates the task knowledge from each client and provides constraint on the local supervised training. The constraint can be realized by aligning the local relationships to the global relationships, which can be achieved by minimizing the Kullback–Leibler (KL) divergence between the global relation matrix and local relation matrix as follows:

$$l_{l\_r} = (\mathcal{L}_{KL}(R_{t-1}^{G} \parallel R_t^{k,L}) + \mathcal{L}_{KL}(R_t^{k,L} \parallel R_{t-1}^{G}))/2, \tag{2}$$

where $\mathcal{L}_{KL}(\cdot)$ represents the KL-divergence, the representation can be expressed as:

$$\mathcal{L}_{KL}(R_{t-1}^{G} \parallel R_t^{k,L}) = \sum_i R_{t-1(i)}^{G} \log \frac{R_{t-1(i)}^{G}}{R_{t(i)}^{k,L}}, \tag{3}$$

where i is the i-th entry of the relation matrix.

• **Prototype-contrastive learning:** The concept of prototype representation, or prototype for short, was first introduced in Snell et al. (2017) to solve the problem of few-shot classification, and has been applied in federated scenarios (Kim et al., 2022; Michieli & Ozay, 2021; Tan et al., 2022). The main idea of prototypical networks is that each class is associated with a specific prototype that serves as a representative point in the embedding space. Therefore, the predicted class of a new sample is the one whose prototype is closest to the sample in the embedding space.

Due to the limited labeled single-channel EEG information and the confusion among sleep stages, the extracted relationship may not be clear-cut. To obtain more distinct relationships, we introduce the prototype and conduct prototype-contrastive learning. This involves shortening the distance between the local prototype and the global prototype of the same sleep stage, while widening the distance from

prototypes of other stages. A point to make is that in our work, we obtain each local prototype only based on labeled data due to its high certainty. We extract prototypes twice: the first is to constrain local supervised training, and the second is to provide a more accurate local prototype for global aggregation. In addition, different from most methods, we only extract prototypes based on labeled data whose predictions by the model are consistent with corresponding true labels, while skipping instances with incorrect predictions. During the supervised training, the first prototype of stage c can be obtained as follows:

$$P^k_{t(c)} = \frac{1}{|N^{k,L'}_{(c)}|} \sum_{\hat{y}_i = y_i} \mathbb{1}_{[y_i=c]} f_{M^G_{t-1}}(x_i), \tag{4}$$

where $N^{k,L'}_{(c)}$ is the set of samples, whose corresponding predictions $\hat{y}_i$ by the model M^G_{t-1} are consistent with true labels of c (i.e. $\hat{y}_i = y_i$, $y_i = c$). $f_{M^G_{t-1}}(\cdot)$ denotes the function with parameters of M^G_{t-1} to obtain the embedding representations.

Building on the work of Sohn (2016) on N-pair loss and the model contrastive loss proposed in Li et al. (2021), we define the prototype-contrastive loss term as follows:

$$l_{\_ptcon} = (\sum_{c=1}^{C} \log \frac{dis_{(c)}(+^*)}{dis_{(c)}(+^*) + dis_{(c)}(-^*)})/C, \tag{5}$$

with

$$dis_{(c)}(+^*) = \exp \frac{\mathrm{MSE}(P^k_{t(c)}, P^G_{t-1(c)})}{\tau_2}, \tag{6}$$

$$dis_{(c)}(-^*) = \sum_{j=1, j\neq c}^{C} \exp \frac{\mathrm{MSE}(P^k_{t(c)}, P^G_{t-1(j)})}{\tau_2}, \tag{7}$$

where MSE(·) (Mean Square Error) measures the distance between prototypes, τ_2 is also denotes a temperature parameter and $\tau_2 < 1$.

Further, after local supervised learning, we extract prototypes again based on the local updated supervised model $M^{k,L}_t$ to obtain a more accurate local prototype for global aggregation. The local prototype of client k can be expressed as $P^k_t = [P^k_{t(1)}, \ldots, P^k_{t(c)}, \ldots, P^k_{t(C)}]$, and $C = 5$ because the sleep stages are divided into five classes in our study.

• **Objective of local supervised training:** Finally, the objective loss of the local supervised training on each client can be summarized as follows:

$$loss_l = \alpha l_{\_ce} + \beta l_{\_r} + \gamma l_{\_ptcon}, \tag{8}$$

where $l_{\_ce}$ is a class-aware loss proposed in Eldele et al. (2021), $l_{\_r}$ is the consistency loss between local labeled relationships and global relationships, and $l_{\_ptcon}$ denotes prototype-contrastive loss. α, β, γ are hyper-parameters used to balance the contribution of different loss terms.

In fact, the number of samples in different sleep stages varies greatly, and using a standard multi-class cross-entropy function may lead to a biased model that prefers the majority classes. To deal with the issue of class imbalance, we adopt the same weighted multi-class cross entropy function as in Eldele et al. (2021), which can be expressed as:

$$l_{\_ce} = -\frac{1}{N^{k,L}} \sum_{c=1}^{C} \sum_{i=1}^{N^{k,L}} w_{(c)} y_{i(c)} \log(\hat{y}_{i(c)}), \tag{9}$$

where $N^{k,L}$ is the number of labeled samples, $y_{i(c)}$ is the label of the i-th instance belonging to sleep stage c while $\hat{y}_{i(c)}$ is the correspond prediction, $w_{(c)}$ is the weight assigned to sleep stage c and can be obtained by:

$$w_{(c)} = \mu_{(c)} \cdot \max(1, \log(\mu_{(c)} N^{k,L} / N^{k,L}_{(c)})), \tag{10}$$

where $N^{k,L}_{(c)}$ is the number of labeled samples in the c-th sleep stage, $\mu_{(c)}$ is used to determine the distinctness of sleep stage c, and $\mu_{(c)} = h/C$ where h is a hyperparameter that varies for different sleep stages (Eldele et al., 2021), and $C = 5$.

So far, the initial local model M^G_{t-1} of round t is updated based on labeled data, resulting in the updated supervised model denoted as $M^{k,L}_t$. At the same time, the local labeled relation matrix $R^{k,L}_t$ and the local prototype P^k_t can be obtained.

3.4. Relationship extraction from unlabeled data based on pseudo-labeling optimization

In this section, we mainly present the proposed pseudo-labeling optimization and the subsequent relationships extraction and alignment.

In this phase, the local supervised model $M^{k,L}_t$ continues to be updated using the unlabeled data, with the guidance of global knowledge contained in R^G_{t-1}. The guidance can be realized by aligning the relationships extracted from the unlabeled data to the global relationships. However, the absence of true labels makes it infeasible to directly extract the relationships among sleep stages from unlabeled data. Moreover, as complex physiological signals, EEG signals are susceptible to even the faintest change. Therefore, it is challenging to perform unsupervised learning based on the consistency of predicted results from the original and transformed data, as is typically done for image data. To overcome this challenge, we adopt the pseudo-labeling mechanism, which does not require additional processing of the data. Traditionally, pseudo-labels for unlabeled data are generated using either the local supervised model or the global model, which can lead to problems such as local overfitting or ignoring of local features. In our work, we optimize the pseudo-labeling algorithm in the following way.

• **Pseudo-labeling optimization:** We incorporate both the local supervised model and the global model in the generation of pseudo-labels, effectively mitigating the potential issues mentioned earlier. The feasibility of using dropout Bayesian network (Gal & Ghahramani, 2016) to approximate the uncertainty of model prediction has been confirmed in Liu et al. (2021b). We extend this method in our work to generate reliable pseudo-labels with high confidence and low uncertainty for unlabeled data. The detailed process is as follows.

For the same unlabeled input x^U, T-time forward propagation is performed under random dropout using the local supervised model $M^{k,L}_t$ and the global model M^G_t. The corresponding predicted probability vectors can be obtained and denoted as $\{\mathrm{p}^{k,U}_j\}^T_{j=1}$ and $\{\mathrm{p}^{G,U}_j\}^T_{j=1}$, respectively. Thus, the uncertainty of the predictions $\mathbf{u}$ can be obtained by the following formulation:

$$\mathbf{u} = -\sum_{c=1}^{C} \bar{\mathrm{p}}^U_{(c)} \log(\bar{\mathrm{p}}^U_{(c)}),$$
$$\text{with} \quad \bar{\mathrm{p}}^U_{(c)} = \frac{1}{2T} \sum_{j=1}^{T} (\mathrm{p}^{k,U}_{j(c)} + \mathrm{p}^{G,U}_{j(c)}), \tag{11}$$

where $\mathrm{p}^{k,U}_{j(c)}$ and $\mathrm{p}^{G,U}_{j(c)}$ are the values of the c-th class of $\mathrm{p}^{k,U}_j$ and $\mathrm{p}^{G,U}_j$, respectively. Accordingly, the pseudo-labels with high confidence and low uncertainty are generated by:

$$\tilde{y} = \mathrm{argmax}(\mathrm{q}^{k,U} \cdot \mathbb{1}_{[(\mathbf{u} \leq h_u) \cdot (\max(\mathrm{q}^{k,U}) \geq h_c)]}), \tag{12}$$

where $\mathrm{q}^{k,U}$ is the predicted probability generated by the local supervised model $M^{k,L}_t$, which is waiting to be updated with the unlabeled data. And h_u and h_c represent the thresholds for uncertainty and confidence of $\mathrm{q}^{k,U}$, respectively.

