DREAM: Domain Invariant and Contrastive Representation for Sleep Dynamics

Seungyeon Lee[1,2,*], Thai-Hoang Pham[1,2,*], Ping Zhang[1,2]
[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Department of Biomedical Informatics, The Ohio State University, USA
{lee.10029,pham.375,zhang.10631}@osu.edu

***Abstract*—Sleep staging is a key challenge in diagnosing and treating sleep-related diseases due to its labor-intensive, time-consuming, costly, and error-prone. With the availability of large-scale sleep signal data, many deep learning methods are proposed for automatic sleep staging. However, these existing methods face several challenges including the heterogeneity of patients' underlying health conditions and the difficulty modeling complex interactions between sleep stages. In this paper, we propose a neural network architecture named DREAM to tackle these issues for automatic sleep staging. DREAM consists of (i) a feature representation network that generates robust representations for sleep signals via the variational auto-encoder framework and contrastive learning and (ii) a sleep stage classification network that explicitly models the interactions between sleep stages in the sequential context at both feature representation and label classification levels via Transformer and conditional random field architectures. Our experimental results indicate that DREAM significantly outperforms existing methods for automatic sleep staging on three sleep signal datasets.**

***Index Terms*—deep learning, domain invariant, contrastive learning, sleep staging, sleep dynamics, EEG analysis**

I. Introduction

Sleep is a vital process for humans that significantly affects brain function. High-quality sleep benefits humans' physical and mental health, and sleep disorders such as insomnia and sleep apnea can affect functions of the immune system, memory, and metabolism [1]. Sleep stage identification, used in the diagnosis and treatment of sleep disorders, has therefore become a critical task for reducing this burden and enhancing life quality. Sleep specialists often rely on polysomnography (PSG) studies, which include physiological signals such as electroencephalography (EEG), electrooculography (EOG), electrocardiogram (ECG), or electromyography (EMG) to identify the sleep stages of sleep segments (i.e., 30-second intervals) [2]. However, the manual classification of these sleep segments is highly challenging: it is labor-intensive, time-consuming, costly, and prone to human errors–especially in the current big data era in which at-home and mobile wearable devices are popular.

Recently, automatic sleep staging–in which machine intelligence assists sleep specialists in labeling sleep segments–has gained significant attraction. In particular, many deep learning architectures have been proposed to automatically extract meaningful features from raw signals to identify the stages of sleep [3]–[8]. Although promising results have been reported in their experiments, existing methods have not sufficiently addressed some critical challenges, thereby hindering the prediction performances. These challenges are as follows:

C1. Physiological signals vary significantly across subjects due to the differences in subjects' underlying health conditions (i.e., each subject can be considered as one domain whose physiological signal values are generated by a distinct distribution). Capturing sleep-relevant and subject-invariant features is key to generalizing prediction performances of automatic sleep staging systems to new subjects.

C2. Sleep stages are known to have a strong transition [9]. The incapability of existing methods to capture the dynamics of sleep stages at both the feature representation and label classification levels can lead to inaccurate results.

To tackle these challenges, we propose a new neural network-based framework named Domain invaRiant and contrastivE representAtion for sleep dynaMics (**DREAM**). DREAM consists of two main components: **(i)** the feature representation network leveraging variational auto-encoder (VAE) and contrastive learning (which addresses **C1**) and **(ii)** the sleep stage classifier constructed from Transformer and conditional random field (CRF) [10] architectures (which addresses **C2**). Specifically, the physiological signal (i.e., EEG) of each sleep segment from the dataset is entered into the VAE-based model to learn the disentanglement representation in which the subject- and sleep-specific information can be separately captured. The contrastive objective functions are also incorporated into the VAE-based model to enhance the robustness of the representations. The trained encoder from this step is used as the feature representation network for training the sleep staging classification network. Given the sleep signal sequence, the feature representation network is used to extract a sleep-relevant and subject-invariant representation sequence. Then, the Transformer model with a multi-head attention mechanism is utilized to capture the dependencies between these representations. Rather than separately finding a label for each sleep segment, the contextualized representation learned from the Transformer is entered into a CRF model to find the best overall label sequence for the input sleep signal sequence. In summary, our contributions are as follows:

- We design a neural network architecture (DREAM) that learns robust representations for sleep signals and models the dynamics of sleep stages for accurate prediction.

* The first two authors contributed equally to this work.

978-1-6654-5099-7/22/$31.00 ©2022 IEEE
DOI 10.1109/ICDM54844.2022.00126

- We develop a feature representation network based on VAE architecture and contrastive learning to effectively extract sleep-relevant and subject-invariant information across diverse subjects from the dataset, resulting in robust representations that generalize well to new subjects.
- We introduce a sleep stage classification network composed of Transformer and CRF models to explicitly captures the dependencies between sleep stages in the sleep signal sequence at both the feature representation and label classification levels.
- We demonstrate the effectiveness of DREAM compared to a wide range of existing approaches to automatic sleep staging.

