Importance Weighting with Adversarial Network for Large-Scale Sleep Staging

Anonymous Authors[1]

Abstract

To develop a generalized automated sleep staging method based on the gold standard modality, electroencephalograms (EEGs), requires a large and accurately labeled training and test set acquired from different individuals with diverse demographics and medical conditions. However, data in the training set may exhibit changes in the EEG patterns that are very different from the data in the test set, due to inherent inter-subject variability, electrode misplacement, and the variability of medication use/response. Training an algorithm on such data without accounting for this diversity can lead to underperformance and a lack of generalizability on novel data. Previous methods have attempted to address this by developing robust representations across all individuals in the dataset using deep transfer learning approaches. However, not all parts of the training data are as relevant as others to the test data. Forcing the alignment of these nontransferable data with the transferable data may lead to a negative impact on the overall performance. This work jointly learns patient-invariant representations and weights features (spectrogram coefficients) to enhance the contribution of relevant features in the final model and decrease the impact of irrelevant features using an unsupervised approach. The proposed method leverages transferable and discriminable knowledge from the training set to the test set. Using a large public database of 42,560 hours of EEG, recorded from 5,793 from Sleep Heart Health Study, we demonstrate that adversarially learning a network with an importance weighting scheme, significantly boosts performance compared to state-of-the-art deep learning approaches in the cross-subject scenario. The proposed method improves, on average, accuracy from 0.81 to 0.94, precision from 0.81 to 0.82, and sensitivity from 0.74 to 0.85.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Approximately one-third of the US population experiences less than the recommended amount of sleep, which in turn, is linked to chronic diseases such as depression, obesity type 2 diabetes, and heart disease (con, 2015). Sleep pathologies are increasingly being recognized as crucial factors in many illnesses, both as effects and causes. In addition, the increasing availability of low-cost sleep monitoring devices and data storage continues to accelerate the field, and the volume of data being collected continues to expand. Since sleep staging and diagnostics is a labor-intensive and expensive process involving highly trained experts, there is therefore a pressing need for automation, particularly in low resource regions of the world. The ground truth for sleep staging remains the multi-lead electroencephalogram (EEG) and the standard rules for sleep staging are still focused on 30-sec windows of data (or 'epochs') with manual labeling by a sleep expert into five stages: Wake (W), Rapid Eye Movement (REM), Non-REM 1 (N1), Non-REM 2 (N2) and Non-REM 3 (N3) (Berry et al., 2012b). In addition to the time and cost involved in manual sleep staging, the significant inter-expert variability remains an issue (Younes & Hanly, 2016). However, the lack of a sizeable public database with heterogeneous populations has limited the development of verifiable algorithms that generalize well across the population. Due to the characteristics and complexities of EEG signals, accurate interpretation of them by human experts requires several years of training. Therefore, developing an accurate classifier with high generalizability on other datasets remains challenging. The non-stationary nature of the EEG signal (Kaplan et al., 2005) and the consequent changes in statistical characteristics of the signal with time, results in poor generalization for a classifier that is trained on a temporally-limited amount of data from an individual recorded at a different time, even for the same subject. Moreover, there exists high inherent inter-subject variability in the characteristics of an EEG due to physiological differences (e.g. skull shape) between individuals,

and because neural activity does not propagate in a similar manner in different subjects. In particular, cortical folding, tissue conductivity, and tissue shapes of brains are different between individuals (Gayraud et al., 2017). Moreover, electrode sensor montages (the points at which the electrodes are attached and the references points) can vary based on the preference of the clinical team or type of underlying ailment under investigation. In addition, each manufacturers' acquisition hardware may filter the EEG differently. Finally, when electrodes are applied, small differences in the locations on the skull may exist, reflecting the EEG technicians' different skill levels or training, or even attentiveness on a given day. All these factors lead to significant variabilities in EEG signals, which lead to different joint distributions, $P(X, Y)$ between different recordings, where X and Y are the feature and label space, respectively.