In the subsequent training process, we regard the obtained pseudo-labels as the true labels of the unlabeled data, and compute the multi-class cross entropy loss as follows:

$$l_{u\_ce} = -\frac{1}{N^{k,U}} \sum_{c=1}^{C} \sum_{i=1}^{N^{k,U}} w^U_{(c)} \tilde{y}_{i(c)} \log(\hat{y}^U_{i(c)}), \tag{13}$$

where $N^{k,U}$ is the number of generated pseudo-labels, $\tilde{y}_{i(c)}$ is the pseudo-label of i-th instance while $\hat{y}^U_{i(c)}$ is corresponding prediction, and $w^U_{(c)}$ is the weight as mentioned in Eq. (10).

• **Relationship extraction and alignment:** In addition, we can obtain the relation matrix $R^{k,U}_t$ from unlabeled data according to Eq. (1), and the guidance from global knowledge can be guaranteed by relationships alignment:

$$l_{u\_r} = (\mathcal{L}_{KL}(R^G_{t-1} \parallel R^{k,U}_t) + \mathcal{L}_{KL}(R^{k,U}_t \parallel R^G_{t-1}))/2. \tag{14}$$

• **Objective of local unsupervised training:** The overall objective of the local unsupervised learning on client k is:

$$loss_u = \delta l_{u\_ce} + \eta l_{u\_r}, \tag{15}$$

where δ and η are hyperparameters to balance the effect of two constraints. Using Eq. (15), we continue to update the obtained supervised model $M^{k,L}_t$ to get the final local model M^k_t. In addition, the local relation matrix is updated by $R^k_t = (R^{k,L}_t + R^{k,U}_t)/2$, which takes into account both the relationships extracted from the labeled data and the unlabeled data.

The overall local training process is outlined in Algorithm 1.

Algorithm 1: The local training of each client

inputs : current federated round t, client k's local dataset $D^k = D^{k,L} \cup D^{k,U}$, initial local model M^G_{t-1}, global relation matrix R^G_{t-1}, global prototype P^G_{t-1}

outputs: Updated local model $\boldsymbol{M^k_t}$, updated local relation matrix $\boldsymbol{R^k_t}$ and updated local prototype $\boldsymbol{P^k_t}$

for t<maximum, client $k = 1, \ldots, K$ in parallel **do**
  for each local epoch $e = 1, 2, \ldots, E$ **do**
    $R^{k,L(e)}_t, M^{k,L(e)}_t \leftarrow$ **LocalSupervisedTraining** $(D^{k,L}, M^G_{t-1}, R^G_{t-1}, P^G_{t-1})$;
    $P^{k(e)}_t \leftarrow$ **GetLocalPrototype**$(D^{k,L}, M^{k,L(e)}_t)$;
    $R^{k,U(e)}_t, M^{k(e)}_t \leftarrow$ **LocalUnsupervisedTraining** $(D^{k,U}, M^{k,L(e)}_t, M^G_{t-1}, R^G_{t-1}, P^G_{t-1})$;
    $R^{k(e)}_t \leftarrow (R^{k,L(e)}_t + R^{k,U(e)}_t)/2$;
    $e \leftarrow e + 1$;
  end
  $\boldsymbol{R^k_t} \leftarrow (\sum_{e=1}^{E} R^{k(e)}_t)/E$;
  $\boldsymbol{P^k_t} \leftarrow (\sum_{e=1}^{E} P^{k(e)}_t)/E$;
  $\boldsymbol{M^k_t} \leftarrow M^{k(E)}_t$;
  $t \leftarrow t + 1$;
end
LocalSupervisedTraining $(D^{k,L}, M^G_{t-1}, R^G_{t-1}, P^G_{t-1})$ **:**
  l_{l_ce} is computed using Eq.(9);
  l_{l_r} is computed using Eq.(2) and $R^{k,L}_t$ is obtained;
  $P^k_t \leftarrow$ **GetLocalPrototype**$(D^{k,L}, M^G_{t-1})$;
  l_{l_ptcon} is computed using Eq.(5) based on P^k_t;
  $M^{k,L(e)}_t \leftarrow M^G_{t-1}$ is updated based on $D^{k,L}$ by using Eq.(8);
GetLocalPrototype$(D^{k,L}, M^{k,L}_t)$ **:**
  choose samples predicted by $M^{k,L}_t$ correctly;
  P^k_t is obtained according to Eq.(4) based on chosen samples;
LocalUnsupervisedTraining $(D^{k,U}, M^{k,L(e)}_t, M^G_{t-1}, R^G_{t-1}, P^G_{t-1})$ **:**
  pseudo-labels are generated using Eq.(12);
  l_{u_ce} is computed using Eq.(13);
  l_{u_r} is computed using Eq.(14) and $R^{k,U(e)}_t$ is obtained;
  $M^{k(e)}_t \leftarrow M^{k,L(e)}_t$ is updated based on $D^{k,U}$ by using Eq.(15);

4. Experiments

In this section, we first introduce the datasets and metrics used to evaluate our method. Then, we describe the experimental setup and implementation details. Finally, we conduct a series of experiments to mainly answer the following questions:

•**RQ1**: Is it beneficial for all participants to participate in collaboration? (Section 4.3.1)

•**RQ2**: Can the proposed method effectively utilize unlabeled data and complete the task of automatic sleep staging? (Section 4.3.2)

•**RQ3**: Is the proposed prototype-contrastive learning helpful to extract more distinct relationships among sleep stages from data? (Section 4.3.3)

•**RQ4**: Is the pseudo-labeling optimization we proposed superior? (Section 4.3.4)

•**RQ5**: Whether our method outperform the existing semi-supervised automatic sleep staging methods based on raw single-channel EEG data? (Section 4.3.5)

•**RQ6**: Is each module of our method effective? (Section 4.3.6)

4.1. Datasets and evaluation metrics

4.1.1. Datasets

In our work, we evaluated the proposed method using raw single-channel EEG data from two public datasets, namely, Sleep Heart Health Study (SHHS) and Sleep-EDF-78.

SHHS. The SHHS is a multi-center cohort study designed to investigate the relationship between sleep-related breathing and an increased risk of coronary heart disease, stroke, all-cause mortality, and hypertension (Quan et al., 1997; Zhang et al., 2018). The SHHS dataset consists of two subsets: SHHS1 and SHHS2. SHHS1 contains 6,441 male and female subjects aged 40 years and older, and SHHS2 records the second polysomnogram of 3,295 of these participants. Following the selection criteria outlined in previous studies (Eldele et al., 2021; Fonseca et al., 2017), we selected 329 subjects with regular sleep from the SHHS1 subset to carry out our work. In more details, we utilized the raw data from the C4-A1 EEG channel with a sampling rate of 125 Hz.

Sleep-EDF-78. We also used an extended version of the Sleep EDF Database from the PhysioBank, referred to as Sleep-EDF-78, which contains data files for 78 subjects (Goldberger et al., 2000; Kemp et al., 2000). To be more specific, it contains 197 whole-night PolySomnoGraphic sleep recordings, among which 153 Sleep Cassette (SC*) files were collected in a study about the effects of age on sleep in healthy Caucasians aged from 25 to 101 years. The remaining 44 Sleep Telemetry (ST*) files were used to investigate the impact of temazepam on sleep in 22 Caucasian males and females who were not taking any other medication. Consistent with previous works (Eldele et al., 2021; Phan et al., 2019a; Supratak et al., 2017), we conducted our experiments using data from a single EEG channel of 153 SC* files. Specifically, we employed the raw data from the Fpz-Cz channel with a sampling frequency of 100 Hz to evaluate the effectiveness of the proposed method.

As with most existing studies (Eldele et al., 2021; Phan et al., 2019a; Supratak et al., 2017), for the aforementioned two datasets, we initially removed any UNKNOWN and MOVEMENT stages, and merged N3 and N4 stages into a single N3 stage according to the AASM manual (Iber et al., 2007). The epochs in both datasets were categorized into five stages: Wake, N1, N2, N3, and REM. In addition, we only included the 30 min of wake periods before and after the entire sleep phase, as our focus was solely on the study of sleep stages.

Table 1 summarizes the number of each sleep stage from all subjects in the two public datasets.

4.1.2. Metrics

In our work, we evaluated the performance of our method using multiple metrics. Specifically, we used the per-class F1-score to assess the performance of the method for each sleep stage. The accuracy (ACC), macro-averaged F1-score (MF1) and Cohen Kappa (κ) were also

Table 1
Number of each sleep satge in two public datasets.

