ConCAD: Contrastive Learning-Based Cross Attention for Sleep Apnea Detection

Guanjie Huang(✉) and Fenglong Ma(✉)

College of Information Sciences and Technology, Pennsylvania State University, State College, PA 16802, USA
{gzh8,fenglong}@psu.edu

Abstract. With recent advancements in deep learning methods, automatically learning deep features from the original data is becoming an effective and widespread approach. However, the hand-crafted expert knowledge-based features are still insightful. These expert-curated features can increase the model's generalization and remind the model of some data characteristics, such as the time interval between two patterns. It is particularly advantageous in tasks with the clinically-relevant data, where the data are usually limited and complex. To keep both implicit deep features and expert-curated explicit features together, an effective fusion strategy is becoming indispensable. In this work, we focus on a specific clinical application, i.e., sleep apnea detection. In this context, we propose a contrastive learning-based cross attention framework for sleep apnea detection (named `ConCAD`). The cross attention mechanism can fuse the deep and expert features by automatically assigning attention weights based on their importance. Contrastive learning can learn better representations by keeping the instances of each class closer and pushing away instances from different classes in the embedding space concurrently. Furthermore, a new hybrid loss is designed to simultaneously conduct contrastive learning and classification by integrating a supervised contrastive loss with a cross-entropy loss. Our proposed framework can be easily integrated into standard deep learning models to utilize expert knowledge and contrastive learning to boost performance. As demonstrated on two public ECG dataset with sleep apnea annotation, `ConCAD` significantly improves the detection performance and outperforms state-of-art benchmark methods.

Keywords: Contrastive learning · Cross attention · Sleep apnea detection

1 Introduction

According to the National Institutes of Health of USA, 50 to 70 million people have chronic sleep disorders [4], and the sleep disorders of sleep can increase the

Y. Dong et al. (Eds.): ECML PKDD 2021, LNAI 12979, pp. 68–84, 2021.
https://doi.org/10.1007/978-3-030-86517-7_5

risk of many related diseases, such as hypertension and cardiovascular pathologies [41]. Sleep apnea is one of the most common sleep disorders, which is an abnormal respiratory activity repeatedly occurring during sleep. The current primary method for diagnosing sleep apnea requires the patient to record the polysomnogram (PSG) in a clinic setup, which is very inconvenient and belated. Thus, how to automatically and effectively detect sleep apnea is a challenge, especially in the earlier stages.

Towards this end, automatic sleep apnea detection methods [3,10,18,30,35] have been developed to simplify the diagnostic procedure, which including traditional machine learning methods and deep learning methods. Existing studies [29,35,38] have shown that deep learning models perform better than the traditional machine learning ones, which require expert knowledge to manually extract features. However, these hand-crafted expert features are still valuable and insightful. While researching the most appropriate hand-crafted features is time-consuming, there are a number of hand-crafted features that can be leveraged right away, as summarized by previous studies over centuries. In this context, we proposed a **cross-attention mechanism** to combine the deep features and the hand-crafted features to take advantage of both of them appropriately.

On the other hand, the regular deep learning methods usually train with the cross-entropy (CE) loss for a classification task. Since the CE loss only focuses on learning the necessary features to solve the classification task over known training data, it can be easily impaired by the mislabeled data [42], which further hinders the quality of the learned representations [25]. To alleviate the problem, a common solution is to collect more data so that the model can learn more general features without excessive discrimination. However, this solution is particularly impractical in clinically relevant data, such as electrocardiography (ECG), where the data are always limited, and the labeling is prone to human errors. As a remedy, we design a novel hybrid loss that integrates the cross-entropy loss with a **contrastive loss**. The contrastive loss helps to learn more general and robust features by clustering similar data and pushing apart dissimilar ones.

To sum up, in this work, we proposed a novel CONtrastive learning-based Cross Attention for sleep apnea Detection (`ConCAD`) using single ECG data. To the best of our knowledge, our work is the first to successfully integrate contrastive learning for sleep apnea detection. Our major contributions of this paper are as follows:

- We propose a cross attention mechanism to combine the deep features and expert knowledge-based features, which automatically fuses the features by emphasizing the important parts based on each other synergistically.
- We design a novel hybrid loss that encompasses both the cross-entropy (CE) loss and the supervised contrastive (SC) loss. The SC loss help to learn more general and robust by minimizing the ratio of intra-class to inter-class similarity while cross-entropy CE loss focus on discovering the useful features to solve the classification task.
- We demonstrate state-of-the-art classification performance on two public ECG datasets outperforming all benchmark methods.

– We show that our proposed framework of contrastive learning-based cross attention has better generalization ability, especially when the number of labeled training data is limited, comparing to a naive deep learning method without it.
– Both the cross attention mechanism and contrastive learning can be painlessly integrated into standard deep learning models.

2 Related Work

In this section, we review the studies related to the proposed `ConCAD` model, including the work on sleep apnea detection, cross attention mechanism for feature fusion, and contrastive learning.

2.1 Sleep Apnea Detection

The standard approach to diagnose sleep apnea requires the patient to sleep overnight at a clinic setup and record the polysomnography (PSG) by various physiological sensors, and then the outputs of PSG are visually inspected by a clinical expert to give a diagnosis [6,19,23]. This process is always inconvenient and uncomfortable. Thus, some studies have begun to simplify the procedure of diagnosing sleep apnea by only using a single physiological data, such as ECG [3,10], EEG [2], and the respiration signal [31]. Among these physiological data, ECG is a less intrusive option and also strongly related to sleep apnea.