Recently, multiple authors have focused on developing automated sleep scoring approaches based on applying deep learning (DL) methods (Biswal et al., 2018; Malafeev et al., 2018; Tagluk et al., 2010; Perslev et al., 2019). Due to the spatio-temporal nature of the information in the EEG, most convolutional and recurrent processing methodologies are quite suitable for EEG analysis. However, these predictive models do not generalize well to unseen patients due to inter-subject variability, as explained earlier. The typical solution is to further fine-tune these networks on new patients, where it is expensive and time-consuming to obtained labeled data from them, and which further reduces their real-life application in a clinical setting. Hence, there is strong motivation to establishing effective algorithms to reduce the labeling consumption by leveraging readily-available labeled data from different, but related patients. As note, due to inherent inter-subject variability, information from some training patients may not transfer well to the test set. Here, a new framework is proposed to quantify the transferability of features in the adversarial network to select relevant features and weight them based on their transferability and discriminability. Using the largest public EEG database for sleep staging, 5,793 patients ($\approx$ 42,560 hours) EEGs from Sleep Heart Health Study (SHHS), an adversarially learned network with a importance weighting scheme is used to significantly boost performance compared to state-of-the-art deep learning approaches in the cross-subject scenario.

2. Related Work

As noted, the spatial shift in data can be caused by the variation of sensors' location on the brain in different datasets or mismatching of electrodes in one dataset. This issue can be partly solved by finding an invariant representation across data-sets (Biswal et al., 2018). In the literature, it has been shown that Symmetric Positive Definite (SPD) matrices provide a strong ability to provide useful representations of brain signals (Congedo et al., 2017; Barachant et al., 2010). The covariance matrix is a typical example of an SPD matrix, which has been employed in several studies (Saifutdinova et al., 2019; Rodrigues et al., 2019; Li et al., 2012). These studies showed that using second-order statistics of multi-channel signals reduces inter-subject and intra-subject variabilities between EEG signals. The spatial covariance matrix is particularly good at separating useful information about the brain's functional connectivity structure (Barachant et al., 2010) and creates a feature space that is comparable across subjects. Moreover, it has been shown that SPD matrices have excellent robustness to the considerable variability of real-world environmental conditions such as instrument noise (Congedo et al., 2017).

Other studies (Li et al., 2019; Ma et al., 2019; Tang & Zhang, 2020) tackled this challenge using domain adaptation techniques to increase generalization of a model that is trained on EEG data and tested on unseen subjects in Brain-Computer Interface (BCI), Motor Imagery (MI), and emotion recognition tasks. In the literature, it has been shown that domain adaption, which can be considered as a particular case of transfer learning, solved dataset bias of domain shifts, which is common in biomedical applications. The key technique of domain adaption is to diminish the discrepancy between these two distributions using the Maximum Mean Discrepancy (MMD) metric (Long et al., 2015). Previous studies, which have employed domain adaptation in biomedical time-series data, bridge the training and test datasets from different individuals by learning subject-invariant representations or estimating feature importance using labeled training features and unlabeled test features (Ma et al., 2019; Li et al., 2019; Jayaram et al., 2016).

Other methods to increase the generalization ability of a model involve transfer learning - finding subsets of known (labeled) subjects to initialize a classifier for training on a new subject (Zanini et al., 2017). Bolagh et al. (Bolagh et al., 2016; 2017) proposed subject-selection and subject clustering to select relevant individuals based on the similarity between the EEG pattern of different individuals. Raza et al. (Raza & Samothrakis, 2019) proposed bagging methods to handle mismatching between training and test distributions. Chai et al. (Chai et al., 2017) proposed an adaptive subspace feature matching to match both the marginal and conditional distributions between EEG data from different sessions/subjects. All of these studies tried to develop a method for reducing inter-subject variability by removing the irrelevant subjects in the training set and enabling efficient knowledge transfer from previous subjects to a new unseen patient.

