Deep Attention Networks With Multi-Temporal Information Fusion for Sleep Apnea Detection

Meng Jiao, Changyue Song, Xiaochen Xian, Shihao Yang, *Graduate Student Member, IEEE*, and Feng Liu

***Abstract*—Sleep Apnea (SA) is a prevalent sleep disorder with multifaceted etiologies that can have severe consequences for patients. Diagnosing SA traditionally relies on the in-laboratory polysomnogram (PSG), which records various human physiological activities overnight. SA diagnosis involves manual scoring by qualified physicians. Traditional machine learning methods for SA detection depend on hand-crafted features, making feature selection pivotal for downstream classification tasks. In recent years, deep learning has gained popularity in SA detection due to its capability for automatic feature extraction and superior classification accuracy. This study introduces a Deep Attention Network with Multi-Temporal Information Fusion (DAN-MTIF) for SA detection using single-lead electrocardiogram (ECG) signals. This framework utilizes three 1D convolutional neural network (CNN) blocks to extract features from R-R intervals and R-peak amplitudes using segments of varying lengths. Recognizing that features derived from different temporal scales vary in their contribution to classification, we integrate a multi-head attention module with a self-attention mechanism to learn the weights for each feature vector. Comprehensive experiments and comparisons between two paradigms of classical machine learning approaches and deep learning approaches are conducted. Our experiment results demonstrate that (1) compared with benchmark methods, the proposed DAN-MTIF exhibits excellent performance with 0.9106 accuracy, 0.9396 precision, 0.8470 sensitivity, 0.9588 specificity, and 0.8909 F_1 score at per-segment level; (2) DAN-MTIF can effectively extract features with a higher degree of discrimination from ECG segments of multiple timescales than those with a single time scale, ensuring a better SA detection performance; (3) the overall performance of deep learning methods is better than the classical machine learning algorithms, highlighting the superior performance of deep learning approaches for SA detection.**

***Index Terms*—Sleep apnea (SA), electrocardiogram, deep learning, convolutional neural network (CNN), multi-head attention.**

***Impact Statement*—To facilitate automatic detection of Sleep Apnea (SA) events based on single-lead electrocardiogram (ECG) signals, we introduced a Deep Attention Network with Multi-Temporal Information Fusion (DAN-MTIF) leveraging ECG signal of multiple temporal scales. The proposed model exhibits superior performance compared to classical machine learning algorithms and deep learning approaches.**

I. Introduction

SLEEP accounts for about one-third of people's daily lives [1], [2]. Sleep quality is closely related to physical and mental health, but unfortunately, nearly 50 to 70 million adults in the United States are afflicted with various sleep disorders [2]. Sleep Apnea (SA) is a common and serious sleep disorder characterized by brief interruptions (10 seconds or more) of breathing during sleep [3]. There are three main types of SA: Obstructive Sleep Apnea (OSA), Central Sleep Apnea (CSA), and Mixed Sleep Apnea (MSA) [3], [4]. The cause of OSA is that muscles in the back of the throat are overly relaxed, blocking or narrowing the airway so that breathing is obstructed. CSA, on the other hand, is caused by the temporary inability of the brain to send signals to muscles that control breathing. MSA refers to a condition in which the patient experiences both OSA and CSA at the same time. Patients with SA suffer both physically and psychologically, as SA causes them to snore loudly during sleep, wake up repeatedly at night, and feel tired or even exhausted during daytime, etc [5]. Those in severe SA conditions even have a much higher risk of cardiovascular disease, mental health illnesses, and psychological distress [6], [7], [8]. The clinical indicators for diagnosing SA vary [9]. Initially, Guilleminault defined "sleep apnea syndrome" as more than 30 apnea events per night [10]. Later, the "Apnea-Hypopnea Index (AHI)", which refers to the number of minutes containing apnea events per hour, was adopted as a proper measurement. Most clinicians regard an AHI below 5 as normal, and an AHI of 10 or more as pathological [11]. Moreover, criteria used in current practice are based not just on AHI but also encompass symptoms and cardiovascular outcomes [9].

As the prevalence of SA has risen substantially over the past few decades, posing an increasing threat to public health

Manuscript received 3 November 2023; revised 25 March 2024, 6 May 2024, and 16 May 2024; accepted 17 May 2024. Date of publication 27 May 2024; date of current version 13 September 2024. This work was supported by the NIBIB of the National Institutes of Health (NIH) under Award R21EB033455. The review of this article was arranged by Editor Michael Khoo. *(Corresponding author: Feng Liu.)*

Meng Jiao, Shihao Yang, and Feng Liu are with the Department of Systems and Enterprises, Stevens Institute of Technology, Hoboken, NJ 07030 USA (e-mail: fliu22@stevens.edu).

Changyue Song is with the Independent Researcher, Irvine, CA 92602 USA.

Xiaochen Xian is with the Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611 USA.

Digital Object Identifier 10.1109/OJEMB.2024.3405666

and safety, it is essential to establish an efficient diagnostic method for SA [12]. A widely accepted gold standard is the polysomnogram (PSG), which involves the overnight recording of different physiologic activities during sleep, for example, brain waves, heart rhythm, eye movements, blood oxygen level, and so on [13], [14]. Typically, the individual who underwent PSG is required to stay overnight in a sleep laboratory monitored by sleep technicians and physicians. Then, the physiological recordings need to be manually annotated and interpreted to ensure detailed and precise results, which limits PSG's utility outside of a laboratory setting [15]. Other factors limiting the wide adoption of PSG include its extensive duration of recording and substantial cost, and the patients are prone to feel uncomfortable due to multiple sensors attached to their skin [16]. In contrast, the increasing utilization of auto-scoring home sleep monitoring devices, characterized by their wearability and portability, has the potential to help mitigate the physician shortage in hospitals while offering enhanced convenience and comfort to patients [17], [18]. These electronics have either single or multiple sensors, which can monitor one or more physiological signals independently or in combination [19], [20]. The commonly measured signals include electroencephalogram (EEG), electrocardiogram (ECG), blood oxygen level, and respiration [21], [22]. Among all these physiological measurements, ECG achieves the best signal-to-noise ratio with a signal strength of 1–2 mV, and the acquisition of ECG is less technical when compared to PSG [23]. In addition, various studies have indicated that sleep distinctly influences ECG in various ways. For example, a notable correlation has been observed between heart rate variability (HRV) and SA events [24], [25]. Due to the above advantages, the use of ECG for SA detection has been extensively studied in recent years. A brief summary and introduction of methods developed during the past decades is provided in Section II.

In this paper, we propose a novel deep network incorporating information at different timescales of SA events through feature fusion, which aims to make full use of the discriminative information implicit in ECG measurements before and after the occurrence of SA. The main contributions of this study are outlined as follows:

1) The proposed deep attention network with multiple temporal information fusion (DAN-MTIF) for SA detection extracts features from R-R intervals and R-peak amplitudes of multiple timescales, enabling the utilization of the information contained in adjacent ECG segments.
2) A dilated convolution filter is used in the convolutional filters, which allows the convolutional filters to extract features from a wider range of the receptive field, allowing better capturing of temporal dependencies without increasing the depth of neural networks. In addition, the multi-head attention module to assign proper weights for features derived from different channels is incorporated, enabling the adaptive integration of information at different timescales.
3) Comprehensive evaluations of the proposed DAN-MTIF framework and its comparison with benchmark methods, including both ML models and DL models, are conducted. The results suggest that DAN-MTIF outperforms the benchmark methods, while deep learning models outperform the classical ML approaches

II. Related Work

The proposed methods for SA events detection from single channel ECG can be mainly divided into two categories respectively, represented by classical machine learning (ML) approaches and deep learning (DL) approaches [26].

ML approaches typically require extracting discriminative features from the recorded ECG data as they are rich in respiratory information about the patients and inevitably contain irrelevant information that cannot reflect whether a patient is in SA condition during sleep [27], [28], [29]. To solve this problem, principal component analysis (PCA) or other dimensionality reduction algorithms are often employed to extract the most discriminative features [30]. The obtained features after dimensionality reduction are then used as the input and fed into a classifier for SA detection [31]. During the past few decades, various ML models have been applied to SA detection. For example, Varon et al. presented a least-squares support vector machine (LS-SVM) for the automatic detection of SA from single-lead ECG, in which the standard deviation and the serial correlation coefficients of the R-R interval time series are used as two input features [32]. Rizal et al. performed SA classification based on a support vector machine (SVM) using eleven HRV features and achieved an accuracy of about 89.5% [33]. Moreover, Salari et al. employed a set of basic popular classifiers, including logistic regression (LR), linear discriminant analysis (LDA), and K-nearest neighbors (KNN), etc., for SA detection and then conducted a comprehensive analysis of the model performance [34].