To this end, several studies [18,32] manually extract hand-crafted features and feed them to classifiers (e.g., random forest, support vector machine) for sleep apnea detection. Recently, with the development of deep learning methods, some studies [3,35] extract RR interval (RRI) and the R-peak envelope (RPE) and build deep learning model to automatically learn representation and detect sleep apnea. Furthermore, several studies [10,30] develop deep learning models to directly learn features from the raw ECG data and detect sleep apnea in an end-to-end style.

2.2 Attention-Based Feature Fusion

Another line of related work is feature fusion, which aims at combining different features to obtain a more effective representation. The frequently-used feature fusion techniques are concatenation [29], summation [11], and multiplication [39]. However, these operations evenly combine all the features together without considering the importance of each feature. Some of the features gathered will help the model make the right decision, while others can lead to significant misjudgment [21].

Recently, the usage of attention learning mechanism has shown remarkable performance improvement for different tasks, such as natural language processing [37], image classification [33], and object tracking [8]. The attention mechanism highlights the effective discriminant parts of features while suppressing the

redundant parts to a certain degree. To further take advantage of the features extracted from multi-modality inputs, a cross attention mechanism has been proposed to derive an attention mask from different inputs mutually. In [22], the authors use one modality (LiDAR) to generate an attention mask that controls the spatial features of a different modality (HSI). In [16], the authors derive cross attention maps for each pair of class features and query sample features to highlight specific regions and make the extracted features more discriminative.

2.3 Contrastive Learning

All of the deep learning methods mentioned above are trained by the cross-entropy loss. The cross-entropy loss is the most commonly-used one in the classification tasks, which calculates the difference between the actual probability distribution of the data and the predicted probability distribution of the model [28]. As we previously introduced, the cross-entropy loss has some limitations. Thus, a supervised contrastive loss is added as an auxiliary regularization to alleviate problems in our proposed framework.

The contrastive loss has recently been widely used in self-surprised learning [5,14,24], which aims at clustering the similar data and pushing apart the dissimilar data. A supervised version of the contrastive loss is proposed by [20] to leverage the label information. Their proposed supervised contrastive learning contains two steps: First, the supervised contrastive loss is used to learn a representation to cluster the data from the same class and separate the data from different classes; Second, they froze the model and add a multi-layer perceptron (MLP) as a classifier on its top for the classification task. Recently, supervised contrasting learning has been used for different applications, such as image classification [20], few-shot classification [25], and semantic segmentation [36].

3 Methodology

The goal of this work is to design an effective framework for leveraging the power of both the deep learning-based features and the expert knowledge-based features simultaneously to enhance the performance of sleep apnea detection. Towards this goal, we propose `ConCAD` as shown in Fig. 1, which is based on the contrastive learning framework to obtain better representations and utilizes a cross-attention mechanism to fuse different types of features.

Concretely, `ConCAD` is achieved by three steps as shown in Fig. 1. Firstly, the original raw data are passed through a feature extractor to learn the deep features. Simultaneously, the expert knowledge is passed through a feature extractor with a relatively shallow network to learn the expert features. Besides, data augmentation can be used before passing the data. Secondly, the deep features and the expert features are fed to a cross-attention module, which automatically fuses the features by emphasizing the important parts based on each other synergistically. Thirdly, the resultant attention-weighted features are mapped

into a projection space for learning a representation with high intra-class similarity and low inter-class similarity to improve the classification accuracy by contrastive learning. Then, the learned representation is fed to the classification modules to output the probability of sleep apnea events.

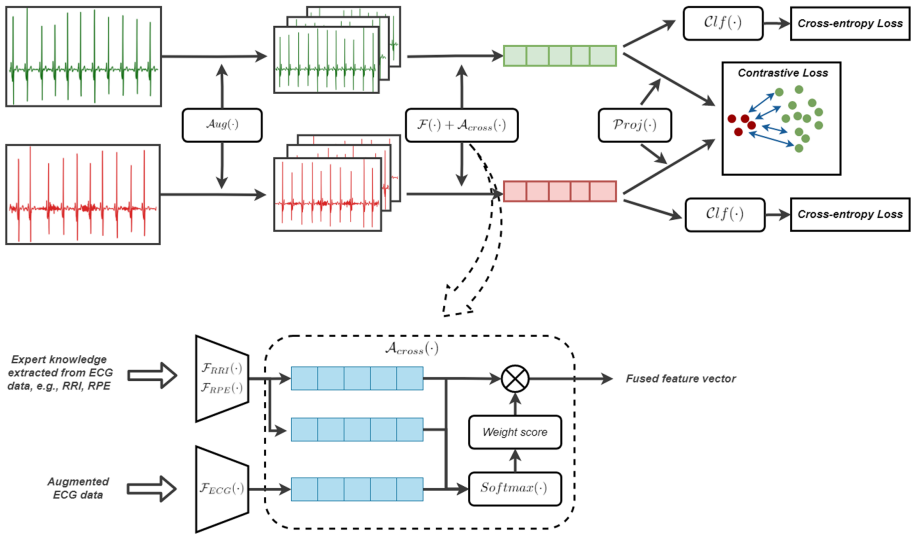


Fig. 1. Overall of our proposed `ConCAD` framework. The cross attention fuses the features by generating an attention weight mask. It can highlight the effective discriminant parts and suppress the irrelevant parts of features from ECG, RRI, and RPE collaboratively. Besides the cross-entropy, the supervised contrastive loss is also computed to optimize the intra-class to inter-class similarity ratio.

As we are going to demonstrate the proposed `ConCAD` on the datasets of ECG-based sleep apnea detection, the ECG data are certainly considered as the raw data input. Furthermore, the RRI and RPE manifest their effectiveness to detect sleep apnea by many research works [1,9]. Consequently, the RRI and RPE are chosen as the expert knowledge input for our proposed framework.