Dataset	Number of each stages					Total
	W	N1	N2	N3	REM	
SHHS	46319 (14.3%)	10304 (3.2%)	142125 (43.7%)	60153 (18.5%)	65953 (20.3%)	324854
Sleep-EDF-78	65951 (33.7%)	21522 (11.0%)	69132 (35.4%)	13039 (6.7%)	25835 (13.2%)	195479

used to appraise the overall performance of our method. The following are the definitions of the metrics we referred to:

$$F1 = \frac{2 * P * R}{P + R}, \quad \text{with} \quad P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN}, \tag{16}$$

where TP, FP, FN represent the number of true positive, false positive and false negative samples, respectively. And MF1 can be expressed as:

$$MF1 = \frac{1}{C}\sum_{c=1}^{C} F1_{(c)}, \tag{17}$$

where $F1_{(c)}$ is the F1-score of the c-th sleep stage and C is the number of sleep stages. In addition,

$$ACC = \frac{TP + TN}{M}, \tag{18}$$

where TN represents the number of true negative samples, M is the total number of samples and $M = TP+FP+TN+FN$. The formulation of κ is:

$$\kappa = \frac{ACC - P_e}{1 - P_e}, \tag{19}$$

where P_e is the accidental consistency error and

$$P_e = \frac{(TP + FP) \times (TP + FN)}{M^2}. \tag{20}$$

4.2. Experiment setup and implementation

In our study, we divided the two datasets by individual subjects, with 70% of them serving as the training set and the remaining 30% as the test set. We set five clients, i.e. $K = 5$, and further randomly divided the training set into five non-overlapping subsets. Table 2 summarizes the results of dividing the two datasets. Subsequently, we conducted a series of experiments using different proportions of the labeled data on each client, respectively.

• **Backbone net:** AttnSleepNet (Eldele et al., 2021), an advanced attention-based model designed for automatic sleep stage classification using single-channel EEG signals, was adopted as the foundational backbone network in our proposed framework. It is composed of three main components: feature extraction, temporal context encoder (TCE) and classification. Specifically, the feature extraction module comprises a multi-resolution CNN (MRCNN) and adaptive feature recalibration (AFR). The MRCNN consists of two branches of convolutional layers with different kernel sizes in their first layers (i.e. 400 and 50), enabling the extraction of features corresponding to both low and high frequencies. The AFR captures inter-dependencies inherent in the extracted features and adaptively selects the most discriminative features by employing a residual squeeze and excitation block. The design of the TCE module focuses on leveraging a multi-head attention mechanism with causal convolutions to capture temporal dependencies among the extracted features. In addition, a class-aware loss function is proposed to address the issue of class imbalance in the classification module. Extensive experiments have validated the feasibility and superiority of the model.

• **Parameter settings:** For our experiments, we set the maximum federated round to 120, with the local training epoch (e) set to 2. Adam optimizer with momentum of 0.9 and 0.99 was adopted in each local training and the batch size was set to 128 for both local supervised and local unsupervised learning. We performed 10-time forward propagation under random dropout to generate reliable pseudo-labels for unlabeled data. In addition, the temperature parameter τ_1 and τ_2 were set to 2.0 and 0.8, respectively, and other hyper-parameters could be adjusted according to the actual situation. It is worth mentioning that in the first 20 rounds, we conducted the federated learning based only on the labeled data of each client to obtain relevant task information, so as to provide reliable global knowledge for subsequent federated semi-supervised learning.

4.3. Experimental results

4.3.1. Beneficial analysis for clients

"Why should I participate in Federated Learning?" Indeed, independent institutions may be reluctant to collaborate with others because they may not benefit from it, or worse, it may lead to a waste of resources and the risk of privacy leakage. In this section, we used ACC, MF1 and κ as metrics to verify that each institution participating in the federated collaboration can benefit from it.

On the one hand, we compared the models trained independently on each client using its private data, with the model trained through federated learning using the same data. Specifically, we conducted independently supervised learning on each client using their own private labeled data, and denoted the obtained models as 'Ck_L' with $k = 1, \ldots, 5$. The model obtained by federated learning using the same labeled data was marked as 'Fed_ L'. The performance comparison of these models is presented in Fig. 3, and we can see that the federated model (Fed_ L) outperforms all client models in terms of ACC, MF1, and κ, across different proportions of labeled data in both two public datasets. It illustrates that although the private raw data cannot be shared directly, the federated global model can obtain more comprehensive task knowledge by aggregating the models of each client, leading to higher performance. The advantages of federated learning are fully demonstrated.

On the other hand, using the proposed federated semi-supervised method, we compared the performance of the obtained federated global model and the local models. To be concise, Fig. 4 only shows the results for the case where the labeled data on each client accounts for 30%. As can be seen from Fig. 4, the performance of the federated global model, marked as 'Global', is superior to all of the local models, denoted as 'Client$k$' with $k = 1, \ldots, 5$. Especially in the Sleep-EDF-78 dataset, the superiority is more conspicuous. We attribute it to the fact that each client holds a small amount of data, with the labeled data being even more scarce, making the local acquired knowledge more biased towards its own data and increasing the likelihood of overfitting in the local model. While the federated learning can effectively mitigate the circumstances by expanding the local data implicitly and facilitating knowledge sharing between clients. The resulting federated global model is distributed to all participants, providing each with a better performing model for their respective classification tasks. Therefore, the benefits of federated collaboration for participants are thus further confirmed. In particular, the benefits are more significant in the case where the labeled data is more limited.

4.3.2. Performance analysis of our method

In this section, we analyzed the effectiveness of the proposed method, specifically, whether the method can effectively utilize unlabeled data to achieve automatic sleep staging, especially when labeled data accounts for a small proportion.

In our work, we first paid close attention to the scenario where the labeled data on each client accounted for less than half of the total data. We compared our method with federated supervised learning only based on labeled data of each client ('Fed_L' method for short), in terms of ACC, MF1, and κ. Table 3 summarizes the comparison results,

Table 2
Summary of two public datasets.

Datasets	EEG channle	Sampling rate	# Total subjects	# Subjects for training on each client	# Subjects for testing
SHHS	C4-A1	125 Hz	329	46	99
Sleep-EDF-78	Fpz-Cz	100 Hz	153	22	43

Table 3
Performance comparison between the proposed method and federated supervised method based on labeled data only (Fed_L).

Dataset1: SHHS								
Overall metrics	10% labeled-subjects		20% labeled-subjects		30% labeled-subjects		40% labeled-subjects	
	Fed_L	Ours	Fed_L	Ours	Fed_L	Ours	Fed_L	Ours
ACC (%)	75.4	**76.9**	76.3	**80.6**	78.1	**80.9**	79.0	**81.6**
MF1 (%)	65.0	**66.3**	65.5	**69.5**	68.4	**70.3**	69.8	**71.8**
κ	0.658	**0.677**	0.662	**0.723**	0.693	**0.727**	0.706	**0.737**
Dataset2: Sleep-EDF-78								
Overall Metrics	10% labeled-subjects		20% labeled- subjects		30% labeled- subjects		40% labeled- subjects	
	Fed_L	Ours	Fed_L	Ours	Fed_L	Ours	Fed_L	Ours
ACC (%)	74.5	**75.2**	75.1	**77.6**	76.2	**78.4**	77.4	**79.3**
MF1 (%)	67.1	**68.5**	68.0	**70.3**	69.9	**70.4**	72.5	**73.2**
κ	0.645	**0.659**	0.660	**0.688**	0.672	**0.695**	0.692	**0.716**

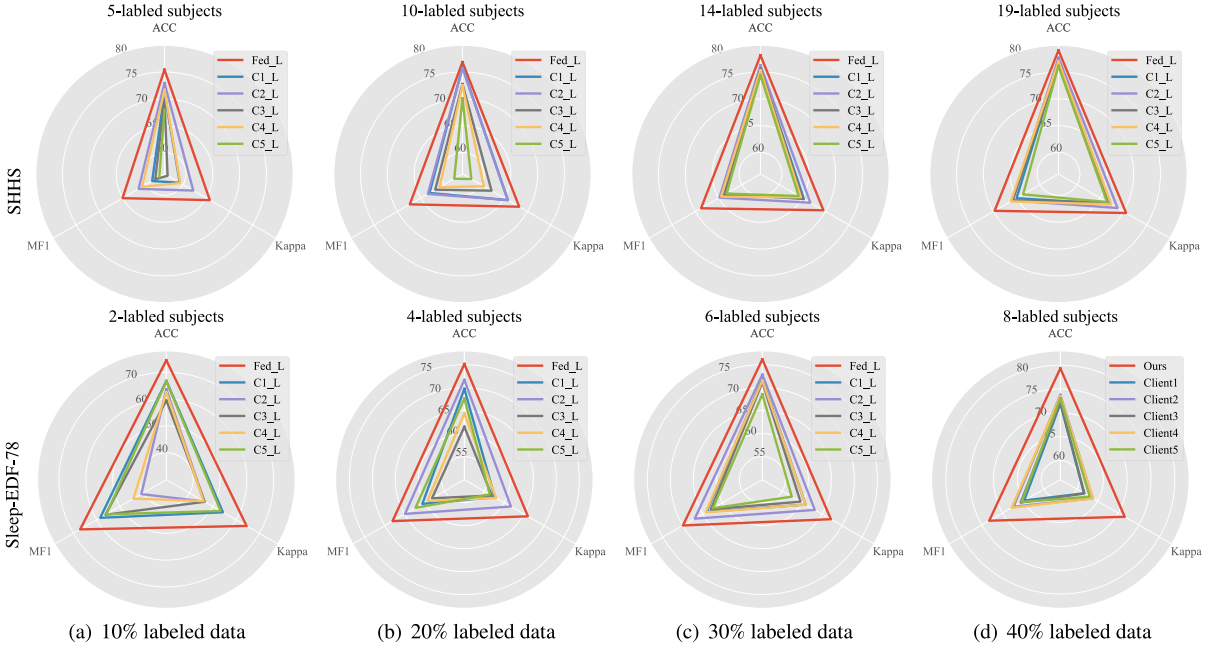


Fig. 3. Performance comparison of federated learning and independent learning, using the same labeled data of each client.