The remainder of the paper is organized as follows. Section II summarizes related works. Section III describes the technical details of the proposed model (DREAM). Section IV presents experimental results and discussions. Finally, Section V concludes the paper.

II. Related Works

Automatic sleep staging. The availability of large-scale public datasets about sleep stages creates chances for applying machine learning methods to automatically identify sleep stages from physiological signals. Deep learning paradigms are mostly based on convolutional neural networks (CNNs) [3]–[5] and recurrent neural networks (RNNs) [6]–[8] for automatic sleep staging. The combinations of CNNs and RNNs in which CNNs are used to extract features from sleep segments and RNNs are used to model temporal relation between them are also popular architectures for this task [11]–[13]. The advantage of these models is that they can automatically extract meaningful features. However, these existing models cannot (i) explicitly model subject-invariant and sleep-relevant features to help them can be generalized well in the new subjects and (ii) capture the transition between sleep stages at both feature representation and label classification levels.

Robust machine learning in data shift. Most machine learning models rely on the strong assumption about data distribution in which data points are independently and identically distributed for both training and testing environments. However, this over-simplified assumption is often violated in practice where machine learning models need to deal with the out-of-distribution problem (e.g., data distribution of train and test subjects are different due to their different underlying health conditions) resulting in significant performance drops in the testing environment. Many methods have been proposed to guarantee the robustness of machine learning models under the change of environment in different scenarios including transfer learning, domain adaptation, and domain generalization. In our study, we design the model for the domain generalization scenario so that it can successfully generalize its sleep stage predictions to the new testing subjects based on information from subjects in the training set.

Contrastive learning. Contrastive learning constructs a representation space in which similar sample pairs stay close to each other while dissimilar ones are far apart to learn robust representation. This method can be applied for both supervised [14] and unsupervised [15] settings. In our study, we leverage the contrastive objective function to learn more robust representations for sleep signal segments from the labeled dataset.

III. Methodology

In this section, we use upper-case and bold letters (e.g., $\mathbf{X}$) for matrices, lower-case and bold letters (e.g., $\mathbf{x}$) for vectors, and lower-case letters (e.g., x) for scalar.

A. Definitions for automatic sleep staging

- **Sleep signal sequence:** The sleep signal (EEG) sequence is divided into 30-second segments in sleep staging identification problem. i^{th} segment of the k^{th} sequence in the dataset is denoted by $\mathbf{x}_{k,i} \in \mathbb{R}^N$ where $N = 30 \times F$ is the number of signal values in the segment $\mathbf{x}_i$ and F is the number of samplings per second. Then the sleep signal sequence can be represented as $\mathbf{X}_k = [\mathbf{x}_{k,1}, \mathbf{x}_{k,2}, \cdots, \mathbf{x}_{k,T}]$ where T is the number of segments in the sequence.
- **Sleep stage:** According to the American Academy of Sleep Medicine (AASM) standard, the sleep segment can be divided into three main groups: Wakefulness (W), rapid eye movement (REM), non-rapid eye movement (NREM). NREM can be further divided into N1, N2, N3, and N4. Due to the infrequency of N4, we merge it into N3. The final sleep stage set $\mathcal{Y}$ used in our study includes W, REM, N1, N2, and N3.
- **Automatic sleep staging:** Given the sleep signal sequence $\mathbf{X}_k = [\mathbf{x}_{k,1}, \mathbf{x}_{k,2}, \cdots, \mathbf{x}_{k,T}]$, the goal of automatic sleep staging is finding the function $f : \mathbb{R}^{T \times N} \rightarrow \mathcal{Y}^T$ that maps from $\mathbf{X}_k$ to the sequence of sleep stages $\mathbf{y}_k = [y_{k,1}, y_{k,2}, \cdots, y_{k,T}]$

B. Proposed model

The proposed framework DREAM consists of two main networks: (i) the feature representation network that transforms the input sleep segment into sleep-relevant and subject-invariant feature representations and (ii) the classification network that captures the dependencies between sleep segments in the sequential context to find the best corresponding sleep stage sequence. Figure 1 visualizes the overall architecture of DREAM and the details of this model are described as follows.

*a) **Feature representation network**:* We leverage VAE framework to learn sleep-relevant and subject-invariant representations for each sleep segment. This approach is based on an assumption about the sleep segment generation process in which $\mathbf{x}$ is generated from two latent vectors $\mathbf{z}_d$ and $\mathbf{z}_y$ that only capture information about the subject and sleep stage, respectively. Then under this assumption, $\mathbf{z}_y$ is sleep-relevant and subject-invariant representations. The training process for the feature representation network is as follows. For dataset $D \equiv \{(\mathbf{x}_k, y_k, d_k)\}_{k=1}^{|D|}$ where d_k is subject id, we apply two data augmentation methods including (i) randomly split the sleep segment into chunks then randomly permute these