3.1 Expert Feature Extraction and Data Augmentation

The expert knowledge is summarized by previous researches over the last centuries. In the field of detecting sleep apnea using ECG data, several previous studies [1,9] have shown that the RR intervals (RRI) and R-peak envelope (RPE) are effective. To prepare the RRI and RPE data, we first detect the locations of the R-peaks by the Hamilton algorithm [13]. Then, we calculate the distance between R-peaks as the RRI and use the amplitudes of the R-peaks as the RPE. Since the RRI can be easily disturbed by unexpected ECG spikes, a median filter

is used to eliminate the disturbance as suggested by [7]. Besides, since the number of the RRI or RPE is not always the same by giving a fixed time duration (e.g., 1 min), cubic interpolation was used to resample them to the same length [35].

We are augmenting the ECG, RRI and RPE by two simple approaches: random time shift and reversion. Given the data $\mathbf{x} = [x_0, x_1, x_2, \ldots, x_n]$, the random time shift will obtain $\mathbf{x}_{shift} = [x_t, x_{1+t}, x_{2+t}, \ldots, x_{n+t}]$, where t is a randomly-generated number and represents the number of data points to shift. The revision will generate $\mathbf{x}_{reverse} = [x_n, x_{n-1}, \ldots, x_2, x_1, x_0]$. The augmentation is conducted in each batch to provide more positives (i.e., instances with the same label) during batch training, which benefits a more robust clustering of the projection space. The augmentation process is presented as $\mathcal{A}ug(\mathbf{x})$.

3.2 Feature Extractor

The feature extractor should be designed case by case. In this study, we have three feature extractors, which are used for learning features from the ECG, RRI and RPE separately, and named $\mathcal{F}_{ECG}$, $\mathcal{F}_{RRI}$ and $\mathcal{F}_{RPE}$.

$\mathcal{F}$ consists of 4 convolution blocks. The first three blocks are made of one convolutional layer, one batch normalization layer, one ReLU activation layer, one maxpooling layer, and one dropout layer. The feature map size of the convolutional layer in the first block is chosen to cover data points of two contiguous beats in case that the patterns between beats get missed. The last convolution block does not contain the maxpooling and dropout layer. The module can be represented as $\mathbf{x}' = \mathcal{F}(\mathbf{x}; \theta_F)$, where $\mathbf{x}$ represents the input data and θ_F denotes the parameters of the module.

Since there are three kinds of data, we have three corresponding extractors, which are $\mathbf{x}'_{ECG} = \mathcal{F}_{ECG}(\mathbf{x}_{ECG}; \theta_{F_{ECG}})$ for ECG data, $\mathbf{x}'_{RRI} = \mathcal{F}_{RRI}(\mathbf{x}_{RRI}; \theta_{F_{RRI}})$ and $\mathbf{x}'_{RPE} = \mathcal{F}_{ECG}(\mathbf{x}_{RPE}; \theta_{F_{RPE}})$ for the expert knowledge-based features. More details of the extractors are described in Appendix A.

3.3 Cross Attention

Not all the deep features and expert features contribute equally to the classification task. Thus, we design a cross-attention module, $\mathcal{A}_{cross}$, to collaboratively learn their importance and concentrate more on the important ones. The cross-attention is designed to ask the model to concentrate on the particular features, which contribute more to distinguish the instances from different classes. Before computing the cross attention, since the outputs of feature extractors are likely to have different dimensions, we need to project them to the same space by a linear transformation. Given $\mathbf{x}' \in \mathbb{R}^{m \times n}$, the transformation is

$$\mathbf{x}'' = \mathbf{u}^\top \mathbf{x}' \mathbf{V} \tag{1}$$

where $\mathbf{u} \in \mathbb{R}^m$ and $\mathbf{V} \in \mathbb{R}^{n \times k}$ are trainable parameters.

After it, $\mathbf{x}''_{ECG}$, $\mathbf{x}''_{RRI}$ and $\mathbf{x}''_{RPE}$ have the same dimension k. Then, we are going to compute the attention weights. Specifically,

$$\begin{aligned} \boldsymbol{\alpha} &= Softmax([\alpha_{ECG}, \alpha_{RRI}, \alpha_{RPE}]) \\ \alpha_i &= \mathbf{w}_i^\top \mathbf{x}''_i + b_i \end{aligned} \tag{2}$$

where $i \in \mathcal{S} = \{ECG, RRI, RPE\}$ and $\boldsymbol{\alpha} \in \mathbb{R}^3$ is the attention weights. $\mathbf{w} \in \mathbb{R}^k$ and $b \in \mathbb{R}$ are trainable parameters. The transformed $\mathbf{x}''$ is passed through an one-layer MLP to learn the importance of different types of features synergistically, and the importance is normalized by a softmax function.

Lastly, we compute the context vector $\mathbf{c}$ by

$$\mathbf{c} = \sum_{i \in \mathcal{S}} \alpha_i \mathbf{x}''_i \tag{3}$$

The context vector is the fused feature vector that is the weighted sum of features from different inputs based on the learned importance. The cross-attention module can be represented as $\mathcal{A}_{cross}([\mathbf{x}'_{ECG}, \mathbf{x}'_{RRI}, \mathbf{x}'_{RPE}]; \theta_{A_{cross}})$.

3.4 Contrastive Learning.

For most of the conventional classification tasks, cross-entropy (CE) loss is commonly used to adjust model weights during training. However, CE loss may be impaired by noisy labels [42] and induce representations with excessive discrimination towards training data [25]. In order words, CE loss is likely to result in sub-optimal generalization.