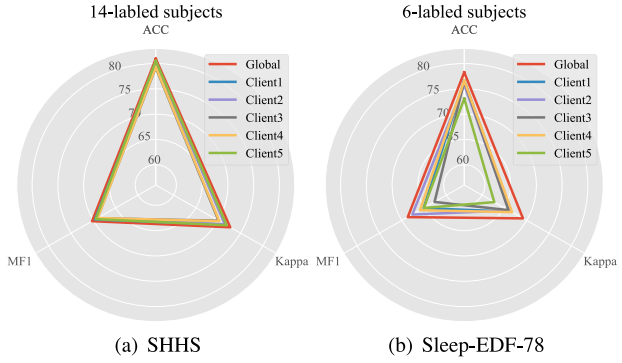
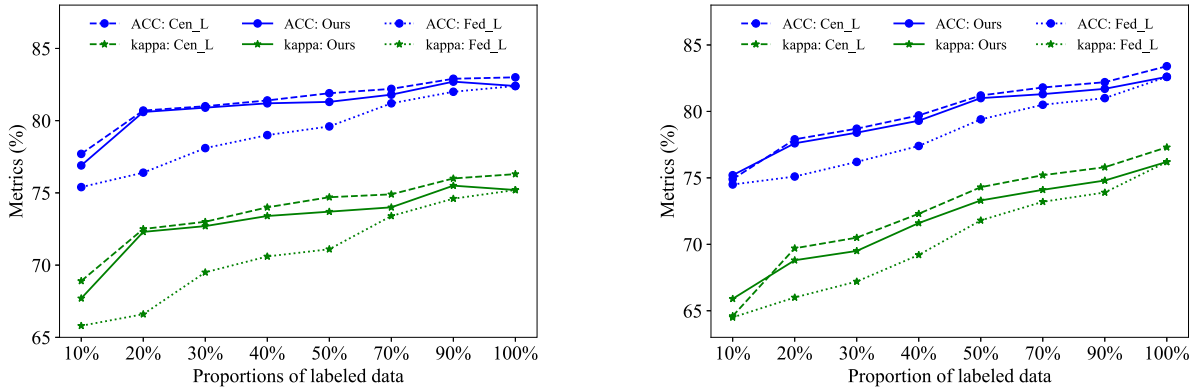


Fig. 4. Performance comparison of global model and local models obtained by federated semi-supervised sleep staging, when the proportion of labeled data on each client is 30%.

which demonstrate that our method outperforms Fed_L in all metrics for different proportions of labeled data. Notably, the accuracy of our method increases by more than 2% compared to the Fed_L method when the labeled data on each client exceeds 10%. It is worth noting that even with just 10% labeled data on each client, our method is able to effectively utilize knowledge extracted from labeled data to guide the training of unlabeled data and alleviate catastrophic forgetting. In summary, the comparison results confirm that our method can effectively utilize unlabeled data and extract relevant task knowledge from it, thus providing strong support for improving the performance of the model.

Additionally, we evaluated the performance of our proposed method when the labeled data comprised more than 50% on each client. We also conducted traditional centralized training based on labeled data from all clients and abbreviated the model as 'Cen_L', whose results were used as the upper bound. To present the results in a concise manner, only the comparison results of two representative metrics (ACC



(a) SHHS (b) Sleep-EDF-78

Fig. 5. Performance analysis of the proposed method based on different proportions of labeled data on each client.

and κ) are shown in Fig. 5. The comparison results confirm that our method consistently outperforms the Fed_L method and illustrate the effectiveness and stability of our proposed method. It can be seen that the performance fluctuation of our method and Cen_L method is consistent, and both metrics improves as the amount of labeled data on each client increases. This is reasonable and proves that our method is effective in extracting more task knowledge as more labeled data becomes available.

Notably, we can also see from Fig. 5 that the advantages of our method are more pronounced when the percentage of labeled data on each client is less than 50%. Specifically, the performance gap between our method and the upper bound is small, or even disappears, when the proportion of labeled data is less than half on each client. Remarkably, for the Sleep-EDF-78 dataset, the performance of our method even surpasses the upper bound when there are only 10% labeled data on each client. Another noteworthy point is that in the SHHS dataset, when the proportion of labeled data is up to 90%, the performance of the method is slightly higher than the results that the data on clients are all labeled data. It illustrate that for the SHHS dataset, when the proportion of labeled data is up to 90%, our method can extract more comprehensive knowledge from both labeled data and unlabeled data than from all of the data directly.

To summarize, our method effectively utilizes the unlabeled data on each client, and achieves a high level of performance in the task of sleep staging with insufficient labeled data.

4.3.3. Effectiveness of prototype-contrastive learning

In our work, we propose prototype-contrastive learning to obtain more distinct relationships among sleep stages. In this section, we examined the effectiveness of the proposed prototype-contrastive learning approach in improving the clarity of the relationships among sleep stages. Taking the case where only 20% of the subjects in each client were labeled, we extracted relation matrices from the test data using proposed method with the prototype-contrastive learning module and the method without the module, respectively. The relation matrices were obtained according to Eq. (1) with $\tau_1 = 2.0$. The results are shown in Fig. 6.

Upon observing and comparing the results in Figs. 6(a) and 6(b), we can see that our proposed method using prototype-contrastive learning produced relation matrices that are more inclined towards the identity matrix, indicating less confusion and more distinct relationships among sleep stages. Specifically, for the SHHS dataset, we can see the most significant improvement in distinguishability of almost 10% for W stage, while the smallest improvement is less than 2% for N1 stage. For the Sleep-EDF-78 dataset, we can observe the most significant improvement in distinguishability for stage N2, with an increase of nearly 50%. The discriminability of N3 and REM is also greatly improved, by nearly 20% and 15% respectively. Therefore, the proposed prototype-contrastive learning is helpful in obtaining more distinguishable relationships among sleep stages.

4.3.4. Superiority of our pseudo-labeling optimization

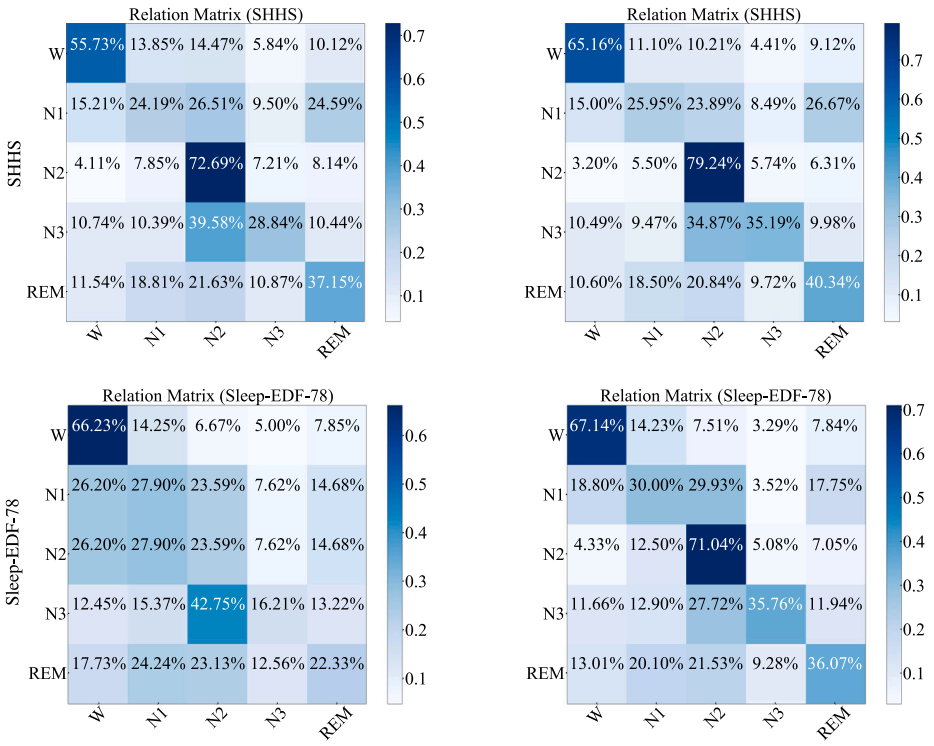
In our work, considering the global comprehensiveness and the local pertinence, we combined the global network and the local network to generate reliable pseudo-labels for unlabeled data on each client. As we have proven in Section 4.3.2 that our method can utilize unlabeled data effectively, indicating that our pseudo-labeling optimization works well to generate reliable pseudo-labels for them. In this section, we conducted a series of experiments with labeled data on each client occupying 20% to validate the superiority of the proposed pseudo-labeling optimization.