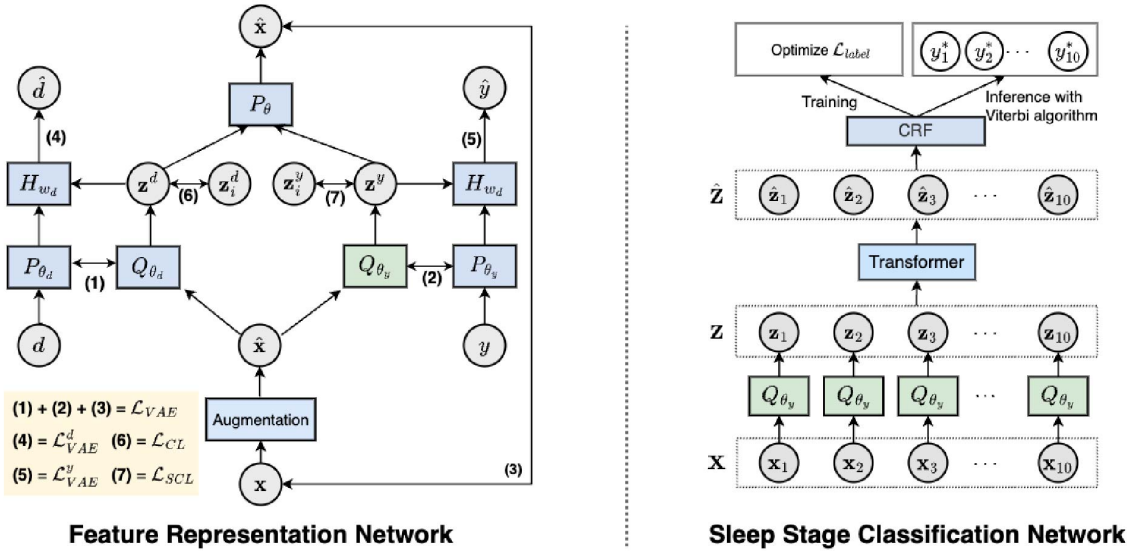

Fig. 1: Overall architecture of **DREAM**. This model consists of two main components: **feature representation network** and **sleep stage classification network**. We apply a 2-stage training process for DREAM. First, the feature representation network is trained on the labeled dataset, then the trained feature representation network is used in the training of the classification network.

chunks, (ii) randomly crop the sleep segment and resize it to the original size with linear interpolation to create two augmented examples $\widehat{\mathbf{x}}_{2k}$ and $\widehat{\mathbf{x}}_{2k+1}$ for each $\mathbf{x}_k$. This process results in the augmented dataset $\widehat{D} \equiv \left\{\left(\widehat{\mathbf{x}}_k, \widehat{y}_k, \widehat{d}_k\right)\right\}_{k=1}^{2|D|}$ where $\widehat{y}_{2k} = \widehat{y}_{2k+1} = y_k$ and $\widehat{d}_{2k} = \widehat{d}_{2k+1} = d_k$. Then, we maximize the ELBO with two latent variables $\mathbf{z}^d_k$ and $\mathbf{z}^y_k$ as a surrogate function for augmented data log-likelihood as follows.

$$\begin{aligned}\mathcal{L}_{VAE} = &\mathbb{E}_{Q_{\phi_d}\left(\mathbf{z}^d_k|\widehat{\mathbf{x}}_k\right), Q_{\phi_y}\left(\mathbf{z}^y_k|\widehat{\mathbf{x}}_k\right)}\left[P_\theta\left(\widehat{\mathbf{x}}_k|\mathbf{z}^d_k, \mathbf{z}^y_k\right)\right] \\ &- \beta\,\mathrm{KL}\left(Q_{\phi_d}\left(\mathbf{z}^d_k|\widehat{\mathbf{x}}_k\right) \parallel P_{\theta_d}\left(\mathbf{z}^d_k|\widehat{d}_k\right)\right) \\ &- \beta\,\mathrm{KL}\left(Q_{\phi_y}\left(\mathbf{z}^y_k|\widehat{\mathbf{x}}_k\right) \parallel P_{\theta_y}\left(\mathbf{z}^y_k|\widehat{y}_k\right)\right)\end{aligned}$$

Q_{ϕ_d} and Q_{ϕ_y} are the two encoder networks constructed from ResNet-50 architecture [16] that map input $\widehat{\mathbf{x}}_k$ to latent representations $\mathbf{z}^d_k$ and $\mathbf{z}^y_k$, P_θ is the decoder (i.e., 2-layer feed-forward network followed by 3 transposed convolutional layers) that reconstructs input $\widehat{\mathbf{x}}_k$ from its latent representations $\mathbf{z}^d_k$ and $\mathbf{z}^y_k$, P_{θ_d} and P_{θ_y} are the two prior networks (i.e., 3-layer feed-forward networks) for $\mathbf{z}^d_k$ and $\mathbf{z}^y_k$, KL is the KL-divergence between two distributions, and β is a weight that controls KL-divergence constraints. Motivated by β-VAE model [17], a larger value of β forces each dimension of $\mathbf{z}^d_k$ and $\mathbf{z}^y_k$ captures one of the conditionally independent factors in $\mathbf{x}$. To force the disentangled representations $\mathbf{z}^d_k$ (resp. $\mathbf{z}^y_k$) only capture information about $\widehat{d}_k$ (resp. $\widehat{y}_k$), we utilize two classifiers H_{ω_d} and H_{ω_y} (i.e., 1-layer feed-forward networks) that predict $\widehat{d}_k$ from $\mathbf{z}^d_k$ and $\widehat{y}_k$ from $\mathbf{z}^y_k$ respectively, and