Recently, with the popularity of artificial neural networks (ANN), various DL frameworks have been developed for SA detection, mainly including convolution neural networks (CNNs), recurrent neural networks (RNNs), hybrids of CNN and RNN architectures [35], [36]. The advantage of DL models over ML models is the latter can automatically extract discriminative features from ECG signals. For example, Wang et al. [37] proposed a modified LeNet-5 CNN for SA detection with R-peak amplitudes and R-R intervals derived from ECG signals as input and achieved better or comparable results when compared with traditional ML methods. Niroshana et al. [38] proposed an image-based method to detect SA events, in which ECG segments were first converted into scalogram images and spectrogram images, then fed into a two-dimensional CNN (2D-CNN) network. Furthermore, to utilize the temporal information of ECG, a variety of RNN-based frameworks and hybrids of CNN and RNN have been proposed during the past several years. Faust et al. [39] presented an automatic SA detection method based on a long short-term memory (LSTM) network with RR intervals as input. Liang et al. [40] designed a deep network using the combination of CNN and LSTM structures to detect OSA events from R-R intervals. What's more, Zarei et al. [41] also presented a hybrid CNN-LSTM framework for OSA detection from single-lead ECG signals, and the presented model significantly outperforms existing state-of-the-art methods. Chen

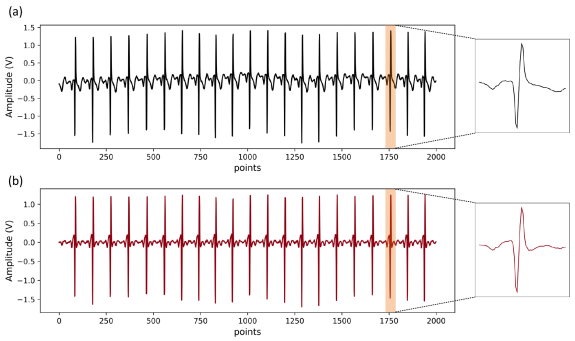

Fig. 1. An illustrative example of preprocessing to denoise: (a) raw ECG signal; (b) filtered ECG signal.

et al. [42] developed a lightweight multi-scaled neural network called SE-MSCNN for SA detection based on single-lead ECG signals and achieved significantly better performance compared to state-of-the-art SA detection methods. Overall, extensive published works demonstrate the potential of DL models in SA detection based on ECG signals [35], [43].

III. Material and Method

A. Data Preprocessing

In general, raw ECG recordings are vulnerable to noises from multiple sources, such as baseline drift, motion artifacts, and electromagnetic interference. These disruptions can be attributed to patients' movements and the data acquisition equipment used in the recording process. In order to mitigate noise contamination from ECG, we perform signal filtering using a bandpass filter with a cutoff frequency of 3 Hz to 45 Hz. The comparison between original and filtered ECG signals is given in Fig. 1. It can be observed that the baseline drift and other high-frequency disturbances are removed effectively after band-pass filtering. Next, all the filtered overnight ECG recordings are segmented into short segments of 1-minute length to detect patients' SA status with varying time stages during sleep. Two ECG features, R-peak amplitudes and R-R intervals, are employed as the model input for SA classification. The Hamilton algorithm [44] is adopted to detect R-peaks, then R-peak positions are further used to extract R-peak amplitudes and compute R-R intervals. Since the heart rate of the subject changes during sleep, the extracted R-peak amplitudes and R-R intervals are of varied lengths. In order to make the derived features meet the input requirements of the model, we fixed the feature length corresponding to 1-, 3-, 5-min long segments to 180, 540, 900 points, respectively. For features with fewer points, the cubic interpolation technique is introduced to extend length [37]. For features with more points, we cropped them at the center to reduce length. The adjacent ECG segments at different timescales are also used for SA detection. The example of adjacent ECG segments and the corresponding R-peak amplitudes and R-R intervals are shown in Fig. 2.

B. The Proposed Architecture

In this section, a unique SA detection framework (DAN-MTIF) based on multi-scaled CNN and multi-head attention

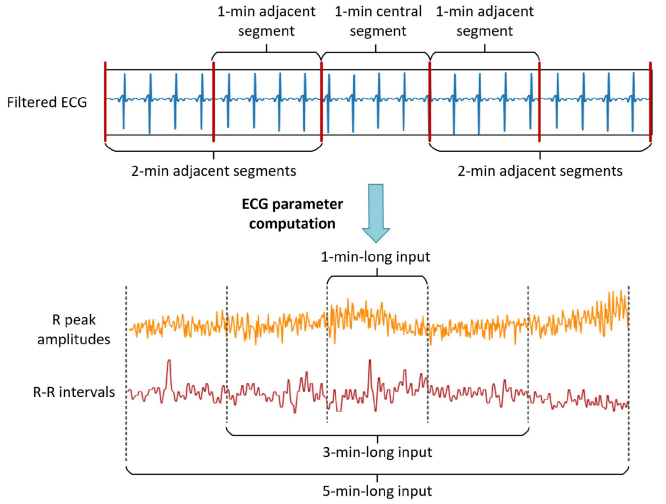

Fig. 2. An illustration of example R-R intervals and R-peak amplitudes.

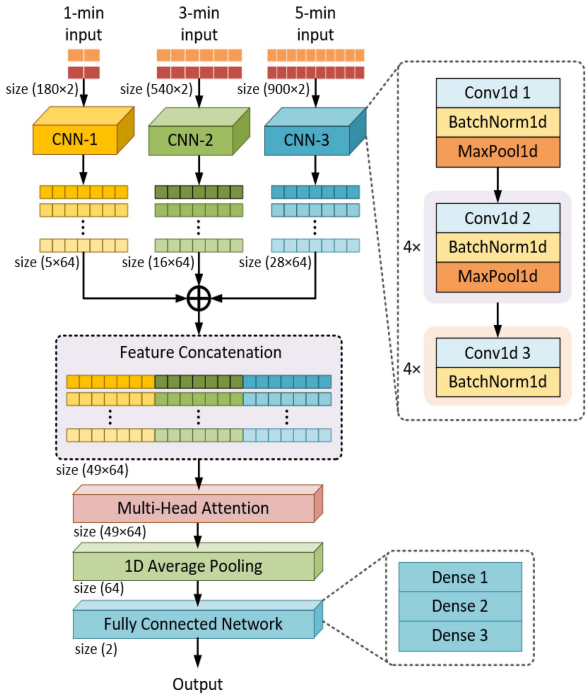

Fig. 3. The structure of the proposed framework.

mechanism is designed for automatic SA detection using single-lead ECG signals. The architecture and configuration details of the proposed DAN-MTIF model are illustrated in Fig. 3 and Table I, which consists of three modules: multi-scaled CNN module, multi-head attention module, and a fully connected network as the SA classifier.

1) Multi-Temporal Scale CNN Module: As illustrated in Fig. 3, the multi-scaled CNN module consists of three CNN blocks (CNN-1, CNN-2, CNN-3), of which the inputs are the derived R-R intervals and R-peak amplitudes with different lengths (180, 540, 900 points) corresponding to ECG segments with different timescales (1-, 3-, 5-min long). This unique design enables the DAN-MTIF model to utilize the information from both central and adjacent ECG segments simultaneously. Moreover, the dilated convolution kernel [45] was employed to expand

TABLE I
DETAILED EXPLANATION OF EACH LAYER IN THE PROPOSED DAN-MTIF

Blocks	Layers	Configuration Details	Parameters
CNN-1	Conv1d 1	in_channels=2, out_channels=64, kernel_size=3, stride=1, padding='same', dilation=1	448
	Conv1d 2	in_channels=64, out_channels=64, kernel_size=3, stride=1, padding='same', dilation=2	12352
	Conv1d 3	in_channels=64, out_channels=64, kernel_size=3, stride=1, padding='same', dilation=2	12352
	MaxPool1d	kernel_size=2, stride=2, padding=0, dilation=1	-
	BatchNorm1d	num_features=64	128
CNN-2	Conv1d 1	in_channels=2, out_channels=64, kernel_size=7, stride=1, padding='same', dilation=1	960
	Conv1d 2	in_channels=64, out_channels=64, kernel_size=7, stride=1, padding='same', dilation=2	28736
	Conv1d 3	in_channels=64, out_channels=64, kernel_size=7, stride=1, padding='same', dilation=2	28736
	MaxPool1d	kernel_size=2, stride=2, padding=0, dilation=1	-
	BatchNorm1d	num_features=64	128
CNN-3	Conv1d 1	in_channels=2, out_channels=64, kernel_size=11, stride=1, padding='same', dilation=1	1472
	Conv1d 2	in_channels=64, out_channels=64, kernel_size=11, stride=1, padding='same', dilation=2	45120
	Conv1d 3	in_channels=64, out_channels=64, kernel_size=11, stride=1, padding='same', dilation=2	45120
	MaxPool1d	kernel_size=2, stride=2, padding=0, dilation=1	-
	BatchNorm1d	num_features=64	128
Attention	-	Q_i,K_i,V_i: in_features=64, out_features=32 (d=49, c=64, d_q=64, d_k=64, d_v=64, h=2)	2080
	-	O: in_features=64, out_features=64 (d_{out}=64)	4160
Pooling	AvgPool1d	kernel_size=49	-
SA Classifier	Dense 1	in_features=64, out_features=64 (ReLU activation)	4160
	Dense 2	in_features=64, out_features=32 (ReLU activation)	2080
	Dense 3	in_features=32, out_features=2	66