We trained semi-supervised sleep staging models independently for each client using the traditional pseudo-labeling method, which used the prediction results of a trained supervised model as the pseudo-labels, and denoted the resulting models as 'Ck_tra' with $k = 1, \ldots, 5$. Then, we compared them with the local models (represented by 'Ck_our' , $k = 1, \ldots, 5$) and federated global models (marked as 'Global_our') obtained using our method. The comparison results are shown in Fig. 7, which demonstrate that for each client, the local models obtained using our method are superior to the semi-supervised models trained using the traditional method. The performance of Global_our is undoubtedly better than Ck_tra, $k = 1, \ldots, 5$, which is in line with the findings presented in Section 4.3.1.

Furthermore, we validated the superiority of the proposed pseudo-labeling optimization by comparing it with the methods that generate pseudo-labels for unlabeled data on clients only based on the global model (denoted as 'P_Global') or the local supervised model (represented as 'P_Local'). The performance of these pseudo-labeling methods is shown in Fig. 8. It can be seen that the P_Global method slightly outperforms the P_Local method, while our proposed optimization method surpasses both methods significantly in terms of ACC, MF1 and κ. The reason for the superiority of the P_Global method over the P_local method is that the local model tends to fall into overfitting, while the global model can effectively avoid this dilemma by drawing on external knowledge. However, relying solely on the global model may introduce incorrect external knowledge and exacerbate the lack of relevant task knowledge within clients. By taking both the global model and the local supervised model into account, our optimization effectively balances the local task knowledge and external knowledge mutually, resulting in improved performance.

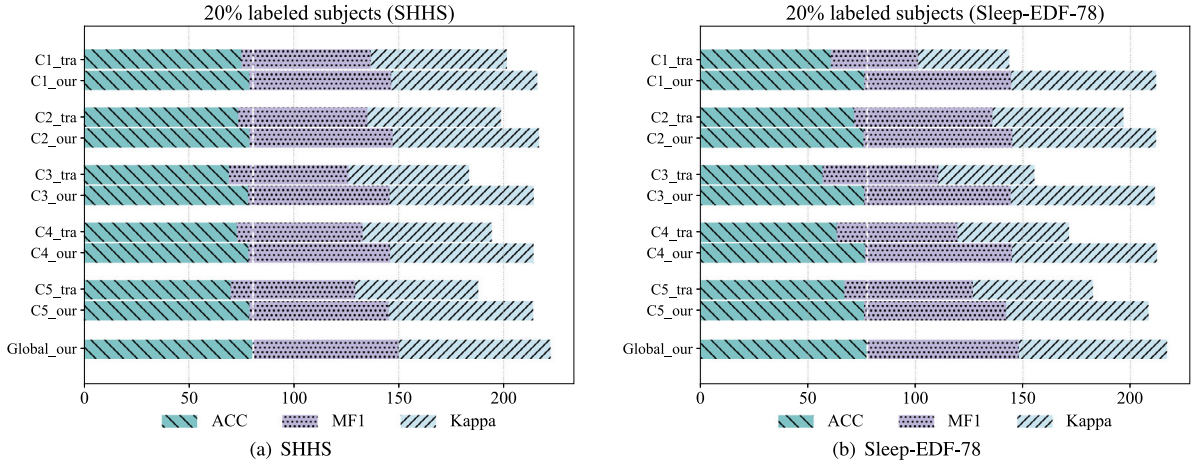
4.3.5. Comparison with state-of-the-art approaches

In this section, we compared our method with the state-of-the-art methods for semi-supervised sleep staging in terms of per-class F1 score, as well as overall metrics such as ACC, MF1, and κ. The following methods were selected for comparison: (1) CNN (Sors et al., 2018); (2) DeepSleepNet (Supratak et al., 2017); (3) CCRRSleepNet (Neng et al., 2021); (4) SHNN (Zhang et al., 2023). It is worth noting that these



(a) Method without prototype-contrastive learning (b) Method with prototype-contrastive learning (Ours)

Fig. 6. Relation matrices extracted by the method without prototype-contrastive learning and our method respectively, with 20% labeled data on each client.



(a) SHHS (b) Sleep-EDF-78

Fig. 7. The comparison of our method with traditional pseudo-labeling methods when the proportion of labeled data on each client is 20%.

methods were originally proposed as supervised automatic sleep staging models, and Zhang et al. introduced their pseudo-labeling algorithm in the training process of above models to enable semi-supervised sleep staging (Zhang et al., 2023). The descriptions for these models are outlined below:

(1) CNN: The CNN for sleep staging based on raw single-channel EEG consists of 12-layer CNN and two fully-connected layers. To enhance the modeling of scoring rules, the model takes a 30-s epoch, along with two preceding epochs and one following epoch, as inputs. Different convolutional layers have distinct kernel sizes, specifically, the first seven layers have a kernel size of 7, layers 7 to 9 have a size of 5, and the remaining layers use a size of 3. The stride of all convolutional layers is set to 2. The convolutional layers are followed by two fully-connected layers with sizes of 256 and 5. Employed as the activation function is a leaky rectified linear unit with a negative slope of 0.1.

(2) DeepSleepNet: Designed for automatic sleep staging, DeepSleepNet is comprised of a representation learning module and a sequential residual learning module. The representation learning module extracts time-invariant features from 30-s EEG epochs using two branches of CNNs. Each branch comprises four convolutional layers and two max-pooling layers. The two CNNs branches utilize small and large convolutional kernels in their first layers, respectively. Specifically, the small kernel size is set to Fs/2 (half of the sampling rate, Fs) with a stride of Fs/16, while the large kernel size is set to Fs ×4 with a stride of Fs/2. The sequential residual learning module learns transition rules among sleep stages using two layers of bidirectional-LSTMs and a shortcut connection.

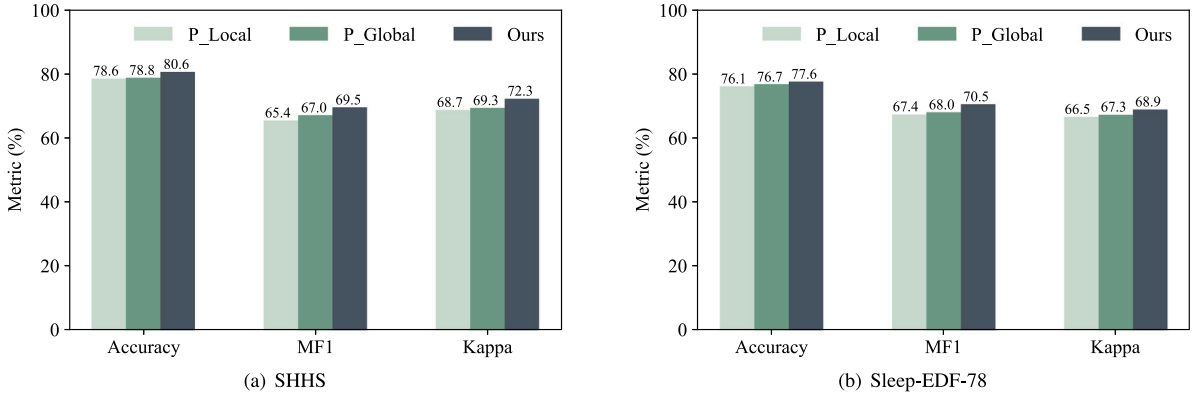


Fig. 8. The comparison of our method with pseudo-labeling methods based on only the global model or only the local supervised model, with 20% subjects are labeled on each client.

Table 4
Quantitative comparisons with state-of-the-art approaches on two public datasets.