Algorithm 1: Two-stage training process for DREAM

Input: Dataset $\mathcal{D}$
Output: feature representation network Q_{θ_y}, sleep stage classification network f

Procedure `Feature_Learning`
  for *epoch* $= 1$ *to* E **do**
    for *batch* $\boldsymbol{b} = \{(\boldsymbol{x}_k, y_k, d_k)\}_{k=1}^{|\boldsymbol{b}|}$ *in* $\mathcal{D}$ **do**
      Generate augmented batch $\widehat{\mathbf{b}}$ by data augmentation;
      Optimize $\mathcal{L}_{feature}$ in Eq. (1);
Procedure `Classification_Learning`
  for *epoch* $= 1$ *to* E **do**
    for *batch* $\boldsymbol{b} = \{(\boldsymbol{X}_k, \boldsymbol{y}_k)\}_{k=1}^{|\boldsymbol{b}|}$ *in* $\mathcal{D}$ **do**
      Optimize $\mathcal{L}_{label}$ in Eq. (2);

optimize the classification losses as follows.

$$\begin{aligned}\mathcal{L}^d_{VAE} &= -\mathbb{E}_{Q_{\phi_d}\left(\mathbf{z}^d_k|\widehat{\mathbf{x}}_k\right)}\left[\log H_{\omega_d}\left(\widehat{d}_k|\mathbf{z}^d_k\right)\right] \\ \mathcal{L}^y_{VAE} &= -\mathbb{E}_{Q_{\phi_y}\left(\mathbf{z}^y_k|\widehat{\mathbf{x}}_k\right)}\left[\log H_{\omega_y}\left(\widehat{y}_k|\mathbf{z}^y_k\right)\right]\end{aligned}$$

To further force the robustness of the latent representations, we apply self-supervised contrastive learning [15] for $\mathbf{z}^d_k$ and supervised contrastive learning [14] for $\mathbf{z}^y_k$. The main idea is to construct an embedding space in which similar representations stay close to each other while dissimilar ones are far apart. For subject-relevant representation $\mathbf{z}^d_k$, its similar representation is the one generated from the same sleep segment while for sleep-relevant representation $\mathbf{z}^y_k$, its similar ones are representations that have the same corresponding sleep stages. In particular, given the set $\{\mathbf{z}^d_k\}_{k=1}^{2|\mathcal{D}|}$, the self-supervised

contrastive objective function is applied for a similar pair $\{\mathbf{z}_k^d, \mathbf{z}_i^d\}$ as follows

$$\mathcal{L}_{CL}^d = -\log \frac{\exp\left(\text{sim}\left(\psi_d\left(\mathbf{z}_k^d\right), \psi_d\left(\mathbf{z}_i^d\right)\right)/\rho\right)}{\sum_{j\in A(k)} \exp\left(\text{sim}\left(\psi_d\left(\mathbf{z}_k^d\right), \psi_d\left(\mathbf{z}_j^d\right)\right)/\rho\right)}$$

where ψ_d is the function that maps $\mathbf{z}_k^d$ to the embedding space, $A(k) \equiv \{1, 2, \cdots, 2|\mathcal{D}|\} \setminus \{k\}$, $\text{sim}(\mathbf{u}, \mathbf{v})$ is the cosine similarity of two vector $\mathbf{u}, \mathbf{v}$ in the embedding space, and ρ denotes the temperature parameters. For $\mathbf{z}_k^y$, we optimize the supervised contrastive objective function as follows

$$\mathcal{L}_{SCL}^y = \frac{1}{|P(k)|} \times \sum_{p\in P(k)} -\log \frac{\exp\left(\text{sim}\left(\psi_y\left(\mathbf{z}_k^y\right), \psi_y\left(\mathbf{z}_p^y\right)\right)/\rho\right)}{\sum_{j\in A(k)} \exp\left(\text{sim}\left(\psi_y\left(\mathbf{z}_k^y\right), \psi_y\left(\mathbf{z}_j^y\right)\right)/\rho\right)}$$

where $P(k) \equiv \{p \in A(k) : y_p = y_k\}$ is the set of indices of all positives in the data batch distinct from k and $|P(k)|$ is its cardinality. In sum, we optimize

$$\mathcal{L}_{feature} = -\mathcal{L}_{VAE} + \alpha_d \mathcal{L}_{VAE}^d + \alpha_y \mathcal{L}_{VAE}^y + \gamma_d \mathcal{L}_{CL}^d + \gamma_y \mathcal{L}_{SCL}^y \quad (1)$$

where $\alpha_d, \alpha_y, \gamma_d, \gamma_y$ are hyper-parameters that control the relative importance of $\mathcal{L}_{VAE}^d$, $\mathcal{L}_{VAE}^y$, $\mathcal{L}_{CL}^d$, $\mathcal{L}_{SCL}^y$, respectively, compared to the $\mathcal{L}_{VAE}$.