As a remedy, contrastive learning is adopted to assist the model to learn more general and robust features by maximizing intra-class similarity while minimizing inter-class similarity. Concretely, we propose a novel hybrid loss, which utilizes the supervised contrastive (SC) loss [20] as an auxiliary regularization to the standard CE loss.

Contrastive Loss. The SC loss aims at simultaneously increasing the agreement among instances in positive pairs and encouraging the difference among instances in negative pairs. The instances with the same label form the positive pairs, and the instances with the different labels are considered as negative pairs. Specifically, the SC loss is computed in two steps. We first project the input, i.e., the fused feature vector, to a lower dimension space by a one-layer MLP, and the low dimension vector is normalized to the unit hypersphere by L2 norm, $\mathbf{z} = \mathcal{P}roj_{SC}(\mathcal{A}_{cross}([\mathbf{x}'_{ECG}, \mathbf{x}'_{RRI}, \mathbf{x}'_{RPE}]; \theta_{A_{cross}})$. Then, the SC loss can be computed by

$$\mathcal{L}_{SC} = -\sum_{i=1}^{N} \frac{1}{N_{y_i}} \log \frac{\sum_{j=1}^{N} \mathbb{1}_{[y_i = y_j]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}, \tag{4}$$

where N is the batch size, and N_{y_i} is the number of samples with the same label in each batch. $\mathbb{1}_{[\cdot]}$ denotes an indicator function. $\text{sim}(\cdot)$ represents the measure

of similarity, and here the cosine similarity is used, i.e., $\text{sim}(u, v) = u \cdot v/\|u\|\|v\|$. τ is a hyperparameter that controls the strength of penalties on negative pairs [34].

In the SC loss formula, the numerator represents the similarity of the positives, and the denominator represents the similarity of everything else in regard to $\mathbf{z}_i$. The optimization of this formula pulls together the positives and pushes apart everything else. That is, instances from the same class will form a closer cluster while the distances between clusters are increased in the projected hypersphere. As a result, the model learns more general features instead of naively learning the features for the classification task over the known training data.

Hybrid Loss. As described in [20], the standard SC loss requires two separate steps for a classification task: first, they train the feature extractor with the SC loss to learn a representation vector; second, they freeze the feature extractor and train a classifier on the vector using the CE loss. However, the SC loss usually requires a very large batch size to achieve decent and stable performance. For example, [20] uses a batch size of 6,144. On the other hand, the CE loss only works in the second step and cannot update the model parameters in the feature extractor, which means that the CE loss does not make any contribution to learning the feature representation.

To alleviate these problems, we use the SC loss as an auxiliary regularization term and integrate it with the CE loss. Specifically, we propose a new hybrid loss, which is the summation of CE and SC losses with a scaling parameter λ to control the contribution of each loss:

$$\mathcal{L}_{hybrid} = \lambda \mathcal{L}_{CE} + (1 - \lambda)\mathcal{L}_{SC}. \tag{5}$$

With the proposed hybrid loss, the model can take advantage of both the CE and SC losses simultaneously. The CE loss can learn effective features for classification tasks with small batch sizes, and the SC loss helps to promote these features to be more general and robust by minimizing the intra-class to inter-class similarity ratio. To train the model with the proposed hybrid loss, we project the fused feature vector to lower dimension hypersphere by $\mathcal{P}roj_{SC}(\cdot)$ to calculate the SC loss. At the same time, the fused feature vector is sent to fully-connected layers $\mathcal{C}lf(\cdot)$ to calculate the CE loss as shown in the last step of Fig. 1. All the layers except the last one in $\mathcal{C}lf(\cdot)$ use the ReLU activation, while the last layer is operated on the softmax activation, and its unit number needs to be identical to the number of classes. In addition, we will discard $\mathcal{P}roj_{SC}(\cdot)$ during prediction so that the proposed model has the same number of parameters as a model with only the CE loss.

4 Experiments and Results

4.1 Datasets

In our experiments, two datasets, i.e., Apnea-ECG [26] and MIT-BIH Polysomnographic [17] obtained from Physionet [12], are used for performance

evaluation and comparison. Both datasets are publicly available and have been used to study sleep apnea detection methods in previous researches.

- **Apnea-ECG**: The apnea-ECG database is provided by Philipps University, which is the most commonly-used dataset for ECG-based sleep apnea studies. It contains 70 single-lead ECG recordings of varying lengths between 7 h to 10 h, sampled at the rate 100 Hz. Each segment of 1 min ECG data is annotated by the expert as either apnea or normal event. The datasets are officially split into two sets by the provider: a released set of 35 recordings and a withheld set of 35 recordings. After removing the data with an unreasonable heartbeat rate, the released set contains 16,888 segments and the withheld set contains 17,120 segments.
- **MIT-BIH Polysomnographic**: The MIT-BIH Polysomnographic database is collected by Boston's Beth Israel Hospital Sleep Laboratory. It contains over 80 h of polysomnographic (PSG) recordings during sleep. Each recording includes a single channel of ECG annotated beat-by-beat and EEG and respiration signals. Each segment of 30 s of data is annotated with respect to sleep stages and apnea. After removing the data with an unreasonable heartbeat rate, the final dataset contains 9,717 segments.