Datasets	Channel	Methods	Per-class F1 score (F1)					Overall metrics		
			W	N1	N2	N3	REM	ACC	MF1	κ
SHHS	C4-A1	CNN*	83.91	15.08	77.75	76.14	70.35	74.99	64.65	0.6475
	C4-A1	DeepSleeNet*	81.51	17.09	72.00	73.10	67.03	72.26	62.15	0.6219
	C4-A1	CCRRSleepNet*	75.80	16.00	78.17	79.39	68.78	73.77	63.63	0.6326
	C4-A1	SHNN*	83.20	18.26	71.38	71.02	71.51	76.58	63.07	0.6655
	C4-A1	Ours	**85.36**	**20.15**	**84.40**	79.39	**78.18**	**80.64**	**69.50**	**0.7227**
Sleep-EDF-78	Fpz-Cz	CNN*	86.88	35.38	78.71	64.42	69.56	74.89	66.95	0.6475
	Fpz-Cz	DeepSleeNet*	90.6	**38.89**	80.1	65.31	**73.71**	77.31	69.72	0.6886
	Fpz-Cz	CCRRSleepNet*	87.63	29.52	80.62	75.51	60.32	74.33	66.72	0.6508
	Fpz-Cz	SHNN*	90.6	35.21	82.05	76.41	71.08	78.62	71.07	0.7051
	Fpz-Cz	Ours	**91.7**	31.7	**86.08**	**79.79**	70.19	**81.08**	**71.90**	**0.7332**

(3) CCRRSleepNet: CCRRSleepNet is an end-to-end network designed for automatic sleep staging, consisting of three blocks: frame-level CNN, epoch-level hybrid neural network and sequence-level RNN. The frame-level CNN block employs two convolutional layers with kernel sizes of 25 and 100, capturing signal characteristics at different frequencies simultaneously. A multi-scale atrous convolution block (MSACB) follows, utilizing convolution layers with varying receptive field sizes and depths to learn features of different attributes. In the epoch-level block, the CNN branch with two convolution layers (kernel sizes of 1 and 3) extracts time-invariant features. Simultaneously, the RNN branch employs Bi-GRU with 512 hidden units to learn time-varying features from the extracted frame-level features. The sequence-level block uses a two-layer BiLSTM with 512 hidden units to capture relationships among sleep stages.

(4) SHNN: The feature extraction module of SHNN incorporates a dual-scale CNN, with each CNN branch housing three convolutional layers and two pooling layers. In the CNN branch focused on fine-time granularity, the kernel sizes of three convolutional layers are 8, 6, and 2, accompanied by strides of 4, 1, and 1, respectively. For the CNN branch emphasizing coarse-time granularity, the kernel sizes of each convolutional layer are 64, 6, and 4, and the strides are set to 12, 1, and 1. Additionally, a Bi-GRU is employed to capture the temporal correlations within sleep sequences.

To ensure a fair comparison, we followed the same data setting as described in Section 4.2 for all methods on the SHHS dataset. For the Sleep-EDF-78 dataset, we followed the same experimental settings as described in Zhang et al. (2023), using 80% of the data as training set and regarding the remaining 20% for testing. Further, within the training set, 26.7% of the data was distributed to each client as labeled data, while the rest was distributed as unlabeled data. We directly compared our results with the results reported in Zhang et al. (2023).

The comparison results are presented in Table 4, and the best performance metrics for all methods are highlighted in bold. As we can see, although the F1 scores obtained by our method for N1 and REM stages are slightly lower on the Sleep-EDF-78 dataset, our method significantly outperforms other methods in F1 scores for N2 and N3 stages. With regard to overall metrics, our method outperforms all other methods on both two public datasets. Particularly, our method achieves the best MF1 and κ on both two datasets, indicating that our method is more competent in handling imbalanced data, and there is relatively high consistency between the prediction results of our method and the classification results by sleep experts (κ is between 0.61 and 0.8).

4.3.6. Ablation experiments

In this section, we conducted the ablation experiments to analyze the contribution of each component in our method. The experiment began with the 'Base' method, which refers to the federated learning based only on the labeled data on each client, i.e. 'Fed_L' method mentioned before. Then, the proposed components, such as Prototype-Contrastive learning (P_C), Pseudo-labeling (P_L), and Relation Matrix (R) were incorporated into the Base method individually or in combinations.

The results are summarized in Table 5 and the importance of each proposed component is demonstrated. It shows that B+P_C method and B+P_L method outperform the Base method, indicating that the proposed prototype-contrastive learning and the pseudo-labeling optimization are effective. The superiority of B+P_L+R method over B+P_L method is obvious, highlighting the contribution of the knowledge contained in the stage relationships to the improved performance of our method. The best performance of B+ALL method (our method) on both the SHHS dataset and the Sleep-EDF-78 dataset suggests that all the proposed components are essential for achieving optimal results.

5. Discussion

In this section, we present our analysis and discussion about the experimental results obtained based on two public datasets: the SHHS

Table 5
Results of ablation experiment on two datasets. P_L, P_C and R are abbreviations of Pseudo-labeling module, Prototype-contrastive module and Relation matrix module, respectively.

Dataset1: SHHS						
Method	Modules			Metrics		
	P_L	P_C	R	ACC(%)	MF1(%)	κ
Base	✗	✗	✗	76.3	65.5	0.662
B+P_C	✗	✓	✗	77.6	67.8	0.683
B+P_L	✓	✗	✗	77.1	67.6	0.679
B+P_L+R	✓	✗	✓	79.0	68.6	0.695
+ All (Ours)	✓	✓	✓	**80.6**	**69.5**	**0.723**
Dataset2: Sleep-EDF-78						
Method	Modules			Metrics		
	P_L	P_C	R	ACC(%)	MF1(%)	κ
Base	✗	✗	✗	75.1	68.0	0.660
B+P_C	✗	✓	✗	76.1	68.7	0.670
B+P_L	✓	✗	✗	75.9	68.2	0.668
B+P_L+R	✓	✗	✓	76.7	68.8	0.676
B+All (Ours)	✓	✓	✓	**77.6**	**70.5**	**0.689**

dataset and the Sleep-EDF-78 dataset. The summary of our findings is as follows:

(1) According to the results in Figs. 3–4 and analysis presented in Section 4.3.1, it is evident that individual institutions benefit significantly from participating in the federated collaboration. The proposed collaboration strategy effectively solves the problem of insufficient labeled data within single institution, also facilitates knowledge sharing among different institution and enabling implicit expansion of labeled data for each participating institution.

(2) The effective utilization of unlabeled data is demonstrated in Table 3 and Fig. 5, confirming the capability of our method to extract valuable task knowledge from unlabeled data. Moreover, even only a small amount of labeled data is available within each client, our approach can rival or even outperforms the traditional centralized training using all labeled data from clients.

(3) The relationships obtained through our method are more distinct, as illustrated in Fig. 6, demonstrating the effectiveness of the proposed prototype-contrastive learning.

(4) The superiority of the proposed pseudo-labeling optimization is well confirmed by the results shown in Figs. 7–8. By generating reliable pseudo-labels for unlabeled data, our method can extract the valuable task knowledge contained therein.

(5) The comparison results presented in Table 4 demonstrate the superiority of our method over current semi-supervised learning methods. Especially in the case where the labeled data is extremely limited, the superiority is highly prominent.

In summary, our method effectively addresses the issue of insufficient labeled data within a single institution, and all of the proposed components are essential for its success.

6. Conclusion

In this paper, we present a novel strategy of fostering secure collaboration among multiple institutions to tackle the problem of insufficient labeled data within a single institution. We adopt federated learning to build the collaboration, which enables implicit augmentation of labeled data and expansion of task knowledge for each participating institution by sharing knowledge with others. The task knowledge is mainly contained in the relationships among sleep stages. Specifically, we propose a federated semi-supervised sleep staging method, which simultaneously utilizes both a few labeled and a large amount of unlabeled single-channel EEG data from each institution, to complete sleep staging accurately and automatically. The prototype-contrastive learning is proposed to obtain more distinct relationships among sleep stages from labeled data. The pseudo-labeling optimization is also put forward to generate reliable pseudo-labels for unlabeled data, enabling the extraction of corresponding relationships. A series experimental results based on two public datasets demonstrate the effectiveness and superiority of our method. In practice, institutions may adopt various automatic sleep staging models according to their own circumstances. Applying traditional federated learning methods directly in such cases presents a significant challenge. How to complete the federated cooperation among these institutions based on multiple different models is the focus of our next research.

CRediT authorship contribution statement

Bian Ma: Conceptualization, Methodology, Writing – original draft. **Lijuan Duan:** Supervision, Writing – review & editing. **Zhaoyang Huang:** Writing – review & editing. **Yuanhua Qiao:** Writing – review & editing. **Bei Gong:** Writing – review & editing.

Declaration of competing interest

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

Acknowledgments

This work was supported in part by the Scientific and Technological Innovation 2030 under Grant 2021ZD0204300, in part by the National Natural Science Foundation of China under Grant 62176009. The authors also acknowledge the joint support provided by the Beijing Municipal Education Commission and the Municipal Natural Science Foundation of China under Grant 23JA002.

References

Abdulla, S., Diykh, M., Laft, R. L., Saleh, K., & Deo, R. C. (2019). Sleep EEG signal analysis based on correlation graph similarity coupled with an ensemble extreme machine learning algorithm. *Expert Systems with Applications*, *138*, Article 112790. http://dx.doi.org/10.1016/j.eswa.2019.07.007.