After training the VAE architecture with $\mathcal{L}_{feature}$, the trained network Q_{ϕ_y} is used as the feature representation network which generates sleep-relevant and subject-invariant inputs for the classification network.

*b) **Classification network**:* Instead of considering one sleep segment as input as in the feature representation network, the classification network generates predictions in the sequential context. Specifically, given the dataset $D = \{(\mathbf{X}_k, \mathbf{y}_k)\}_{k=1}^{|D|}$ where $\mathbf{X}_k = [\mathbf{x}_{k,1}, \mathbf{x}_{k,2}, \cdots, \mathbf{x}_{k,T}]$ is the sleep signal sequence, $\mathbf{y}_k = [y_{k,1}, y_{k,2}, \cdots, y_{k,T}]$ is the corresponding sleep stage sequence, and T is the number of elements in $\mathbf{X}_k$ and $\mathbf{y}_k$, first we transform $\mathbf{X}_k$ to the sequence of features as follows.

$$\begin{aligned}\mathbf{Z}_k &= [\mathbf{z}_{k,1}, \mathbf{z}_{k,2}, \cdots, \mathbf{z}_{k,T}] \\ &= \left[Q_{\phi_y}(\mathbf{x}_{k,1}), Q_{\phi_y}(\mathbf{x}_{k,2}), \cdots, Q_{\phi_y}(\mathbf{x}_{k,T})\right]\end{aligned}$$

Then the 4-layer encoder network of Transformer model with 8 attention heads for each layer is used to learn the contextualized representations for $\mathbf{Z}_k$ as follows.

$$\widehat{\mathbf{Z}}_k = [\widehat{\mathbf{z}}_{k,1}, \widehat{\mathbf{z}}_{k,2}, \cdots, \widehat{\mathbf{z}}_{k,T}] = \text{Transformer\_Encoder}\left(\mathbf{Z}_k\right)$$

Explicitly, each attention head in Transformer_Encoder generates the contextualized representation sequence $\mathbf{Z}$ for its input sequence $\mathbf{X}$ as follows

$$\mathbf{Z} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{m}}\right)\mathbf{V}$$

where the three matrices $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{T\times m}$ is the query, key, value matrices calculated from $\mathbf{X}$ and m is the dimension of the embedding space. Then the output of each Transformer_Encoder layer is concatenated from the outputs of attention heads in that layer. To generate accurate prediction from the contextualized representation $\widehat{\mathbf{Z}}_k$, we leverage the CRF model to find the best overall sleep stage sequence instead of predicting the sleep stage for each sleep segment in the sequence independently. In particular, CRF considers the correlations between sleep stages in neighborhoods and jointly decodes the best chain of stages for $\mathbf{X}_k$. It will be beneficial for automatic sleep staging in the sequential context. For example, it is not likely that the N3 stage follows W or REM stages (this constraint is not trivial to capture and can be violated if the model predicts the sleep stage independently). Formally, the CRF model is used to calculate the probability of the output sleep stage sequence $\mathbf{y}_k$ as follows.

$$P_\varphi\left(\mathbf{y}_k|\widehat{\mathbf{Z}}_k\right) = \frac{\prod_{i=1}^T \mathcal{S}_\varphi\left(y_{k,i-1}, y_{k,i}, \widehat{\mathbf{Z}}_k\right)}{\sum_{\mathbf{y}'\in\mathcal{Y}^T} \prod_{i=1}^T \mathcal{S}_\varphi\left(y'_{i-1}, y'_i, \widehat{\mathbf{Z}}_k\right)}$$

where $\mathcal{S}_\varphi\left(y_{k,i-1}, y_{k,i}, \widehat{\mathbf{Z}}_k\right)$ is the potential function constructed from 1-layer feed-forward network that captures the correlation between $y_{k,i-1}$ and $y_{k,i}$ given $\widehat{\mathbf{Z}}_k$. In the training, we minimize the negative conditional log-likelihood which is given as

$$\mathcal{L}_{label} = -\log P_\varphi\left(\mathbf{y}_k|\widehat{\mathbf{Z}}_k\right) \quad (2)$$

while in the inference, the Viterbi algorithm [18] is used to find the best label sequence with the highest conditional probability

$$\mathbf{y}_k^* = \arg\max_{\mathbf{y}'\in\mathcal{Y}^T} P_\varphi\left(\mathbf{y}'|\widehat{\mathbf{Z}}_k\right)$$

*c) **Optimization**:* To train the proposed model DREAM, we use a 2-stage optimization process as follows. First the feature representation network Q_{ϕ_y} is trained by minimizing $\mathcal{L}_{feature}$. The weights of the trained network Q_{ϕ_y} are fixed and then used in the training of the classification network f in which $\mathcal{L}_{label}$ is minimized. The details of this 2-stage optimization process are shown in Algorithm 1.