4.2 Compared Methods

To valid the performance of our framework, we use several state-of-the-art methods as our benchmark methods:

- Support Vector Machine (**SVM**), Random Forest (**RF**), K-Nearest Neighbor (**KNN**), and Multi-Layer Perception (**MLP**) is adopted with 10 popular hand-crafted features from ECG data (e.g., RMSSD, NN50, etc.) as benchmark methods according to the work by [18].
- **LeNet-5**: In [35], a LeNet-5 convolutional neural network is used to learn features from RRI and RPE for sleep apnea detection.
- **CNN+LSTM**: In [3], three different deep learning architectures are proposed. We adopt their best performing architecture, i.e., CNN+LSTM, as one of our benchmark methods.
- **ResNet**: In [38], a strong baseline model with the ResNet structure is proposed for time series classification, including ECG classification. So, we also use it for comparison.
- **CNN-4**: In [10], a four-layer CNN-based model with a novel pooling layer is proposed to detect sleep apnea from ECG data directly, and we compare it with our proposed method as well.
- **CNN-6**: In [30], several models with a different number of convolutional layers are designed to predict sleep apnea with ECG data. We adopt their best performing one, which contains 6 convolutional layers for our comparison.

We also compare the following approaches to show the improvements of the proposed framework step by step:

- $\mathcal{F}_{ECG} + \mathcal{C}lf$: It employs a CNN-based feature extractor to learn the deep features from the raw ECG data and then classifies the targets by several fully-connected layers. It can be considered as a standard architecture of a naive deep learning model.
- $\mathcal{F}_{ECG} + \mathcal{F}_{RRI} + \mathcal{F}_{RPE} + \mathcal{C}lf$: Besides the raw ECG data, RRI and RPE are used as expert knowledge inputs. A simple concatenation is used to combine the features from ECG, RRI, and RPE. Then the concatenated features are sent for classification.
- $\mathcal{F}_{ECG}+\mathcal{F}_{RRI}+\mathcal{F}_{RPE}+\mathcal{A}_{cross}+\mathcal{C}lf$: Instead of using the simple concatenation, a cross-attention mechanism is proposed to collaboratively fuse the features from ECG, RRI, and RPE.
- $\mathcal{F}_{ECG} + \mathcal{F}_{RRI} + \mathcal{F}_{RPE} + \mathcal{A}_{cross} + \mathcal{P}roj_{SC} + \mathcal{C}lf$: The proposed hybrid loss is used to update the model's parameter by adding an auxiliary projection during training to learn more general and useful features.
- `ConCAD` ($\mathcal{A}ug + \mathcal{F}_{ECG} + \mathcal{F}_{RRI} + \mathcal{F}_{RPE} + \mathcal{A}_{cross} + \mathcal{P}roj_{SC} + \mathcal{C}lf$): Data augmentation is used with the architecture mentioned above to help learn more general and robust features to boost performance.

4.3 Experiment Setup

For the Apnea-ECG dataset, we train all the models, including the proposed `ConCAD` method and benchmark methods, on the released dataset and test them on the withheld dataset. For the MIT-BIH PSG dataset, 10-fold cross-validation is used to examine the performance as there is no predefined training and test set. Moreover, since some existing studies [29,35,40] have shown that adjacent segment information helps analyze the sleep-related problems, the labeled segment with its surrounding ± 2 segments of the ECG data is also included in our study. Thus, we will examine segments of 1 and 5 min on the Apnea-ECG dataset and test segments of 0.5 and 2.5 min on the MIT-BIH PSG dataset. In addition, some of the deep learning-based benchmark methods (i.e., [10,30,38]) are modified by increasing the pooling size and replacing flatten layer with GlobalAveragePooling for the input of 5 min and 2.5 min as they do no have a version to handle data with adjacent segments, and processing very long vector with their original structures exceeds our hardware memory limitation.

The proposed `ConCAD` model is trained by the AMSGrad optimizer [27], and all its parameters are initialized using HeNormal initializer [15]. An initial learning rate of 0.005 is chosen and it decreases to 0.001 after certain epochs (e.g. 200 epochs). Moreover, the L2 regularization is added to the feature extractor $\mathcal{F}$ to prevent the model from overfitting into the noise or artifacts.

4.4 Results and Discussions

We compare the performance of `ConCAD` with other state-of-the-art benchmark methods, and the results are listed in Table 1. Our proposed framework achieves an accuracy of 88.75% with the 1 min segment input and 91.22% with the 5 min segment input on the Apnea-ECG dataset, and 82.50% with the 1 min segment

input and 83.47% with the 5 min segment input on the MIT-BIH PSG dataset, which outperforms other benchmark methods. Besides, we can see deep learning methods can adaptively learn features from a different length of input while the machine learning methods with hand-crafted features are more sensitive to the change of the input length. With the adjacent segments, the deep learning model can learn more effective features for classification tasks.

We also examine the proposed framework step by step to show the effectiveness of each step in Table 2. We can see that the performance can be worse if we simply concatenate the deep features with the expert features as some of the features will help the model make the better judgment possible, while others are likely to act as noise and thereby lead to more errors. With the cross attention module $\mathcal{A}_{cross}$, we can see that the model learns a better-fused feature representation by learning an attention mask synergistically from each other. The new fused feature representation maintains the effective discriminant parts of features while suppressing the irrelevant parts. Specifically, the accuracy improves

Table 1. Accuracy of the proposed framework with other state-of-the-art methods on Apnea-ECG and MIT-BIH PSG datasets.