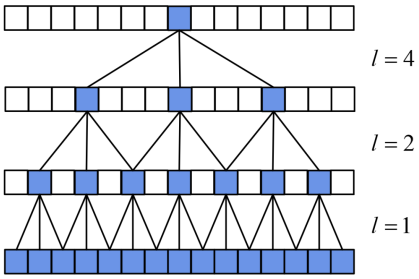

Fig. 4. Illustration of the receptive field for the dilated convolution.

the receptive field of DAN-MTIF. The receptive field refers to the area size of the original input corresponding to a single element in the extracted feature map. Shallow CNNs tend to have small receptive fields as the extracted features contain more local information instead of global information. As the depth of CNN increases, the proportion of global information contained in the feature map increases accordingly, so that the receptive field expands. As illustrated in Fig. 4, for a dilated convolution layer, there are empty "spaces" between kernel elements, which allows the convolutional kernel to directly extract information from a wider range without increasing the number of parameters. In this way, the receptive field can be expanded exponentially without increasing network depth, thus avoiding the increase of model parameters. The dilation rate l is defined to indicate how much the kernel can be expanded. The normal convolution can be seen as a dilated convolution with $l = 1$.

In this study, the three blocks in the multi-temporal scale CNN module share the same structure, and each block contains 8 1D-CNN layers with 64 output channels. The dilation rate l is set to 2, and the kernel sizes are respectively set to 3×1, 7×1, and 11×1, corresponding to three timescales. The feature vectors extracted from a single channel in three CNN blocks were first concatenated into a long vector x containing information from multiple timescales, then the concatenated feature vectors corresponding to different output channels were stacked together producing an output X with a size of $d \times c$, where d is the dimension of the concatenated feature vector x, and c is the number of channels in each CNN block.

2) Multi-Head Attention Module: Since the concatenated feature vectors contain information from multiple sources, which are of varied contributions to the classification results, a multi-head attention module with a self-attention mechanism is employed to assign a proper weight for each feature vector.

With the concatenated feature X as input, the attention score based on self-attention can be computed as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where $Q \in \mathbb{R}^{d\times d_k}, K \in \mathbb{R}^{d\times d_k}, V \in \mathbb{R}^{d\times d_v}$ are three matrices known as the *Query*, *Key* and *Value*, which can be obtained respectively by multiplying $X \in \mathbb{R}^{d\times c}$ with three learnable nonlinear matrices $W^Q \in \mathbb{R}^{c\times d_k}$, $W^K \in \mathbb{R}^{c\times d_k}$, and $W^V \in \mathbb{R}^{c\times d_v}$. d_k is the dimension of the key vector in K, and d_v is the dimension of the value vector in V.

In the multi-head attention module, the Query, Key, Value matrices derived from the concatenated feature X get logically split across multiple heads, which enables the structure to capture richer interpretations from X. The attention score can be computed as follows:

$$head_i = Attention(Q_i, K_i, V_i), i = 1, 2, \ldots, h. \quad (2)$$

$$MultiHead = Concat(head_1, \ldots, head_h)W^O \quad (3)$$

where h is the number of heads, $Q_i \in \mathbb{R}^{d\times \frac{d_k}{h}}, K_i \in \mathbb{R}^{d\times \frac{d_k}{h}}, V_i \in \mathbb{R}^{d\times \frac{d_v}{h}}$ can be obtained respectively by multiplying $X \in \mathbb{R}^{d\times c}$ with projection matrices $W_i^Q \in \mathbb{R}^{c\times \frac{d_k}{h}}$, $W_i^K \in \mathbb{R}^{c\times \frac{d_k}{h}}$, and $W_i^V \in \mathbb{R}^{c\times \frac{d_v}{h}}$. $W^O \in \mathbb{R}^{d_v\times d_{out}}$ is the output matrix, and d_{out} is the dimension of the outputs. By introducing multi-head attention, the model is enabled to simultaneously focus on different parts of the input and potentially capture various types of dependencies and relationships within the data.

By applying the generated attention score to the extracted feature X, the proposed DAN-MTIF framework is enabled to adaptively make full use of information from ECG segments at different timescales. Finally, the features obtained from the

multi-head attention module are fed into a global average pooling layer and followed by a fully connected (FC) module for SA classification.

IV. Experiments and Results

In this section, we conducted experiments to validate the effectiveness of the proposed DAN-MTIF model for SA detection.

A. Dataset and Evaluation Metrics

The ECG data used in this study is from the PhysioNet Apnea-ECG database provided by Philipps University [46]. This database contains 70 single-lead ECG recordings collected from 35 individuals with a sampling frequency of 100 Hz. The time duration of each ECG recording varies between 401 min to 578 min and each recording was broken down into a set of 1-min long ECG segments. Each ECG segment was annotated as A (Apnea) or N (Normal) by a human expert, indicating whether apnea occurred at that time. In this dataset, the proportions of normal and apnea segments are 62% (N) and 38% (A), respectively. These ECG recordings were further classified into three classes: Class A (Apnea), Class B (Borderline), and Class C (Control). The number of recordings in each category is 40, 10, and 20, respectively. Since the ECG recordings in Class B meet part rather than all of the SA criteria used in current practice, which means it is difficult to classify them unambiguously, following other researchers who studied this dataset, only recordings from Class A and Class C were employed for SA detection. These recordings were then divided into a learning dataset and a testing dataset each containing 20 class A recordings and 10 class C recordings. 70% of the ECG segments in the learning set were used for model training and 30% for model validation. Further, all the ECG segments in the testing set were applied for model testing.

For classical ML models, the HRV features [47], including time domain features, frequency domain features, and nonlinear features, are employed for model learning. In this study, the HRV features were extracted using the *hrvanalysis* package [48], [49], [50], which provides methods to remove outliers and ectopic beats from R-R intervals to get normal to normal intervals (NN-intervals). Then, the HRV features can be calculated from the filtered NN intervals. To drop redundant features, the random forest (RF) [51] and the extreme gradient boosting (XGBoost) [52] are adopted to calculate the permutation feature importance. The permutations are conducted five times for each model, and the mean values are used for the final ranking (Fig. 5). According to the ranking results, the first 25 features are used for ML model learning, and the grid search technique is adopted for hyperparameter tuning.

The benchmark algorithms used for comparison include (i) classical ML models: LR [53], LDA [54], KNN [55], SVM [56], LS-SVM [57], RF [51], gradient-boosting decision tree (GBDT) [58], XGBoost [52] and (ii) DL models: LeNet-5 [59], ZFNet [60], AlexNet [61], GRU [62], LSTM [63], BiLSTM [64], CNN-LSTM [26], SE-MSCNN [42]. For LR, we use the regularization strength to be 10. For KNN, we use the Manhattan distance, and the number of neighbors K is set to 70. For SVM, the RBF kernel with a kernel coefficient of

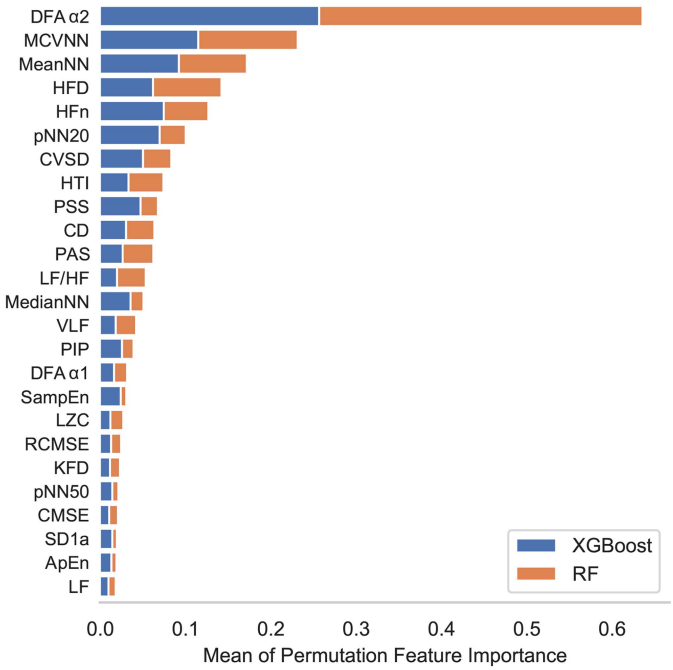

Fig. 5. The HRV feature ranking for SA detection.