IV. Experiments

In this section, we evaluate the performances of DREAM on three sleep staging datasets and compare its results with existing models to demonstrate the effectiveness of our method for automatic sleep staging.

A. Datasets

We conduct our experiments on three sleep staging datasets including **SleepEDF-20**, **SleepEDF-78** (i.e., expansion of **SleepEDF-20**) [19], [20] and **SHHS** [21], [22]. **SleepEDF** datasets include PSG studies for healthy subjects aged from 25 to 101 to investigate age effects on sleep. According to previous works [12], [23], [24], we extract data from single-channel EEG (Fpz-Cz), which is easy to collect and contains the most information about the sleep stage, with a sampling rate of 100 Hz. **SHHS** is a multi-center cohort study used to investigate the effect of sleep-disordered breathing on cardiovascular diseases. We only selected the subjects who are considered to have regular sleep by the studies [23], [25], so 329 of 6,441 subjects were finally selected to construct the

TABLE I: Comparison of prediction performance measured by accuracy, F_1 MF_1, and k scores on three sleep staging datasets. We report the average scores and their corresponding standard deviation between folds under the cross-validation setting.

	Method	Overall Metrics			F_1-score per class				
		Accuracy	MF_1	k	W	N1	N2	N3	R
SleepEDF-20	DREAM	**83.91 ± 5.62**	**75.72 ± 6.06**	**0.77 ± 0.08**	**87.94 ± 8.07**	37.22 ± 13.77	**86.92 ± 6.20**	85.32 ± 8.36	**81.18 ± 9.15**
	AttnSleep	81.31 ± 7.36	75.06 ± 7.47	0.74 ± 0.09	85.28 ± 11.07	**39.08 ± 13.31**	86.69 ± 7.96	**88.28 ± 4.51**	75.95 ± 11.00
	DeepSleepNet	80.75 ± 7.64	74.37 ± 7.45	0.74 ± 0.10	85.38 ± 9.73	37.11 ± 12.69	85.83 ± 9.57	87.92 ± 5.58	75.63 ± 10.69
	U-time	78.81 ± 8.34	69.71 ± 8.14	0.70 ± 0.11	80.79 ± 10.83	28.58 ± 10.72	83.45 ± 10.56	84.06 ± 10.82	71.64 ± 16.73
	ResNet+LSTM	78.95 ± 7.55	66.92 ± 7.15	0.70 ± 0.10	82.73 ± 9.76	13.15 ± 12.34	84.40 ± 8.60	84.55 ± 7.25	69.75 ± 12.68
	ResNet+GRU	78.52 ± 7.46	65.03 ± 7.23	0.70 ± 0.10	81.37 ± 9.99	5.95 ± 8.93	84.53 ± 8.21	83.42 ± 8.47	69.87 ± 12.78
	GBM	57.21 ± 6.19	47.66 ± 5.76	0.40 ± 0.08	54.91 ± 11.00	12.34 ± 6.50	66.32 ± 7.69	67.66 ± 12.94	37.10 ± 9.07
SleepEDF-78	DREAM	**81.94 ± 2.18**	**74.79 ± 2.56**	**0.75 ± 0.03**	**91.82 ± 2.04**	41.77 ± 4.02	**84.68 ± 2.81**	78.38 ± 5.57	**77.32 ± 6.18**
	AttnSleep	79.09 ± 2.67	73.59 ± 2.75	0.71 ± 0.04	90.92 ± 2.44	43.66 ± 4.11	82.62 ± 4.04	**78.71 ± 7.81**	72.02 ± 5.97
	DeepSleepNet	76.58 ± 4.70	72.65 ± 4.12	0.68 ± 0.06	86.80 ± 5.49	**46.49 ± 5.43**	79.54 ± 6.79	77.17 ± 6.19	73.26 ± 8.49
	U-time	74.84 ± 2.56	63.81 ± 3.34	0.65 ± 0.04	89.19 ± 2.61	20.57 ± 3.42	80.27 ± 3.87	74.03 ± 7.88	55.00 ± 6.34
	ResNet+LSTM	78.14 ± 2.59	69.88 ± 2.87	0.70 ± 0.04	89.45 ± 2.55	34.30 ± 4.24	82.60 ± 3.03	76.87 ± 7.94	66.17 ± 6.39
	ResNet+GRU	78.11 ± 2.28	69.54 ± 2.92	0.69 ± 0.03	89.86 ± 2.46	33.30 ± 4.85	82.40 ± 3.02	75.72 ± 7.61	66.41 ± 6.50
	GBM	57.43 ± 6.73	43.60 ± 11.69	0.38 ± 0.13	68.89 ± 12.37	13.50 ± 7.11	64.76 ± 5.42	48.14 ± 25.13	22.69 ± 11.75
SHHS	DREAM	**83.90 ± 0.78**	**75.72 ± 1.08**	**0.77 ± 0.01**	**85.16 ± 1.68**	**39.19 ± 2.16**	**85.78 ± 0.68**	82.59 ± 1.35	**85.87 ± 1.43**
	AttnSleep	79.30 ± 2.25	71.25 ± 2.28	0.71 ± 0.03	81.35 ± 2.15	30.00 ± 3.10	84.59 ± 0.54	**82.84 ± 2.26**	77.45 ± 4.55
	DeepSleepNet	76.81 ± 3.15	70.33 ± 2.36	0.68 ± 0.04	76.91 ± 2.93	33.73 ± 4.78	79.63 ± 3.75	79.22 ± 1.94	82.16 ± 3.44
	U-time	79.22 ± 1.56	65.33 ± 3.04	0.70 ± 0.02	76.24 ± 3.81	10.47 ± 12.90	83.24 ± 1.02	76.01 ± 1.22	80.67 ± 2.64
	ResNet+LSTM	80.56 ± 0.53	65.01 ± 0.61	0.72 ± 0.01	80.83 ± 1.32	0.00 ± 0.00	83.67 ± 0.25	81.59 ± 1.97	78.98 ± 0.87
	ResNet+GRU	80.35 ± 0.93	64.72 ± 0.98	0.72 ± 0.01	79.23 ± 2.12	0.00 ± 0.00	83.73 ± 0.48	82.39 ± 0.76	78.27 ± 2.53
	GBM	63.80 ± 0.33	51.00 ± 0.36	0.48 ± 0.00	58.68 ± 1.23	2.53 ± 0.24	69.48 ± 0.60	73.05 ± 1.19	51.27 ± 1.61