Methods		Ref	Apnea-ECG		MIT-BIH PSG	
			1 min	5 min	0.5 min	2.5 min
Feature Based Machine Learning (ML)	SVM	[18]	74.57	67.52	70.02	70.30
	RF		74.86	72.30	70.54	68.15
	KNN		71.81	67.80	69.51	68.12
	MLP		74.81	70.59	71.28	71.20
Deep Learning (DL)	LeNet-5	[35]	83.17	87.25	72.49	78.82
	CNN+LSTM	[3]	82.77	86.12	75.80	80.79
	ResNet	[38]	83.57	85.33	77.29	79.23
	CNN-4	[10]	81.65	84.42	73.56	76.92
	CNN-6	[30]	82.12	84.37	79.69	82.25
Proposed method	`ConCAD`		**88.75**	**91.22**	**82.50**	**83.47**

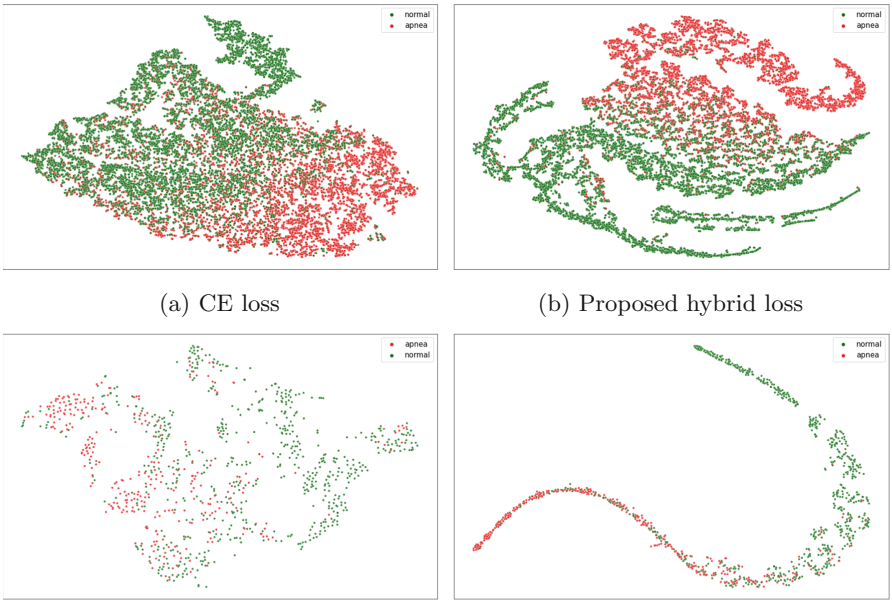
Table 2. Accuracy of different architectures of the proposed framework on Apnea-ECG and MIT-BIH PSG datasets

Architectures	Apnea-ECG		MIT-BIH PSG	
	1 min	5 min	0.5 min	2.5 min
$\mathcal{F}_{ECG} + \mathcal{C}lf$	83.41	85.48	78.83	80.11
$\mathcal{F}_{ECG} + \mathcal{F}_{RRI} + \mathcal{F}_{RPE} + \mathcal{C}lf$	83.23	87.64	79.42	80.60
$\mathcal{F}_{ECG} + \mathcal{F}_{RRI} + \mathcal{F}_{RPE} + \mathcal{A}_{cross} + \mathcal{C}lf$	85.35	89.43	80.22	81.77
$\mathcal{F}_{ECG} + \mathcal{F}_{RRI} + \mathcal{F}_{RPE} + \mathcal{A}_{cross} + \mathcal{P}roj_{SC} + \mathcal{C}lf$	87.16	90.85	81.83	82.83
`ConCAD`	**88.75**	**91.22**	**82.50**	**83.47**

from 83.41% to 85.35% with the 1 min segment input and from 85.48% to 89.43% with the 5 min segment input on the Apnea-ECG dataset. On the MIT-BIH PSG dataset, the accuracy improves from 78.83% to 80.22% with the 0.5 min segment input and from 80.11% to 81.77% with the 2.5 min segment input.

In Table 2, we can also see a further improvement by using the proposed hybrid loss, which takes advantage of both CE loss and SC loss. The accuracy increases to 87.16% with the 1 min segment input and 89.43% with the 5 min segment input on the Apnea-ECG dataset. On the MIT-BIH PSG dataset, the accuracy increases to 81.83% with the 1 min segment input and 81.77% with the 5 min segment input. The hybrid loss boosts the performance by promoting the model to learn more general and discriminant feature representations in case that the model overfits into the training data by learning features with excessive discrimination.

In Fig. 2, the t-SNE plots show the learned feature representation with the CE loss and the proposed hybrid loss. We can see that the hybrid loss promotes to more compact clustering of the instances from the same class while the representation with CE loss is more scattered. We also think the attention module benefits contrastive learning by focusing on parts of features when increasing the agreement among instances in positive pairs and encouraging the difference



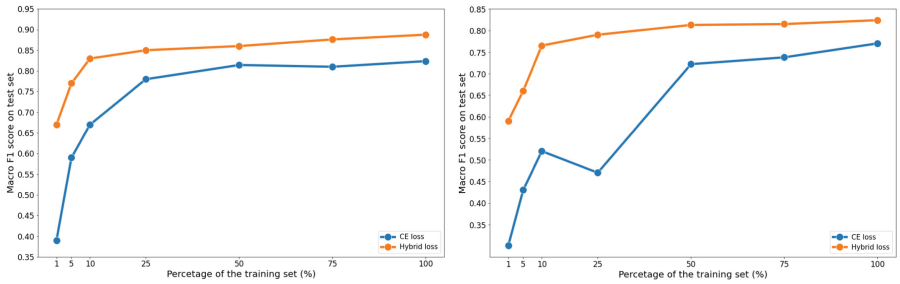
(a) CE loss (b) Proposed hybrid loss

(c) CE loss (d) Proposed hybrid loss

Fig. 2. The t-SNE plots of the fused feature vector on the withheld set of Apnea-ECG (a and b) and the validation set of MIT-BIH PSG (c and d), comparing the cross-entropy (CE) loss with the proposed hybrid loss.

among instances in negative pairs. It is similar to human behavior that human usually tends to recognize an unseen data by comparing the most relevant parts with known ones. By using data augmentation, the proposed model can learn more general feature representation with a more clear boundary and achieve better performance.