0.1 is adopted for nonlinear mapping, and the regularization strength of 10 is used. For LS-SVM, an RBF kernel with a kernel coefficient of 0.1 is adopted for nonlinear mapping and the regularization strength of 0.1 is used. For RF, the number of trees is set to 100, with the maximum depth being 8. The function of measuring the quality of a split is the Gini impurity. For GBDT, the maximum depth of the individual DT is set to 2, and the learning rate used to shrink the contribution of each tree is set to 0.1. The number of boosting stages to perform is 100. For XGBoost, the maximum depth of a tree is set to 2, the number of trees in the ensemble is set to 90, and the learning rate is set to 0.1.

The detailed configuration of the deep learning benchmark methods is given below:

1) *LeNet-5:* The LeNet-5 is comprised of 32 kernels of size 5×1 stacked to a 3×1 max-pooling, followed by a convolution layer with 64 kernels of size 5×1, followed by a 3×1 max-pooling and a Dropout layer with a rate of 0.2. Data were then flattened and fed to two fully connected layers with 32 and 2 nodes.
2) *ZFNet:* The ZF-Net is comprised of 96 kernels of size 7×1 stacked to a 3×1 max-pooling followed by a batch normalization in the first layer, a convolution layer with 256 kernels of size 5×1 followed by a 3×1 max-pooling and batch normalization in the second layer, and three convolution layers with 512 kernels of size 3×1, 1024 kernels of size 3×1, and 512 kernels of size 3×1 stacked together followed by a 3×1 max-pooling in the third layer. Data were then flattened and fed to two fully connected layers with 418 and 2 nodes.
3) *AlexNet:* The developed AlexNet comprised of a convolution layer with 96 kernels of size 11×1 stacked to a batch normalization layer followed by a 3×1 max-pooling operation in the first layer, a convolution layer with 256

kernels of size 5×1 stacked to a batch normalization layer followed by a 3×1 max-pooling operation in the second layer, and two convolution layers with 384 kernels of size 3×1, and a convolution layer with 256 kernels of size 3×1 each stacked to a batch normalization layer followed by a 3×1 max-pooling in the third layer. Data were then flattened and fed to two fully connected layers with 209 and 2 nodes.

4) *LSTM:* LSTM is a novel RNN architecture designed to avoid long-term dependencies with 2 "cell states" and three "gates". In this study, the developed LSTM model comprised two hidden layers with 90 and 22 LSTM units. The outputs of the LSTM block were then flattened and fed to two fully connected layers with 256 and 2 nodes.
5) *BiLSTM:* BiLSTM is an extension of LSTM with two layers of LSTM units processing information from opposite directions. In this study, the developed BiLSTM model comprised two bidirectional hidden layers with 90 and 22 LSTM units in each layer. The outputs of the BiLSTM block were then flattened and fed to two fully connected layers with 256 and 2 nodes.
6) *GRU:* GRU is a variant of LSTM with fewer gates and fewer parameters. In this study, the developed GRU model comprised two hidden layers with 90 and 22 GRU units. The outputs of the GRU block were then flattened and fed to two fully connected layers with 256 and 2 nodes.
7) *CNN-LSTM:* CNN-LSTM is a hybrid model comprised of a convolution layer with 32 kernels of size 5×1 followed by a 3×1 max-pooling operation in the first layer, a convolution layer with 64 kernels of size 5×1 followed by a 3×1 max-pooling operation in the second layer, and an LSTM layer with 16 units in the third layer. The outputs of the LSTM layer were then flattened and fed to two fully connected layers with 64 and 2 nodes.
8) *SE-MSCNN:* SE-MSCNN is a multi-scaled fusion network comprised of 3 CNN modules, and each module includes two convolution layers with 16 and 24 kernels of size 11×1 in the first and second layer followed by a 3×1 max-pooling operation in the third layer. In the first and second CNN modules, the fourth layer is a convolution layer with 32 kernels of size 11×1, and in the third CNN module, the fourth layer is a convolution layer with 32 kernels of size 1×1. The extracted feature vectors from three CNN modules were stacked together and further used to produce proper weights through a channel-wise attention module. The weighted features were then fed to two fully connected layers with 96 and 2 nodes.

For DL models, the cross-entropy loss is chosen as the loss function, and the Adam optimizer is used for model optimization. The number of epochs is set to 100, the batch size is set to 32, and the learning rate is set to 0.001, which will be reduced by a factor of 0.2 if there is no improvement in the model after 2 epochs. All the experiments were conducted on a Windows PC with an i9 CPU and 64 GB memory, and NVIDIA V100 with 32 GB memory was used to train DL models. The metrics used to quantitatively evaluate the performance of each algorithm for SA detection include accuracy (Acc), precision ($Prec$), sensitivity ($Sens$), specificity ($Spec$), and F_1, which are defined as below:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Prec = \frac{TP}{TP + FP} \tag{5}$$

$$Sens = \frac{TP}{TP + FN} \tag{6}$$

$$Spec = \frac{TN}{FP + TN} \tag{7}$$

$$F_1 = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{8}$$

where TP is the number of times that the model correctly predicts the positive class, TN is the number of times that the model correctly predicts the negative class, FP is the number of times that the model incorrectly predicts the positive class and FN is the number of times that the model incorrectly predicts the negative class.

Here, the ECG segment/recording corresponding to the SA state is classified into the positive class, and the ECG segment/recording corresponding to the normal state is classified into the negative class. Better performance for SA detection is expected if these metrics are close to 1.

In this study, the indicator used to judge whether an individual is under the SA condition is AHI:

$$\text{AHI} = \frac{60}{T} \times N \tag{9}$$

where T is the number of 1-min long ECG segments, N is the number of ECG segments which is under SA condition. AHI refers to the average number of 1-min long SA segments per hour. If AHI is greater than or equal to 5, the individual was regarded as under SA condition; otherwise, under normal condition.

B. Classification Performance

The performance comparison between the proposed DAN-MTIF and benchmark algorithms in per-segment SA detection is summarized in Table II and Fig. 6. The correlation between the AHI detected by all algorithms and the real AHI of ECG recordings in the test set are given in Fig. 7. The confusion matrices corresponding to DAN-MTIF on per-segment and per-recording SA detection are shown in Fig. 8. It can be seen from the results that:

- The best performance of classical ML models is achieved by LS-SVM with 0.8332 accuracy, 0.7964 precision, 0.8236 sensitivity, 0.8405 specificity, and 0.8098 F_1 at per-segment level. This indicates that the selected HRV features are not representative enough to accurately distinguish normal and apnea ECG segments.
- DL models significantly outperform classical ML models in SA detection. The proposed DAN-MTIF exhibits the best performance among all models with 0.9106 accuracy, 0.9396 precision, 0.8470 sensitivity, 0.9588 specificity, and 0.8909 F_1 at per-segment level. Moreover, according to the confusion matrices, DAN-MTIF achieved 100% in

TABLE II
PERFORMANCE COMPARISON OF ML MODELS AND DL MODELS AT PER-SEGMENT LEVEL

	Model	Acc	Prec	Sens	Spec	F1
ML	LR	0.7415	0.6349	0.9424	0.5893	0.7586
	LDA	0.6849	0.5818	0.9568	0.4790	0.7236
	KNN	0.8124	0.7513	0.8444	0.7882	0.7951
	SVM	0.8008	0.7104	0.9081	0.7195	0.7972
	LS-SVM	0.8332	0.7964	0.8236	0.8405	0.8098
	RF	0.8203	0.7532	0.8675	0.7846	0.8063
	GBDT	0.7773	0.6783	0.9196	0.6695	0.7807
	XGBoost	0.8119	0.7560	0.8322	0.7965	0.7923
DL	LeNet-5	0.8686	0.8786	0.8066	0.9156	0.8411
	ZFNet	0.8810	0.8510	0.8777	0.8836	0.8641
	AlexNet	0.8741	0.8810	0.8186	0.9162	0.8487
	GRU	0.8389	0.8317	0.7853	0.8796	0.8078
	LSTM	0.8505	0.8781	0.7586	0.9202	0.8140
	BiLSTM	0.8525	0.8478	0.8017	0.8910	0.8241
	CNN-LSTM	0.8960	0.8955	0.8590	0.9241	0.8769
	SE-MSCNN	0.9039	0.9063	0.8667	0.9321	0.8860
	Proposed	0.9106	0.9396	0.8470	0.9588	0.8909