dataset used in our experiments. We also extract data from the single-channel EEG (C4-A1) with a sampling rate of 125 Hz.

B. Experimental Setup

Baseline Models. To validate the performance of the proposed model, we compare it with several state-of-the-art models. The details of these models are presented as follows.

- **Gradient Boosting Machine (GBM)**. A classical ensemble model whose prediction is the average computed from predictions of a number of decision tree classifiers.
- **ResNet + Gated Recurrent Unit (ResNet+GRU).** A neural network model with ResNet for feature representation and GRU for prediction.
- **ResNet + Long Short-Term Memory (ResNet+LSTM).** Similar to ResNet+GRU but using LSTM.
- **DeepSleepNet [26].** A neural network model composed of CNN for feature representation and LSTM for prediction.
- **U-time [27].** A neural network model composed of CNN layers with skip connections. Similar to our proposed model, this model makes predictions in a sequential context.
- **AttnSleep [23].** A neural network model composed of CNN layers with different kernel sizes and a multi-head attention mechanism to capture temporal dependencies between different time steps.

Implementation Details. All neural network-based architectures are implemented by PyTorch. For GBM model, we use its Python implementations from Scikit-Learn. We use ADAM optimizer for neural network-based models. Hyperparameters used in the baseline models are from the authors' implementations. For our proposed model, the batch size is set as 64 and the initial learning rate is 0.001. For **SleepEDF** datasets, $\alpha_y = 3500, \alpha_d = 10500, \gamma_d = \gamma_y = 20000$, $\beta = 1$ with KL cost annealing in first ten epochs. For **SHHS** dataset, $\alpha_y = 1000, \alpha_d = 3000, \gamma_d = \gamma_y = 2000$.

Evaluation Metric. We conduct experiments under a cross-validation setting for all three datasets. For **SleepEDF-20** dataset which consists of 20 subjects, we conduct 20-fold cross-validation in which each fold is constructed from one subject. The ratio for train : dev : test is 15 : 4 : 1. For **SleepEDF-78** dataset, 10-fold cross-validation is conducted with the ratio for train : dev : test is approximately 62 : 8 : 8 (each fold consists 7 or 8 subjects). For **SHHS** dataset, experiments are performed under 5-fold cross-validation with the ratio for train : dev : test is approximately 197 : 66 : 66. The accuracy, macro-averaged F_1 score (MF_1), and Cohen's Kappa (k) are used to measure the performances of prediction models for automatic sleep staging. Besides overall evaluations, we also report F_1 scores for each sleep stage category.

C. Results

As shown in Table I, DREAM achieves the best performances compared to other baselines for automatic sleep staging measured by accuracy, MF_1, and k scores. Specifically, it achieves an accuracy of 83.91%, 81.94%, and 83.90% on **SleepEDF-20**, **SleepEDF-78**, and **SHHS** datasets, respectively, which is averagely 4% better than the best baseline model (i.e., AttnSleep). Looking into each sleep stage class, we also observe that DREAM consistently outperforms other methods for W, N2, and R on all three datasets. For N1 class,

DREAM achieves the first-, second- and third-best performances on **SHHS**, **SleepEDF-20** and **SleepEDF-78** datasets, respectively. For N3 class, DREAM achieves the second-best performance on **SleepEDF-78** dataset and the third-best performances on **SleepEDF-20** and **SHHS** datasets. Such improvements indicate the advantage of DREAM by using (i) VAE-based architecture and contrastive learning to learn sleep-relevant and subject-invariant feature representation for sleep segments and (ii) Transformer and CRF models to capture the dynamics of the sleep stage in sequential context at both feature representation and sleep stage classification levels.