In addition, the hybrid loss enables the classification tasks with limited training labeled data. The limitation of labeled data is a prevalent and critical problem in the healthcare field. We train the model by using a fraction of the training set and test it on the entire test set. The results are shown in Fig. 3 in terms of the macro F1 score. F1 score can clearly show the quality of the model when the dataset is imbalanced. With only 1% of the training data, the proposed `ConCAD` model can still achieve an F1 score of 0.67 on the Apnea-ECG dataset and 0.59 on the MIT-BIH PSG dataset. However, the CE loss performs poorly and skews into the majority class. Furthermore, the proposed model only requires 10% of the training data to make a reasonable classification while the naive deep learning model needs more than 50% to get decent performance. Hence, the proposed model with hybrid loss largely outperforms a naive deep learning approach with CE loss on smaller datasets.



(a) Apnea-ECG dataset (b) MIT-BIH PSG dataset

Fig. 3. Impact of number of training data on the performance of the sleep apnea detection with the cross entropy (CE) loss and the proposed hybrid loss.

5 Conclusions and Future Work

In this paper, we propose a contrastive learning-based cross attention framework, named `ConCAD`. The cross attention leverages the expert knowledge and fuses it with deep features by highlighting each other collaboratively. The novel hybrid loss that encompasses the cross-entropy loss and supervised contrastive loss helps learn more robust features by clustering the same class data and pushing apart data of different classes in projection space. Moreover, we show the proposed framework achieves state-of-the-art results on two public ECG

datasets. Furthermore, we show that the proposed framework has better generalization ability with limited labeled training data. We conclude that the ECG data with adjacent segments helps to detect the sleep apnea occurrence through the experiment.

In future work, we plan to study more ECG data augmentation techniques that would help contrastive learning to generate better representations. We also plan to develop more interfaces to allow different formats of expert knowledge (e.g., electronic health records) to be integrated into our framework.

Appendix A

The feature extractor are different for different data and tasks. In this study, we design a CNN-based extractors for ECG, RRI and RPE separately. The structure of the extractor for two dataset are also different as their ECG data have different sampling frequency and noise. The details are shown in the table below. The ConvBlock(number of filters, kernel size, stride) is made of one convolutional layer, one batch normalization layers, one ReLU activation layer (Table 3).

Table 3. The details of the feature extractors used for ECG, RRI and RPE on Apnea-ECG and MIT-BIH PSG.

$\mathcal{F}_{ECG}$ (Apnea-ECG)	$\mathcal{F}_{ECG}$ (MIT-BIH PSG)	$\mathcal{F}_{RRI}$, $\mathcal{F}_{RPE}$
ConvBlock(64,100,20)-MaxPool(2)-Dropout(0.5)-ConvBlock(64,8,4)-MaxPool(2)-Dropout(0.5)-ConvBlock(128,4,2)-MaxPool(2)-Dropout(0.5)-ConvBlock(128,4,2)	ConvBlock(64,60,5)-MaxPool(2)-Dropout(0.5)-ConvBlock(128,8,3)-ConvBlock(128,8,3)-MaxPool(2)-Dropout(0.5)-ConvBlock(256,4,2)-ConvBlock(256,4,2)–MaxPool(2)-Dropout(0.5)-ConvBlock(128,4,1)-ConvBlock(128,4,1)	ConvBlock(64,8,4)-MaxPool(2)-Dropout(0.5)-ConvBlock(64,4,2)-MaxPool(2)-Dropout(0.5)-ConvBlock(128,2,1)-MaxPool(2)-Dropout(0.5)-ConvBlock(128,2,1)

References

1. Al-Abed, M.A., Manry, M., Burk, J.R., Lucas, E.A., Behbehani, K.: Sleep disordered breathing detection using heart rate variability and r-peak envelope spectrogram. In: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 7106–7109. IEEE (2009)
2. Almuhammadi, W.S., Aboalayon, K.A., Faezipour, M.: Efficient obstructive sleep apnea classification based on eeg signals. In: 2015 Long Island Systems, Applications and Technology, pp. 1–6. IEEE (2015)
3. Almutairi, H., Hassan, G.M., Datta, A.: Detection of obstructive sleep apnoea by ecg signals using deep learning architectures. In: 2020 28th European Signal Processing Conference (EUSIPCO), pp. 1382–1386. IEEE (2021)
4. Altevogt, B.M., Colten, H.R., et al.: Sleep Disorders and Sleep Deprivation: An Unmet Public Health roblem. National Academies Press (2006)
5. Banville, H., Chehab, O., Hyvarinen, A., Engemann, D., Gramfort, A.: Uncovering the structure of clinical eeg signals with self-supervised learning. J. Neural Eng. **18**, 046020 (2020)