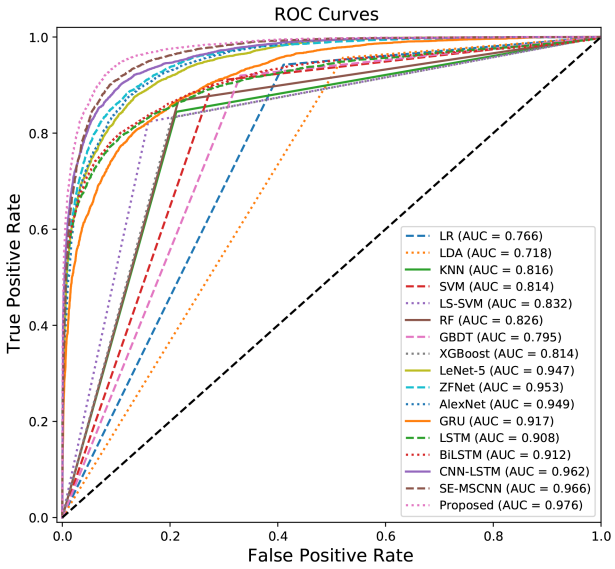

Fig. 6. The ROC curves with different SA detection methods at per-segment level.

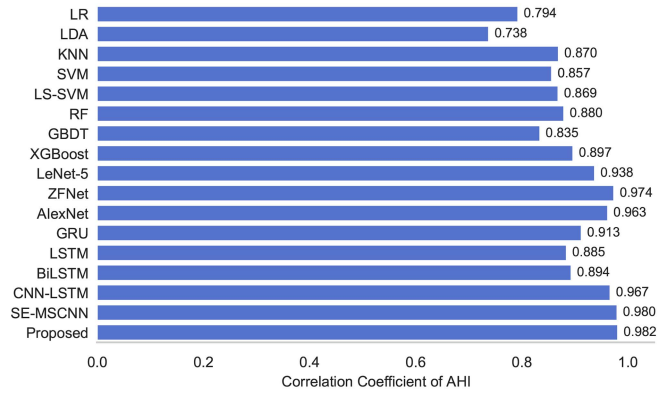

Fig. 7. The correlation coefficients of AHI with different SA detection methods at per-recording level.

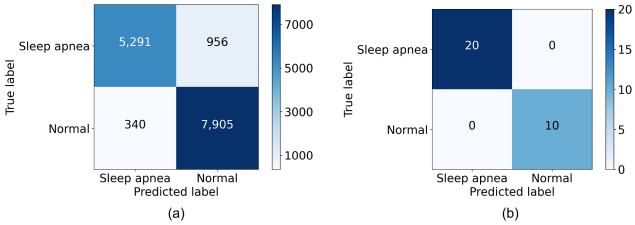

Fig. 8. Confusion matrix for proposed DAN-MTIF: (a) per-segment; (b) per-recording.

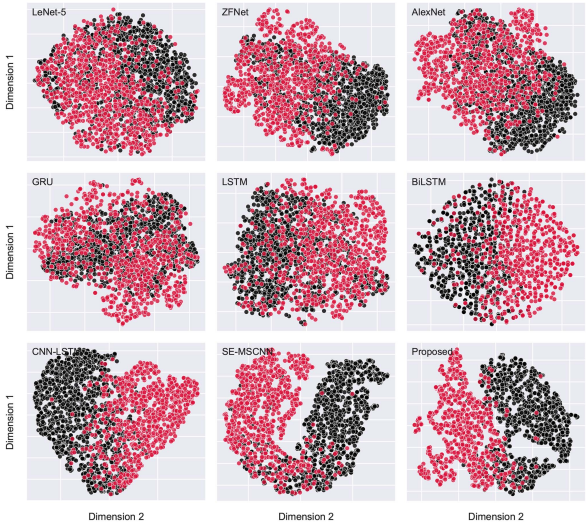

Fig. 9. T-SNE plots for learned features from DL models.

per-recording SA detection, which suggests that DAN-MTIF can effectively extract highly discriminative features from ECG signals.

- As illustrated in Fig. 7, RNN-based models (GRU, LSTM, BiLSTM) showed decreased AHI correlation coefficients when compared to CNN-based models (LeNet-5, ZFNet, AlexNet), while incorporating CNN with LSTM (CNN-LSTM) showed significant improvement in AHI correlation. This further demonstrates the superiority of CNN structure in feature extraction.

To assess the ability of DL models in extracting distinctive features for further SA detection, the t-distributed stochastic neighbor embedding (t-SNE) [65], which is a powerful nonlinear dimensionality reduction technique, is introduced to visualize embeddings learned from ECG segments by projecting high-dimensional embeddings into a low-dimensional (2D) space (Fig. 9). Each data point in t-SNE plots corresponds to an individual embedding entry, and different colors indicate distinct labels of the original ECG segment (A: red, N: black). The overlapping range of data points in two different colors indicates the clustering difficulty of features in the high-dimensional space.

- For instance, the t-SNE plots for LeNet-5 show considerable overlap among data points, implying the features it learned are not easily distinguishable.
- Interpolating an LSTM layer into LeNet-5 (CNN-LSTM) enables the model to capture temporal information from

TABLE III
PERFORMANCE COMPARISON OF DL MODELS AT PER-SEGMENT LEVEL WITH AN INCREASING PROPORTION OF TRAINING SET

	Proportion of Training Set				
Model	0.5	0.6	0.7	0.8	0.9
LeNet-5	0.8537 ± 0.0052	0.8591 ± 0.0058	0.8655 ± 0.0024	0.8615 ± 0.0018	0.8671 ± 0.0049
ZFNet	0.8607 ± 0.0054	0.8619 ± 0.0071	0.8738 ± 0.0059	0.8719 ± 0.0055	0.8742 ± 0.0088
AlexNet	0.8620 ± 0.0143	0.8670 ± 0.0073	0.8611 ± 0.0118	0.8570 ± 0.0163	0.8687 ± 0.0090
GRU	0.8183 ± 0.0063	0.8300 ± 0.0066	0.8411 ± 0.0113	0.8294 ± 0.0093	0.8345 ± 0.0047
LSTM	0.8278 ± 0.0039	0.8384 ± 0.0093	0.8406 ± 0.0054	0.8421 ± 0.0054	0.8437 ± 0.0112
BiLSTM	0.8398 ± 0.0109	0.8404 ± 0.0084	0.8458 ± 0.0122	0.8486 ± 0.0060	0.8476 ± 0.0109
CNN-LSTM	0.8907 ± 0.0030	0.8937 ± 0.0037	0.8949 ± 0.0026	0.8958 ± 0.0019	0.8967 ± 0.0025
SE-MSCNN	0.8942 ± 0.0054	0.8960 ± 0.0024	0.8991 ± 0.0032	0.8933 ± 0.0030	0.8927 ± 0.0047
Proposed	0.9117 ± 0.0133	0.9141 ± 0.0082	0.9154 ± 0.0057	0.9099 ± 0.0031	0.9152 ± 0.0047

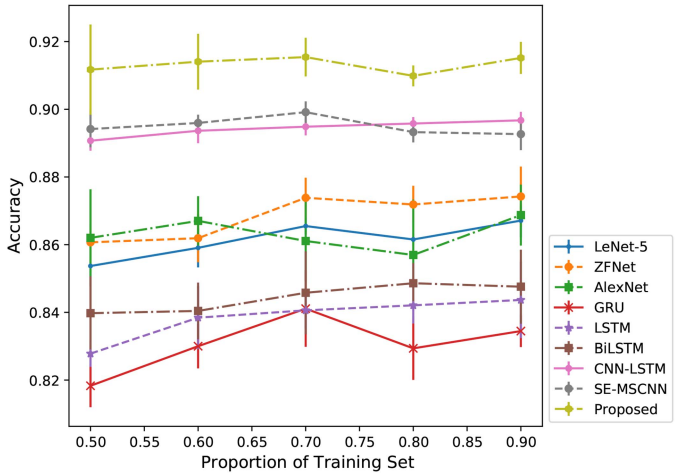

Fig. 10. The accuracy comparison of DL models at per-segment level with an increasing proportion of training set.

TABLE IV
DESIGN OF ABLATION EXPERIMENTS

Model	CNN-1	CNN-3	CNN-5	Attention	Dilated Rate
M1	✓				1
M2		✓			1
M3			✓		1
M4	✓	✓	✓		1
M5	✓				2
M6		✓			2
M7			✓		2
M8	✓	✓	✓		2
M9	✓			✓	1
M10		✓		✓	1
M11			✓	✓	1
M12	✓	✓	✓	✓	1
M13	✓			✓	2
M14		✓		✓	2
M15			✓	✓	2
M16	✓	✓	✓	✓	2

ECG. The extracted features display a trend towards forming two clusters, yet a distinct boundary to differentiate the two clusters is lacking.