V. Conclusions

Automatic sleep staging is crucial in sleep-related disease diagnosis and treatment. In this paper, we propose a novel neural network architecture (DREAM) that leverages EEG signals to predict sleep stages in the sequential context. To tackle the issues of existing methods, DREAM learns sleep-relevant and subject-invariant feature representations and then employs Transformer and CRF models to effectively model sleep dynamics at both feature representation and label prediction levels. DREAM is optimized by the 2-stage training process in which its feature representation network is first trained and then is used in the training of its classification network. We evaluate the prediction performances of DREAM on three sleep staging datasets. The experimental results demonstrate that our proposed model outperforms other state-of-the-art models for automatic sleep staging.

Acknowledgment

This work was funded in part by the National Science Foundation under award numbers IIS-2145625 and CBET-2037398.

References

[1] A. Zarei and B. M. Asl, "Automatic detection of obstructive sleep apnea using wavelet transform and entropy-based features from single-lead ecg signal," *IEEE journal of biomedical and health informatics*, vol. 23, no. 3, pp. 1011–1021, 2018.

[2] P. Memar and F. Faradji, "A novel multi-class eeg-based sleep stage classification system," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 1, pp. 84–95, 2017.

[3] E. Fernandez-Blanco, D. Rivero, and A. Pazos, "Convolutional neural networks for sleep stage scoring on a two-channel eeg signal," *Soft Computing*, vol. 24, no. 6, pp. 4067–4079, 2020.

[4] M. Dut, M. Goodwin, and C. W. Omlin, "Automatic sleep stage identification with time distributed convolutional neural network," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–7.

[5] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 4, pp. 758–769, 2018.

[6] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen, and M. De Vos, "Automatic sleep stage classification using single-channel eeg: Learning sequential features with attention-based recurrent neural networks," in *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2018, pp. 1452–1455.

[7] H. Phan, F. Andreotti, N. Cooray, O. Chen, and M. De Vos, "Seqsleepnet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 3, pp. 400–410, 2019.

[8] A. Guillot, F. Sauvet, E. H. During, and V. Thorey, "Dreem open datasets: Multi-scored sleep datasets to compare human and automated sleep staging," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 9, pp. 1955–1965, 2020.

[9] J. Kim, J.-S. Lee, P. Robinson, and D.-U. Jeong, "Markov analysis of sleep dynamics," *Physical review letters*, vol. 102, no. 17, p. 178104, 2009.

[10] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[11] H. Phan, O. Y. Chen, M. C. Tran, P. Koch, A. Mertins, and M. De Vos, "Xsleepnet: Multi-view sequential model for automatic sleep staging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[12] S. Mousavi, F. Afghah, and U. R. Acharya, "Sleepeegnet: Automated sleep stage scoring with sequence to sequence deep learning approach," *PloS one*, vol. 14, no. 5, p. e0216456, 2019.

[13] A. Supratak and Y. Guo, "Tinysleepnet: An efficient deep learning model for sleep stage scoring based on raw single-channel eeg," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 641–644.

[14] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 661–18 673, 2020.

[15] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," 2016.

[18] G. D. Forney, "The viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.

[19] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. Kamphuisen, and J. J. Oberye, "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 9, pp. 1185–1194, 2000.

[20] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[21] S. F. Quan, B. V. Howard, C. Iber, J. P. Kiley, F. J. Nieto, G. T. O'Connor, D. M. Rapoport, S. Redline, J. Robbins, J. M. Samet *et al.*, "The sleep heart health study: design, rationale, and methods," *Sleep*, vol. 20, no. 12, pp. 1077–1085, 1997.

[22] G.-Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley, and S. Redline, "The national sleep research resource: towards a sleep data commons," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1351–1358, 2018.

[23] E. Eldele, Z. Chen, C. Liu, M. Wu, C.-K. Kwoh, X. Li, and C. Guan, "An attention-based deep learning approach for sleep stage classification with single-channel eeg," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 809–818, 2021.

[24] M. Sokolovsky, F. Guerrero, S. Paisarnsrisomsuk, C. Ruiz, and S. A. Alvarez, "Deep learning for automated feature discovery and classification of sleep stages," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 17, no. 6, pp. 1835–1845, 2019.

[25] P. Fonseca, N. Den Teuling, X. Long, and R. M. Aarts, "Cardiorespiratory sleep stage detection using conditional random fields," *IEEE journal of biomedical and health informatics*, vol. 21, no. 4, pp. 956–966, 2016.

[26] A. Supratak, H. Dong, C. Wu, and Y. Guo, "Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 1998–2008, 2017.

[27] M. Perslev, M. Jensen, S. Darkner, P. J. Jennum, and C. Igel, "U-time: A fully convolutional network for time series segmentation applied to sleep staging," *Advances in Neural Information Processing Systems*, vol. 32, 2019.