Assefa, S. Z., Diaz-Abad, M., Wickwire, E. M., & Scharf, S. M. (2015). The functions of sleep. *AIMS Neuroscience*, *2*(3), 155–171. http://dx.doi.org/10.3934/neuroscience.2015.3.155.

Balaji, A., Tripathi, U., Chamola, V., Benslimane, A., & Guizani, M. (2023). Toward safer vehicular transit: Implementing deep learning on single channel EEG systems for microsleep detection. *IEEE Transactions on Intelligent Transportation Systems*, *24*(1), 1052–1061. http://dx.doi.org/10.1109/tits.2021.3125126.

Banluesombatkul, N., Ouppaphan, P., Leelaarporn, P., Lakhan, P., Chaitusaney, B., Jaimchariyatam, N., Chuangsuwanich, E., Chen, W., Phan, H., Dilokthanakul, N., & Wilaiprasitporn, T. (2021). MetaSleepLearner: A pilot study on fast adaptation of bio-signals-based sleep stage classifier to new individual subject using meta-learning. *IEEE Journal of Biomedical and Health Informatics*, *25*(6), 1949–1963. http://dx.doi.org/10.1109/jbhi.2020.3037693.

Becker, N. B., de Jesus, S. N., Viseu, J. N., Stobäus, C. D., Guerreiro, M., & Domingues, R. B. (2018). Depression and quality of life in older adults: Mediation effect of sleep quality. *International Journal of Clinical and Health Psychology*, *18*(1), 8–17. http://dx.doi.org/10.1016/j.ijchp.2017.10.002.

Benjafield, A. V., Ayas, N. T., Eastwood, P. R., Heinzer, R., Ip, M. S. M., Morrell, M. J., Nunez, C. M., Patel, S. R., Penzel, T., Pépin, J.-L., Peppard, P. E., Sinha, S., Tufik, S., Valentine, K., & Malhotra, A. (2019). Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *The Lancet Respiratory Medicine*, *7*(8), 687–698. http://dx.doi.org/10.1016/s2213-2600(19)30198-5.

Chapelle, O., Schölkopf, B., & Zien, A. (2006). Introduction to semi-supervised learning. In *Semi-supervised learning* (pp. 1–12).

Chellappa, S. L., & Aeschbach, D. (2022). Sleep and anxiety: From mechanisms to interventions. *Sleep Medicine Reviews*, *61*, Article 101583. http://dx.doi.org/10.1016/j.smrv.2021.101583.

Chen, Y., Zhu, X., & Chen, W. (2015). Automatic sleep staging based on ECG signals using hidden Markov models. In *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, http://dx.doi.org/10.1109/embc.2015.7318416.

Deng, J., Wang, Y., Li, J., Wang, C., Shang, C., Liu, H., Rajasekaran, S., & Ding, C. (2021). TAG: Gradient attack on transformer-based language models. In *Findings of the association for computational linguistics: EMNLP 2021*. Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.findings-emnlp.305.

Dou, Q., Liu, Q., Heng, P. A., & Glocker, B. (2020). Unpaired multi-modal segmentation via knowledge distillation. *IEEE Transactions on Medical Imaging*, *39*(7), 2415–2425. http://dx.doi.org/10.1109/tmi.2019.2963882.

Eldele, E., Chen, Z., Liu, C., Wu, M., Kwoh, C.-K., Li, X., & Guan, C. (2021). An attention-based deep learning approach for sleep stage classification with single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *29*, 809–818. http://dx.doi.org/10.1109/tnsre.2021.3076234.

Fiorillo, L., Puiatti, A., Papandrea, M., Ratti, P.-L., Favaro, P., Roth, C., Bargiotas, P., Bassetti, C. L., & Faraci, F. D. (2019). Automated sleep scoring: A review of the latest approaches. *Sleep Medicine Reviews*, *48*, Article 101204. http://dx.doi.org/10.1016/j.smrv.2019.07.007.

Fonseca, P., den Teuling, N., Long, X., & Aarts, R. M. (2017). Cardiorespiratory sleep stage detection using conditional random fields. *IEEE Journal of Biomedical and Health Informatics*, *21*(4), 956–966. http://dx.doi.org/10.1109/jbhi.2016.2550104.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In M. F. Balcan, & K. Q. Weinberger (Eds.), *Proceedings of machine learning research*: *Vol. 48*, *Proceedings of the 33rd international conference on machine learning* (pp. 1050–1059). New York, New York, USA: PMLR, URL https://proceedings.mlr.press/v48/gal16.html.

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet. *Circulation*, *101*(23), http://dx.doi.org/10.1161/01.cir.101.23.e215.

Goshtasbi, N., Boostani, R., & Sanei, S. (2022). SleepFCN: A fully convolutional deep learning framework for sleep stage classification using single-channel electroencephalograms. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *30*, 2088–2096. http://dx.doi.org/10.1109/tnsre.2022.3192988.

Grimaldi, D., Carter, J. R., Cauter, E. V., & Leproult, R. (2016). Adverse impact of sleep restriction and circadian misalignment on autonomic function in healthy Young adults. *Hypertension*, *68*(1), 243–250. http://dx.doi.org/10.1161/hypertensionaha.115.06847.

Guillot, A., & Thorey, V. (2021). RobustSleepNet: Transfer learning for automated sleep staging at scale. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *29*, 1441–1451. http://dx.doi.org/10.1109/tnsre.2021.3098968.

Güneş, S., Polat, K., & Yosunkaya, Ş. (2010). Efficient sleep stage recognition system based on EEG signal using k-means clustering based feature weighting. *Expert Systems with Applications*, *37*(12), 7922–7928. http://dx.doi.org/10.1016/j.eswa.2010.04.043.

Hanaoka, M., Kobayashi, M., & Yamazaki, H. (2001). Automated sleep stage scoring by decision tree learning. *Vol. 2*, In *2001 conference proceedings of the 23rd annual international conference of the IEEE engineering in medicine and biology society* (pp. 1751–1754). http://dx.doi.org/10.1109/IEMBS.2001.1020556.

Hepsomali, P., & Groeger, J. A. (2021). Diet, sleep, and mental health: Insights from the UK biobank study. *Nutrients*, *13*(8), 2573. http://dx.doi.org/10.3390/nu13082573.

Iber, C., Ancoli-Israel, S., Chesson, A. L., & Quan, S. F. (2007). The AASM manual for the scoring of sleep and associated events: Rules, terminology and technical specifications.

Jin, Y., Wei, X., Liu, Y., & Yang, Q. (2020). Towards utilizing unlabeled data in federated learning: A survey and prospective. arXiv:2002.11545.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., ... Zhao, S. (2019). Advances and open problems in federated learning. arXiv preprint arXiv:1912.04977, arXiv:1912.04977.

Kemp, B., Zwinderman, A., Tuk, B., Kamphuisen, H., & Oberye, J. (2000). Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, *47*(9), 1185–1194. http://dx.doi.org/10.1109/10.867928.

Kim, W., Park, K., Sohn, K., Shu, R., & Kim, H.-S. (2022). Federated semi-supervised learning with prototypical networks. (27), http://dx.doi.org/10.48550/arXiv.2205.13921, arXiv:2205.13921.

Le-Dong, N.-N., Martinot, J.-B., Coumans, N., Cuthbert, V., Tamisier, R., Bailly, S., & Pépin, J.-L. (2021). Machine learning–based sleep staging in patients with sleep apnea using a single mandibular movement signal. *American Journal of Respiratory and Critical Care Medicine*, *204*(10), 1227–1231. http://dx.doi.org/10.1164/rccm.202103-0680le.

Li, Q., He, B., & Song, D. (2021). Model-contrastive federated learning. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. IEEE, http://dx.doi.org/10.1109/cvpr46437.2021.01057.

Liew, S. C., & Aung, T. (2021). Sleep deprivation and its association with diseases-a review. *Sleep Medicine*, *77*, 192–204. http://dx.doi.org/10.1016/j.sleep.2020.07.048.

Liu, Q., Chen, C., Qin, J., Dou, Q., & Heng, P.-A. (2021). FedDG: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. IEEE, http://dx.doi.org/10.1109/cvpr46437.2021.00107.

Liu, Q., Yang, H., Dou, Q., & Heng, P. A. (2021). Federated semi-supervised medical image classification via inter-client relation matching. In *Medical image computing and computer assisted intervention – MICCAI 2021*.

Lou, G., Liu, Y., Zhang, T., & Zheng, X. (2021). STFL: A temporal-spatial federated learning framework for graph neural networks. arXiv:2111.06750.

Mathur, S., & Dinakarpandian, D. (2012). Finding disease similarity based on implicit semantic similarity. *Journal of Biomedical Informatics*, *45*(2), 363–371. http://dx.doi.org/10.1016/j.jbi.2011.11.017.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. *vol. 54*, In *20th international conference on artificial intelligence and statistics (AISTATS) 2017* (pp. 1273–1282). arXiv:1602.05629.