- By contrast, SE-MSCNN and the proposed DAN-MTIF utilized the attention mechanism to assign proper weights for embeddings extracted from ECG segments at different timescales, resulting in a relatively clear border between the two clusters.

These findings align with the evaluation results in Table II, further verifying the superior performance of the proposed method.

C. Performance Stability Validation

Since the performance of DL models is clearly superior to ML models, in order to further evaluate the stability of the proposed DAN-MTIF and benchmark DL models (LeNet-5, ZFNet, AlexNet, GRU, LSTM, BiLSTM, CNN-LSTM, and SE-MSCNN), we compared the model performance under a varied training dataset size. The proportion of the training set was set to 0.5, 0.6, 0.7, 0.8, and 0.9, respectively, resulting in a reduction in the size of the validation set. Then, the DAN-MTIF and benchmark DL models were trained separately in different cases. To test the robustness of performance, we repeatedly trained each model 5 times, and all the classification results are summarized in Table III and Fig. 10.

It can be seen from these results that:

- The performance of the proposed DAN-MTIF consistently outperforms benchmark DL models, regardless of the size of the training set, which demonstrates the superior stability and robustness of DAN-MTIF.
- Most models show the best performance when the training proportion is 0.7 and show a small drop in accuracy as the training proportion continues to increase, which indicates the potential overfitting of the model.

D. Ablation Study

In order to verify the effectiveness of each component in DAN-MTIF, we performed a set of ablation experiments. The experimental settings are shown in Table IV. To validate the stability of the performance, we repeatedly conducted each experiment five times, and all SA detection results as well as the computational time of the model are summarized in Figs. 11-12 and Table V.

It can be concluded from these results that:

- By the comparison within M1 to M4, M5 to M8, M9 to M12, and M13 to M16, we can see that expanding the time scale of ECG segments from 1-min long to 3-min long significantly improved the SA detection accuracy of the model, and expanding the time scale to 5-min long

TABLE V
PERFORMANCE COMPARISON IN PER-SEGMENT SA DETECTION IN ABLATION STUDY

Model	Acc	Prec	Sens	Spec	F1	Time (s)
M1	0.8451 ± 0.0021	0.8267 ± 0.0056	0.8107 ± 0.0115	0.8711 ± 0.0067	0.8185 ± 0.0038	1.3268 ± 0.0109
M2	0.9031 ± 0.0025	0.9121 ± 0.0023	0.8580 ± 0.0072	0.9374 ± 0.0021	0.8842 ± 0.0034	1.3369 ± 0.0053
M3	0.9122 ± 0.0027	0.9284 ± 0.0048	0.8631 ± 0.0114	0.9495 ± 0.0043	0.8945 ± 0.0041	1.3648 ± 0.0031
M4	0.9087 ± 0.0023	0.9300 ± 0.0032	0.8524 ± 0.0064	0.9514 ± 0.0026	0.8895 ± 0.0031	3.5084 ± 0.0100
M5	0.8459 ± 0.0026	0.8302 ± 0.0088	0.8081 ± 0.0124	0.8746 ± 0.0095	0.8189 ± 0.0037	1.2945 ± 0.0033
M6	0.9032 ± 0.0019	0.9154 ± 0.0035	0.8545 ± 0.0049	0.9402 ± 0.0029	0.8839 ± 0.0025	1.3382 ± 0.0077
M7	0.9099 ± 0.0024	0.9304 ± 0.0046	0.8549 ± 0.0076	0.9515 ± 0.0037	0.8910 ± 0.0033	1.3634 ± 0.0045
M8	0.9096 ± 0.0021	0.9298 ± 0.0063	0.8548 ± 0.0020	0.9511 ± 0.0048	0.8907 ± 0.0021	3.5218 ± 0.0170
M9	0.8635 ± 0.0038	0.8414 ± 0.0100	0.8425 ± 0.0142	0.8795 ± 0.0104	0.8418 ± 0.0049	1.4685 ± 0.0064
M10	0.9057 ± 0.0145	0.8942 ± 0.0428	0.8903 ± 0.0259	0.9173 ± 0.0419	0.8910 ± 0.0125	1.4936 ± 0.0049
M11	0.9012 ± 0.0269	0.8816 ± 0.0582	0.8974 ± 0.0211	0.9041 ± 0.0573	0.8878 ± 0.0248	1.5379 ± 0.0143
M12	0.9088 ± 0.0097	0.9165 ± 0.0187	0.8680 ± 0.0188	0.9397 ± 0.0150	0.8913 ± 0.0114	3.7051 ± 0.0120
M13	0.8626 ± 0.0203	0.8433 ± 0.0147	0.8364 ± 0.0417	0.8825 ± 0.0095	0.8395 ± 0.0271	1.4215 ± 0.0067
M14	0.9124 ± 0.0038	0.9313 ± 0.0021	0.8604 ± 0.0099	0.9519 ± 0.0018	0.8944 ± 0.0051	1.4955 ± 0.0055
M15	0.9150 ± 0.0043	0.9334 ± 0.0102	0.8648 ± 0.0214	0.9530 ± 0.0089	0.8975 ± 0.0069	1.5299 ± 0.0056
M16	0.9137 ± 0.0052	0.9299 ± 0.0088	0.8652 ± 0.0215	0.9503 ± 0.0078	0.8961 ± 0.0079	3.6312 ± 0.0158

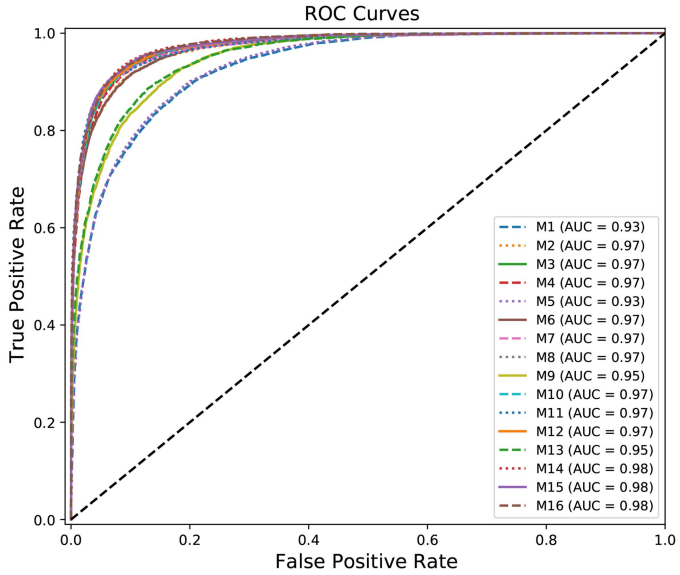

Fig. 11. The ROC curves with SA detection methods in ablation study.

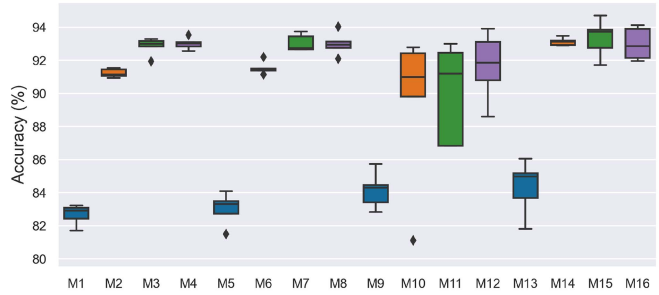

Fig. 12. The accuracy comparison of per-segment SA detection according to the ablation study.

resulting in a less obvious improvement in accuracy, which suggesting the temporal information from adjacent ECG segments in a proper time range indeed contributes to SA detection.

- By comparing M9 to M16 vs M1 to M8, we can observe the application of the multi-head attention module can improve model performance when the time scale of ECG segments is 1-min long, but can lead to a decrease in the model performance when the time scale of the ECG segments is of other temporal scales.
- By the comparison of M5 to M8 vs M1 to M4, and M13 to M16 vs M9 to M12, we can see the dilated rate of l has a certain influence on the model performance when the time scale of ECG segments is 1-min long. We can also notice that there is a variance in accuracy for M10, M11, and M12 where l is set to 1 and adjacent segments are employed as the model input. This configuration enables the model to learn more details on the local features from ECG, which may contribute to overfitting. However, when we change the dilated rate l to 2, a substantial increase in accuracy can be observed for M14, M15, M16. This is because a larger dilated rate allows the model to capture more global information, which can help prevent overfitting.
- The model with a dilated rate $l = 2$ and multi-head attention shows the highest accuracy under different ECG timescales. The combination use of three ECG segments (1-, 3-, 5-min long) segments suggests a similar or even slightly deteriorated performance when compared with the case that only 5-min long ECG segments are used. What's more, an obvious increase in the computational time can be observed as the 1D-CNN blocks in the model increased from one to three.