6. Bloch, K.E.: Polysomnography: a systematic review. Technol. Health Care **5**(4), 285–305 (1997)
7. Chen, L., Zhang, X., Song, C.: An automatic screening approach for obstructive sleep apnea diagnosis based on single-lead electrocardiogram. IEEE Trans. Autom. Sci. Eng. **12**(1), 106–115 (2014)
8. Chu, Q., Ouyang, W., Li, H., Wang, X., Liu, B., Yu, N.: Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4836–4845 (2017)
9. De Chazal, P., Heneghan, C., Sheridan, E., Reilly, R., Nolan, P., O'Malley, M.: Automatic classification of sleep apnea epochs using the electrocardiogram. In: Computers in Cardiology 2000, vol. 27 (Cat. 00CH37163), pp. 745–748. IEEE (2000)
10. Dey, D., Chaudhuri, S., Munshi, S.: Obstructive sleep apnoea detection using convolutional neural network based deep learning framework. Biomed. Eng. Lett. **8**(1), 95–100 (2017). https://doi.org/10.1007/s13534-017-0055-y
11. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1933–1941 (2016)
12. Goldberger, A.L., et al.: Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation **101**(23), e215–e220 (2000)
13. Hamilton, P.: Open source ecg analysis. In: Computers in Cardiology, pp. 101–104. IEEE (2002)
14. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738 (2020)
15. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)
16. Hou, R., Chang, H., Ma, B., Shan, S., Chen, X.: Cross attention network for few-shot classification. arXiv preprint arXiv:1910.07677 (2019)
17. Ichimaru, Y., Moody, G.: Development of the polysomnographic database on cd-rom. Psychiatry Clin. Neurosci. **53**(2), 175–177 (1999)
18. Jezzini, A., Ayache, M., Elkhansa, L., Al Abidin Ibrahim, Z.: ECG classification for sleep apnea detection. In: 2015 International Conference on Advances in Biomedical Engineering (ICABME), pp. 301–304. IEEE (2015)
19. Kapur, V.K., et al.: Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: an American academy of sleep medicine clinical practice guideline. J. Clin. Sleep Med. **13**(3), 479–504 (2017)
20. Khosla, P., et al.: Supervised contrastive learning. arXiv preprint arXiv:2004.11362 (2020)
21. Lin, C.J., Lin, C.H., Jeng, S.Y.: Using feature fusion and parameter optimization of dual-input convolutional neural network for face gender recognition. Appl. Sci. **10**(9), 3166 (2020)
22. Mohla, S., Pande, S., Banerjee, B., Chaudhuri, S.: Fusatnet: dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 92–93 (2020)

23. Nikolaidis, K., Kristiansen, S., Goebel, V., Plagemann, T., Liestøl, K., Kankanhalli, M.: Augmenting physiological time series data: a case study for sleep apnea detection. In: Brefeld, U., Fromont, E., Hotho, A., Knobbe, A., Maathuis, M., Robardet, C. (eds.) ECML PKDD 2019. LNCS (LNAI), vol. 11908, pp. 376–399. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-46133-1_23
24. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
25. Ouali, Y., Hudelot, C., Tami, M.: Spatial contrastive learning for few-shot classification. arXiv preprint arXiv:2012.13831 (2020)
26. Penzel, T., Moody, G.B., Mark, R.G., Goldberger, A.L., Peter, J.H.: The apnea-ecg database. In: Computers in Cardiology 2000, vol. 27 (Cat. 00CH37163), pp. 255–258. IEEE (2000)
27. Reddi, S.J., Kale, S., Kumar, S.: On the convergence of adam and beyond. arXiv preprint arXiv:1904.09237 (2019)
28. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**(6088), 533–536 (1986)
29. Supratak, A., Dong, H., Wu, C., Guo, Y.: Deepsleepnet: a model for automatic sleep stage scoring based on raw single-channel eeg. IEEE Trans. Neural Syst. Rehabil. Eng. **25**(11), 1998–2008 (2017)
30. Urtnasan, E., Park, J.U., Joo, E.Y., Lee, K.J.: Automated detection of obstructive sleep apnea events from a single-lead electrocardiogram using a convolutional neural network. J. Med. Syst. **42**(6), 1–8 (2018)
31. Van Steenkiste, T., Groenendaal, W., Deschrijver, D., Dhaene, T.: Automated sleep apnea detection in raw respiratory signals using long short-term memory neural networks. IEEE J. Biomed. Health Inf. **23**(6), 2354–2364 (2018)
32. Varon, C., Caicedo, A., Testelmans, D., Buyse, B., Van Huffel, S.: A novel algorithm for the automatic detection of sleep apnea from single-lead ecg. IEEE Trans. Biomed. Eng. **62**(9), 2269–2278 (2015)
33. Wang, F., et al.: Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2017)
34. Wang, F., Liu, H.: Understanding the behaviour of contrastive loss. arXiv preprint arXiv:2012.09740 (2020)
35. Wang, T., Lu, C., Shen, G., Hong, F.: Sleep apnea detection from a single-lead ecg signal with automatic feature-extraction through a modified lenet-5 convolutional neural network. PeerJ **7**, e7731 (2019)
36. Wang, W., Zhou, T., Yu, F., Dai, J., Konukoglu, E., Van Gool, L.: Exploring cross-image pixel contrast for semantic segmentation. arXiv preprint arXiv:2101.11939 (2021)
37. Wang, Y., Huang, M., Zhu, X., Zhao, L.: Attention-based lstm for aspect-level sentiment classification. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 606–615 (2016)
38. Wang, Z., Yan, W., Oates, T.: Time series classification from scratch with deep neural networks: a strong baseline. In: 2017 International Joint Conference on Neural Networks (IJCNN), pp. 1578–1585. IEEE (2017)
39. Wu, L., Wang, Y., Li, X., Gao, J.: What-and-where to match: deep spatially multiplicative integration networks for person re-identification. Pattern Recogn. **76**, 727–738 (2018)
40. Yadollahi, A., Moussavi, Z.: Acoustic obstructive sleep apnea detection. In: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 7110–7113. IEEE (2009)

41. Young, T., Peppard, P.E., Gottlieb, D.J.: Epidemiology of obstructive sleep apnea: a population health perspective. Am. J. Resp. Crit. care Med. **165**(9), 1217–1239 (2002)
42. Zhang, Z., Sabuncu, M.R.: Generalized cross entropy loss for training deep neural networks with noisy labels. arXiv preprint arXiv:1805.07836 (2018)