Michieli, U., & Ozay, M. (2021). Prototype guided federated learning of visual feature representations. (19), http://dx.doi.org/10.48550/arXiv.2105.08982, arXiv:2105.08982.

Neng, W., Lu, J., & Xu, L. (2021). CCRRSleepNet: A hybrid relational inductive biases network for automatic sleep stage classification on raw single-channel EEG. *Brain Sciences*, *11*(4), 456. http://dx.doi.org/10.3390/brainsci11040456.

Oerton, E., Roberts, I., Lewis, P. S. H., Guilliams, T., & Bender, A. (2018). Understanding and predicting disease relationships through similarity fusion. In J. Wren (Ed.), *Bioinformatics*, *35*(7), 1213–1220. http://dx.doi.org/10.1093/bioinformatics/bty754.

Pan, Q., Brulin, D., & Campo, E. (2022). Wrist movement analysis for long-term home sleep monitoring. *Expert Systems with Applications*, *187*, Article 115952. http://dx.doi.org/10.1016/j.eswa.2021.115952.

Phan, H., Andreotti, F., Cooray, N., Chén, O. Y., & Vos, M. D. (2018). Automatic sleep stage classification using single-channel EEG: Learning sequential features with attention-based recurrent neural networks. In *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (pp. 1452–1455). http://dx.doi.org/10.1109/EMBC.2018.8512480.

Phan, H., Andreotti, F., Cooray, N., Chen, O. Y., & Vos, M. D. (2019). Joint classification and prediction CNN framework for automatic sleep stage classification. *IEEE Transactions on Biomedical Engineering*, *66*(5), 1285–1296. http://dx.doi.org/10.1109/tbme.2018.2872652.

Phan, H., Andreotti, F., Cooray, N., Chen, O. Y., & Vos, M. D. (2019). SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *27*(3), 400–410. http://dx.doi.org/10.1109/tnsre.2019.2896659.

Phan, H., Chen, O. Y., Tran, M. C., Koch, P., Mertins, A., & Vos, M. D. (2021). XSleepNet: Multi-view sequential model for automatic sleep staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1. http://dx.doi.org/10.1109/tpami.2021.3070057.

Phyo, J., Ko, W., Jeon, E., & Suk, H.-I. (2022). TransSleep: Transitioning-aware attention-based deep neural network for sleep staging. *IEEE Transactions on Cybernetics*, 1–11. http://dx.doi.org/10.1109/TCYB.2022.3198997.

Qu, W., Wang, Z., Hong, H., Chi, Z., Feng, D. D., Grunstein, R., & Gordon, C. (2020). A residual based attention model for EEG based sleep staging. *IEEE Journal of Biomedical and Health Informatics*, *24*(10), 2833–2843. http://dx.doi.org/10.1109/jbhi.2020.2978004.

Quan, S., Howard, B., Iber, C., Kiley, J., Nieto, F., O'Connor, G., Rapoport, D., Redline, S., Robbins, J., Samet, J., & Wahl, P. (1997). The sleep heart health study: design, rationale, and methods. *Sleep*, *20*(12), 1077–1085, URL http://europepmc.org/abstract/MED/9493915.

Radhakrishnan, B. L., Kirubakaran, E., Jebadurai, I. J., Selvakumar, A. I., & Peter, J. D. (2022). Efficacy of single-channel EEG: A propitious approach for in-home sleep monitoring. *Frontiers in Public Health*, *10*, http://dx.doi.org/10.3389/fpubh.2022.839838.

Seifpour, S., Niknazar, H., Mikaeili, M., & Nasrabadi, A. M. (2018). A new automatic sleep staging system based on statistical behavior of local extrema using single channel EEG signal. *Expert Systems with Applications*, *104*, 277–293. http://dx.doi.org/10.1016/j.eswa.2018.03.020.

Snell, J., Swersky, K., & Zemel, R. S. (2017). Prototypical networks for few-shot learning. (15), http://dx.doi.org/10.48550/arXiv.1703.05175, arXiv:1703.05175.

Sohn, K. (2016). Improved deep metric learning with multi-class N-pair loss objective. In *Proceedings of the 30th international conference on neural information processing systems* (pp. 1857–1865). Red Hook, NY, USA: Curran Associates Inc..

Sors, A., Bonnet, S., Mirek, S., Vercueil, L., & Payen, J.-F. (2018). A convolutional neural network for sleep stage scoring from raw single-channel EEG. *Biomedical Signal Processing and Control*, *42*, 107–114. http://dx.doi.org/10.1016/j.bspc.2017.12.001.

Stephansen, J. B., Olesen, A. N., Olsen, M., Ambati, A., Leary, E. B., Moore, H. E., Carrillo, O., Lin, L., Han, F., Yan, H., Sun, Y. L., Dauvilliers, Y., Scholz, S., Barateau, L., Hogl, B., Stefani, A., Hong, S. C., Kim, T. W., Pizza, F., ... Mignot, E. (2018). Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nature Communications*, *9*(1), 5229. http://dx.doi.org/10.1038/s41467-018-07229-3.

Sun, C., Chen, C., Li, W., Fan, J., & Chen, W. (2020). A hierarchical neural network for sleep stage classification based on comprehensive feature learning and multi-flow sequence learning. *IEEE Journal of Biomedical and Health Informatics*, *24*(5), 1351–1366. http://dx.doi.org/10.1109/jbhi.2019.2937558.

Supratak, A., Dong, H., Wu, C., & Guo, Y. (2017). DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *25*(11), 1998–2008. http://dx.doi.org/10.1109/tnsre.2017.2721116.

Tan, Y., Long, G., LIU, L., Zhou, T., Lu, Q., Jiang, J., & Zhang, C. (2022). FedProto: Federated prototype learning across heterogeneous clients. *Vol. 36*, In *Proceedings of the AAAI conference on artificial intelligence* (8), (pp. 8432–8440). Association for the Advancement of Artificial Intelligence (AAAI), http://dx.doi.org/10.1609/aaai.v36i8.20819.

Thorey, V., Hernandez, A. B., Arnal, P. J., & During, E. H. (2019). AI vs humans for the diagnosis of sleep apnea. In *2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, http://dx.doi.org/10.1109/embc.2019.8856877.

Vaquerizo-Villar, F., Alvarez, D., Kraemer, J. F., Wessel, N., Gutierrez-Tobal, G. C., Calvo, E., del Campo, F., Kheirandish-Gozal, L., Gozal, D., Penzel, T., & Hornero, R. (2021). Automatic sleep staging in children with sleep apnea using photoplethysmography and convolutional neural networks. In *2021 43rd annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, http://dx.doi.org/10.1109/embc46164.2021.9629995.

Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning. *ACM Transactions on Intelligent Systems and Technology*, *10*(2), 1–19. http://dx.doi.org/10.1145/3298981.

Yang, X., Song, Z., King, I., & Xu, Z. (2021). A survey on deep semi-supervised learning. arXiv:2103.00550.

Yi, J., Wu, F., Wu, C., Liu, R., Sun, G., & Xie, X. (2021). Efficient-FedRec: Efficient federated learning framework for privacy-preserving news recommendation. In *Proceedings of the 2021 conference on empirical methods in natural language processing*. Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.emnlp-main.223.

Yoo, C., Lee, H. W., & Kang, J.-W. (2022). Transferring structured knowledge in unsupervised domain adaptation of a sleep staging network. *IEEE Journal of Biomedical and Health Informatics*, *26*(3), 1273–1284. http://dx.doi.org/10.1109/jbhi.2021.3103614.

Zhang, Y., Cao, W., Feng, L., Wang, M., Geng, T., Zhou, J., & Gao, D. (2023). SHNN: A single-channel EEG sleep staging model based on semi-supervised learning. *Expert Systems with Applications*, *213*, Article 119288. http://dx.doi.org/10.1016/j.eswa.2022.119288.

Zhang, G.-Q., Cui, L., Mueller, R., Tao, S., Kim, M., Rueschman, M., Mariani, S., Mobley, D., & Redline, S. (2018). The national sleep research resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*, *25*(10), 1351–1358. http://dx.doi.org/10.1093/jamia/ocy064.

Zhang, L., Fabbri, D., Upender, R., & Kent, D. (2019). Automated sleep stage scoring of the sleep heart health study using deep neural networks. *Sleep*, *42*(11), http://dx.doi.org/10.1093/sleep/zsz159.

Zhang, C., Yu, W., Li, Y., Sun, H., Zhang, Y., & Vos, M. D. (2022). CMS2-net: Semi-supervised sleep staging for diverse obstructive sleep apnea severity. *IEEE Journal of Biomedical and Health Informatics*, *26*(7), 3447–3457. http://dx.doi.org/10.1109/jbhi.2022.3156585.

Zhu, X., & Goldberg, A. B. (2009). Overview of semi-supervised learning. In *Introduction to semi-supervised learning* (pp. 9–19). Springer International Publishing, http://dx.doi.org/10.1007/978-3-031-01548-9_2.