Overall, the results of ablation experiments demonstrate the advantages of multi-temporal information fusion and the multi-head attention mechanism in deep learning models for apnea event detection. However, further simplifying the proposed model remains a challenge, because a simpler model architecture will lead to shorter computational time and lower power consumption, which is a key factor for the application of the model in wearable devices. Although the proposed model exhibits performance improvement over benchmark models such as SE-MSCNN, it has no advantage over SE-MSCNN when considering structural complexity as the SE-MSCNN is a lightweight model. Given this, identifying a strategy to simplify the model architecture without adversely affecting its performance will be an essential focus of our future research.

E. Cross Validation

To evaluate the generalizability of the proposed DAN-MTIF, we further conducted SA detection with the MIT-BIH polysomnographic database [66]. This database contains 18 multiple physiologic recordings from 16 male subjects during sleep. All recordings were acquired in Boston's Beth Israel Hospital Sleep Laboratory for evaluation of chronic OSA syndrome with a sampling frequency of 250 Hz, and the duration of each recording is between 2 and 7 hours. The single-lead ECG signals were annotated every 30 seconds by clinical experts, and based on that 16 recordings have an AHI value above 5, while only 2 recordings have an AHI value of 5 or less. To align with the labeling rule in the Apnea-ECG database, we first removed segments annotated as "awake" and "hypopnea", then categorized the remaining segments into "normal" and "apnea". Next, we combined every two 30-second-long segments into a single segment, labeling it as normal (N) only if both sub-segments were identified as normal; otherwise, apnea (A). Finally, a total of 55 hours of ECG segments (N: 43%, A: 57%) derived from 18 recordings were used for SA detection with the proposed DAN-MTIF.

The evaluation results at per-segment level are 0.7125 accuracy, 0.7401 precision, 0.7599 sensitivity, 0.6503 specificity, and 0.7499 F_1. The evaluation results at per-recording level are 0.9375 accuracy, 0.9333 precision, 1.0 sensitivity, 0.5 specificity, and 0.9655 F_1. It is evident from the results that there is a gap between the performance at two different levels, indicating the model tends to identify normal segments as apnea segments. This could be attributed to a range of factors: (i) the variance of data acquisition equipment and environment in different laboratories; (ii) the limited diversity and quantity of ECG recordings in the training set; (iii) the design of the model itself makes it less generalizable and requires further improvement in the future.

V. Conclusion

In this study, a multiple temporal scale CNN framework with the multi-head attention mechanism (DAN-MTIF) is proposed for SA detection based on single-lead ECG signals. A unique module with three 1D-CNN blocks is designed to extract features from R-R intervals and R-peak amplitudes at different timescales (1-, 3-, 5-min long). The dilated convolutional kernel is employed to expand the receptive field while avoiding deepening the network structure. The features extracted from multiple channels are then concatenated together, and a multi-head attention module with a self-attention mechanism is introduced to adaptively generate proper weights for the concatenated features. Extensive experiments on the Apnea-ECG database demonstrate that (1) DAN-MTIF exhibits excellent performance with the highest accuracy for SA detection when compared to benchmark methods; (2) DAN-MTIF can effectively extract features with a higher degree of discrimination from ECG segments of multiple timescales compared to those of a single timescale. The results of ablation experiments suggest directions for further work, and future research will focus on (1) simplifying the model architecture without adversely affecting its performance; (2) improving the generalizability of the model to adapt to external datasets.

Data and Code Availability

The data analyzed in this study has open access from [46], [66]. The code is publicly accessible on GitHub (https://github.com/BAGL-lab/DAN-MTIF/).

Author Contributions

Jiao: data curation, formal analysis, methodology, software, visualization, writing - original draft, Song: conception, validation, methodology, writing - review & editing, Xian: conception, validation, methodology, writing - review & Editing, Yang: data curation, writing - review & editing, Liu: conceptualization, supervision, resources, methodology, writing - original draft, review & editing.

Conflict of Interest Statement

The authors declare no conflict of interest.

Acknowledgment

Research reported in this publication was supported by the NIBIB of the National Institutes of Health (NIH) under Award Number R21EB033455. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1] W. B. Mendelson, *The Science of Sleep: What It Is, How It Works, and Why It Matters*. Chicago, IL, USA: Univ. Chicago Press, 2020.
[2] Bruce M. Altevogt and Harvey R. Colten, *Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem*, Washington, DC, USA: Nat. Academies Press, 2006.
[3] L. M. Prisant, T. A. Dillard, and A. R. Blanchard, "Obstructive sleep apnea syndrome," *J. Clin. Hypertension*, vol. 8, no. 10, pp. 746–750, 2006.
[4] N. N. Finer, K. J. Barrington, B. J. Hayes, and A. Hugh, "Obstructive, mixed, and central apnea in the neonate: Physiologic correlates," *J. Pediatrics*, vol. 121, no. 6, pp. 943–950, 1992.
[5] M. W. Johns, "Daytime sleepiness, snoring, and obstructive sleep apnea: The epworth sleepiness scale," *Chest*, vol. 103, no. 1, pp. 30–36, 1993.
[6] J. M. Golbin, V. K. Somers, and S. M. Caples, "Obstructive sleep apnea, cardiovascular disease, and pulmonary hypertension," *Proc. Amer. Thoracic Soc.*, vol. 5, no. 2, pp. 200–206, 2008.
[7] P. M. Macey, M. A. Woo, R. Kumar, R. L. Cross, and R. M. Harper, "Relationship between obstructive sleep apnea severity and sleep, depression and anxiety symptoms in newly-diagnosed patients," *PLoS One*, vol. 5, no. 4, 2010, Art. no. e10211.
[8] A. Kales, A. B. Caldwell, R. J. Cadieux, A. Vela-Bueno, L. G. Ruch, and S. D. Mayes, "Severe obstructive sleep apnea-II: Associated psychopathology and psychosocial consequences," *J. Chronic Dis.*, vol. 38, no. 5, pp. 427–434, 1985.
[9] P. Kohli, J. S. Balachandran, and A. Malhotra, "Obstructive sleep apnea and the risk for cardiovascular disease," *Curr. Atherosclerosis Rep.*, vol. 13, pp. 138–146, 2011.
[10] C. Guilleminault, F. L. Eldridge, F. B. Simmon, and W. C. Dement, "Sleep apnea syndrome: Can it induce hemodynamic changes?," *Western J. Med.*, vol. 123, no. 1, pp 7–16, 1975.
[11] M. Naresh et al., "Sleep-disordered breathing and mortality: A prospective cohort study," *PLoS Med.*, vol. 6, no. 8, 2009, Art. no. e1000132.
[12] K. A. Franklin and E. Lindberg, "Obstructive sleep apnea is a common disorder in the population-a review on the epidemiology of sleep apnea," *J. Thoracic Dis.*, vol. 7, no. 8, 2015, Art. no. 1311.
[13] F. Mendonca, S. S. Mostafa, A. G. Ravelo-Garcia, F. Morgado-Dias, and T. Penzel, "A review of obstructive sleep apnea detection approaches," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 2, pp. 825–837, Mar. 2019.
[14] J. V. Rundo and R. Downey III, "Polysomnography," in *Handbook of Clinical Neurology*, vol. 160. Amsterdam, The Netherlands: Elsevier, 2019, pp. 381–392.
[15] M. Bruyneel et al., "Sleep efficiency during sleep studies: Results of a prospective study comparing home-based and in-hospital polysomnography," *J. Sleep Res.*, vol. 20, no. 1pt2, pp. 201–206, 2011.

[16] P. Jennum and R. L. Riha, "Epidemiology of sleep apnoea/hypopnoea syndrome and sleep-disordered breathing," *Eur. Respir. J.*, vol. 33, no. 4, pp. 907–914, 2009.

[17] K. Vishesh 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.*, vol. 13, no. 3, pp. 479–504, 2017.

[18] J. N. Miller, P. Schulz, B. Pozehl, D. Fiedler, A. Fial, and A. M. Berger, "Methodological strategies in using home sleep apnea testing in research and practice," *Sleep Breathing*, vol. 22, pp. 569–577, 2018.

[19] F. Mendonça, S. S. Mostafa, A. G. Ravelo-García, F. Morgado-Dias, and T. Penzel, "Devices for home detection of obstructive sleep apnea: A review," *Sleep Med. Rev.*, vol. 41, pp. 149–160, 2018.

[20] S. Kwon, H. Kim, and W.-H. Yeo, "Recent advances in wearable sensors and portable electronics for sleep monitoring," *Iscience*, vol. 24, no. 5, 2021, Art. no. 102461.

[21] J. M. Kelly, R. E. Strecker, and M. T. Bianchi, "Recent developments in home sleep-monitoring devices," *Int. Scholarly Res. Notices*, vol. 2012, 2012, Art. no. 768794.

[22] H. ElMoaqet, M. Eid, M. Glos, M. Ryalat, and T. Penzel, "Deep recurrent neural networks for automatic detection of sleep apnea from single channel respiration signals," *Sensors*, vol. 20, no. 18, 2020, Art. no. 5037.

[23] K. Kesper, S. Canisius, T. Penzel, T. Ploch, and W. Cassel, "ECG signal analysis for the assessment of sleep-disordered breathing and sleep pattern," *Med. Biol. Eng. Comput.*, vol. 50, pp. 135–144, 2012.

[24] P. K. Stein and Y. Pu, "Heart rate variability, sleep and sleep disorders," *Sleep Med. Rev.*, vol. 16, no. 1, pp. 47–66, 2012.

[25] D.-H. Park et al., "Correlation between the severity of obstructive sleep apnea and heart rate variability indices," *J. Korean Med. Sci.*, vol. 23, no. 2, pp. 226–231, 2008.

[26] M. Bahrami and M. Forouzanfar, "Sleep apnea detection from single-lead ECG: A comprehensive analysis of machine learning and deep learning algorithms," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 4003011.

[27] D. Alvarez, R. Hornero, D. Abásolo, F. D. Campo, and C. Zamarrón, "Nonlinear characteristics of blood oxygen saturation from nocturnal oximetry for obstructive sleep apnoea detection," *Physiol. Meas.*, vol. 27, no. 4, 2006, Art. no. 399.

[28] A. Bhattacharjee, S. Saha, S. A. Fattah, W.-P. Zhu, and M. O. Ahmad, "Sleep apnea detection based on Rician modeling of feature variation in multiband EEG signal," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 3, pp. 1066–1074, May 2019.

[29] P. Janbakhshi and M. B. Shamsollahi, "Sleep apnea detection from single-lead ecg using features based on ecg-derived respiration (EDR) signals," *Irbm*, vol. 39, no. 3, pp. 206–218, 2018.

[30] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscipl. Rev.: Comput. Statist.*, vol. 2, no. 4, pp. 433–459, 2010.

[31] J. Chen, M. Shen, W. Ma, and W. Zheng, "A spatio-temporal learning-based model for sleep apnea detection using single-lead ECG signals," *Front. Neurosci.*, vol. 16, 2022, Art. no. 972581.

[32] C. Varon, A. Caicedo, D. Testelmans, B. Buyse, and S. V. Huffel, "A novel algorithm for the automatic detection of sleep apnea from single-lead ECG," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 9, pp. 2269–2278, Sep. 2015.

[33] A. Rizal, F. D. A A. Siregar, and H. T. Fauzi, "Obstructive sleep apnea (OSA) classification based on heart rate variability (HRV) on electrocardiogram (ECG) signal using support vector machine (SVM)," *Traitement du Signal*, vol. 39, no. 2, pp. 469–474, 2022.

[34] N. Salari et al., "Detection of sleep apnea using machine learning algorithms based on ECG signals: A comprehensive systematic review," *Expert Syst. Appl.*, vol. 187, 2022, Art. no. 115950.

[35] S. S. Mostafa, F. Mendonça, A. G. Ravelo-García, and F. Morgado-Dias, "A systematic review of detecting sleep apnea using deep learning," *Sensors*, vol. 19, no. 22, 2019, Art. no. 4934.

[36] B. Samadi, S. Samadi, M. Samadi, S. Samadi, M. Samadi, and M. Mohammadi, "Systematic review of detecting sleep apnea using artificial intelligence: An insight to convolutional neural network method," *Arch. Neurosci.*, vol. 11, no. 1, 2024, Art. no. e144058.

[37] T. Wang, C. Lu, G. Shen, and F. Hong, "Sleep apnea detection from a single-lead ecg signal with automatic feature-extraction through a modified LeNet-5 convolutional neural network," *PeerJ*, vol. 7, 2019, Art. no. e7731.

[38] S. M. I. Niroshana, X. Zhu, K. Nakamura, and W. Chen, "A fused-image-based approach to detect obstructive sleep apnea using a single-lead ECG and a 2D convolutional neural network," *PLoS One*, vol. 16, no. 4, 2021, Art. no. e0250618.

[39] O. Faust, R. Barika, A. Shenfield, E. J. Ciaccio, and U. R. Acharya, "Accurate detection of sleep apnea with long short-term memory network based on RR interval signals," *Knowl.-Based Syst.*, vol. 212, 2021, Art. no. 106591.

[40] X. Liang, X. Qiao, and Y. Li, "Obstructive sleep apnea detection using combination of CNN and LSTM techniques," in *Proc. IEEE 8th Joint Int. Inf. Technol. Artif. Intell. Conf.*, 2019, pp. 1733–1736.

[41] A. Zarei, H. Beheshti, and B. M. Asl, "Detection of sleep apnea using deep neural networks and single-lead ECG signals," *Biomed. Signal Process. Control*, vol. 71, 2022, Art. no. 103125.

[42] X. Chen, Y. Chen, W. Ma, X. Fan, and Y. Li, "Toward sleep apnea detection with lightweight multi-scaled fusion network," *Knowl.-Based Syst.*, vol. 247, 2022, Art. no. 108783.

[43] U. Erdenebayar, Y. J. Kim, J. U. Park, E Y. Joo, and K.-J. Lee, "Deep learning approaches for automatic detection of sleep apnea events from an electrocardiogram," *Comput. Methods Programs Biomed.*, vol. 180, 2019, Art. no. 105001.

[44] P. Hamilton, "Open source ECG analysis," in *Proc. IEEE Comput. Cardiol.*, 2002, pp. 101–104.

[45] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.

[46] T. Penzel, G. B. Moody, R. G. Mark, A. L. Goldberger, and J. H. Peter, "The apnea-ECG database," in *Proc. IEEE Comput. Cardiol.*, 2000, pp. 255–258.

[47] V. C. C. Sequeira, P. M. Bandeira, and J. C. M. Azevedo, "Heart rate variability in adults with obstructive sleep apnea: A systematic review," *Sleep Sci.*, vol. 12, no. 3, pp. 214–221, 2019.

[48] R. Champseix, L. Ribiere, and C. L. Couedic, "A Python package for heart rate variability analysis and signal preprocessing," *J. Open Res. Softw.*, vol. 9, no. 1, 2021, Art. no. 28.

[49] T. Pham, Z. J. Lau, A. S. H. Chen, and D. Makowski, "Heart rate variability in psychology: A review of HRV indices and an analysis tutorial," *Sensors*, vol. 21, no. 12, 2021, Art. no. 3998.

[50] G. Martin Frasch, "Comprehensive HRV estimation pipeline in Python using Neurokit2: Application to sleep physiology," *MethodsX*, vol. 9, 2022, Art. no. 101782.

[51] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[52] T. Chen et al., "Xgboost: Extreme gradient boosting," *R Package Version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.

[53] R. E. Wright, "Logistic regression," in *Reading and understanding multivariate statistics*. Amer. Psychol. Assoc., 1995, pp. 217–244.

[54] S. Balakrishnama and A. Ganapathiraju, "Linear discriminant analysis – A brief tutoria," *Inst. Signal Inf. Process.*, vol. 18, no. 1998, pp. 1–8, 1998.

[55] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, 2009, Art. no. 1883.

[56] W. S. Noble, "What is a support vector machine?," *Nature Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, 2006.

[57] J. A, K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, pp. 293–300, 1999.

[58] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, 2001.

[59] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[60] D. M. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. 3rd Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[61] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[62] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (GRU) neural networks," in *Proc. IEEE 60th Int. Midwest Symp. Circuits Syst.*, 2017, pp. 1597–1600.

[63] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[64] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[65] M. C. Cieslak, A. M. Castelfranco, V. Roncalli, P. H. Lenz, and D. K. Hartline, "T-distributed stochastic neighbor embedding (t-sne): A tool for eco-physiological transcriptomic analysis," *Mar. Genomic.*, vol. 51, 2020, Art. no. 100723.

[66] Y. Ichimaru and G. B. Moody, "Development of the polysomnographic database on CD-ROM," *Psychiatry Clin. Neurosci.*, vol. 53, no. 2, pp. 175–177, 1999.