Sors et al (Sors et al., 2018) used a 14-layer convolutional neural network (CNN) which used an epoch of raw data from channel C4-A1, along with the next and previous

two epochs to achieve an accuracy of 87% on the SHHS dataset. Phan et al (Phan et al., 2019) trained a CNN to simultaneously classify one epoch and its neighbors from the short-time Fourier transform of the C4-A1 EEG channel, ROC-LOC EOG channel and Chin1-Chin2 EMG channel, then used multiplicative voting to aggregate each classification, which achieved an accuracy 82.3% on the Sleep-EDF database and 83.6% on the MASS dataset. Biswal et al. (Biswal et al., 2018) used a recurrent convolutional neural network on the spectrogram of the EEG in each epoch to achieve an accuracy of 77.7% when using the C4-A1 and C3-A2 channels of the SHHS dataset, 81.9% accuracy using the C4-A1 and C3-A2 of their own private dataset and 87.5% accuracy using the F3-M2, F4-M1, C3-M2, C4-M1, O1-M2 and O2-M1 channels of their own private dataset. Zhang et al. (Zhang et al., 2019) fed spectrograms into CNN layers and an LSTM layer to assess the generalization capability of their model by testing their model on two different datasets. Their model achieved F1-score of 0.81 and Cohen's Unweighted kappa of $\kappa = 0.82$. These methods have recently gained attention since they simplify processing pipelines through end-to-end learning, removing the need for domain-specific knowledge for feature engineering. This is clearly appealing, but it presents some dangers, and ignoring the nature of the EEG and how it is acquired, has limited the impact of DL in this domain. Although DL architectures have been very successful in processing complex data such as images, text, and audio signals (Liu et al., 2017; Hershey et al., 2017), the generalization and interpretation of a DL method across different patients are still the main challenges for using DL in most clinical applications. DL architectures are hard to 'trust' due to their complexity and extreme non-linearity, which further reduces their real-life application in a clinical setting.

Recently, the use of generative adversarial networks (GANs) (Goodfellow et al., 2014) to handle temporal and spatial shifts has received more attention (Tzeng et al., 2017; Sankaranarayanan et al., 2018; Liu et al., 2019). Notably, Ganin et al. constructed a two-player minimax game (rather like the appriach of GANs), in which the first player discriminates between training and test sets and the second player is adversarially trained to deceive the discriminator and extract transferable features (Ganin et al., 2016). These networks try to align the representations extracted from all EEG channels across all subjects. It is evident that some parts of the brain are more involved in a given task (or are more active during a given state), thus all channels are not equally transferable. Moreover, some parts of the EEG pattern are significantly dissimilar across subjects. Those patterns might be related to the specific health history of the patient, which could affect EEG patterns. Therefore, forcing the use of the irrelevant channels, and their EEG patterns, may have a large impact on overall performance. An attention mechanism (Vaswani et al., 2017) is an effective method to focus on essential regions of data, with numerous successes in deep learning tasks such as classification, segmentation, and detection.

3. Methods

Ganin et al. inspired the idea of GANs and used the same idea for the domain adaptation problem, where adaptation behavior is achieved via adversarial training (Ganin et al., 2016). The feature extractor, similar to the generator in GANs, tries to perform some transformation on data from two domains such that the transformed features have the same distribution. The second network, (a discriminator network), similar to GANs, should be able to classify the domains as source (i.e. training features) and target (i.e. test features). This is achieved by training two networks in such a way that the feature extractor is trying to confuse the domain discriminator via adversarial training. The key idea of domain-adversarial training is to use a Gradient Reversal Layer (GRL), placed between feature extractor and domain discriminator. The GRL acts like an identity function during forwarding propagation and multiplies the gradient by a certain negative constant during the backpropagation, leading to the opposite of gradient descent. The adversarial network has three components; a feature extractor ($G_f(\cdot, \theta_f)$), a label classifier ($G_y(\cdot, \theta_y)$), and a domain discriminator ($G_d(\cdot, \theta_d)$). The feature extractor is a neural network that learns an invariant representation across domains by finding a robust transformation. The label classifier is a neural network that classifies extracted features from the source (labeled) domain. Finally, the domain discriminator is a neural network that predicts whether the feature is coming from the source domain or target domain. The optimization of this framework can be written as follows:

$$C(\theta_f, \theta_y, \theta_d) = \frac{1}{n_{tr}} \sum_{\mathbf{x}_i \in \mathcal{D}_{tr}} L_y(G_y(G_f(\mathbf{x}_i)), y_i) - \frac{\lambda}{n} \sum_{\mathbf{x}_i \in \mathcal{D}_{tr} \cup \mathcal{D}_{te}} L_d(G_d(G_f(\mathbf{x}_i)), d_i) \tag{1}$$

where $n = n_{tr} + n_{te}$, n_{tr} and n_{te} are number of sample in training (source) and test (target) sets, respectively, and λ is a hyper-parameter that trades-off the domain discriminator loss L_d with the classification loss L_y corresponding to the training classifier G_y.

As mentioned earlier, these networks try to align the extracted features from all EEG from the whole population. It is obvious that forcefully aligning the feature from two dissimilar patients might inject negative information to the network. Therefore, in this work, we develop an algorithm to transfer useful extracted features from the training set to test set while mitigating from irrelevant features. The

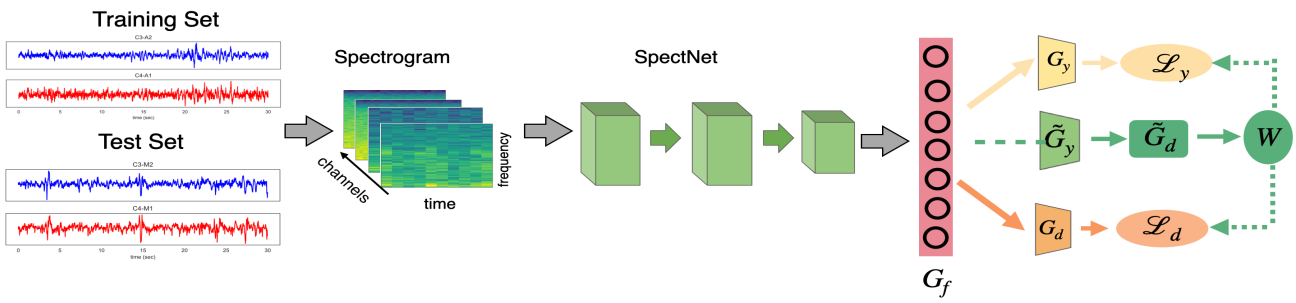


Figure 1. The proposed method for generalized sleep staging problem, where G_f is the feature extractor, G_y is the training classifier, G_d is domain discriminator (involved in adversarial training) for alignment features from traing and test; $\tilde{G}_d$ is the auxiliary domain discriminator (uninvolved in adversarial training) that quantifies the transferability W of each training feature, and $\tilde{G}_y$ is the auxiliary label predictor encoding the discriminative information to the auxiliary domain discriminator $\tilde{G}_d$. Best viewed in color

proposed method uses the adversarial network and combines it with a weighting scheme. Weights automatically measure the transferability and discriminability. Let $w(\mathbf{x}_i^{tr})$ be the weight of each training feature $\mathbf{x}_i^{tr}$, which measures its transferability to test set; thus, features with high weights contribute more to the final model, and the impact of features with lower weight is decreased. The entropy minimization principle encourages the low-density separation between classes by minimizing the entropy of class-conditional distribution on the test set, which is useful for refining the classifier adaptation. In this work, we use this principle to quantify the uncertainty of a test feature's predicted label. Let $\hat{y} = G_y(G_f(x_j^{te})) \in \Re^C$, the entropy loss to quantify the uncertainty of a test features's predicted label is $H(G_y(G_f(x_j^{te}))) = -\sum_{c=1}^{C} \hat{y}_{j,c}^{te} \log \hat{y}_{j,c}^{te}$.

By Re-weighting training features in the loss of the discriminator G_d, and the training classifier G_y, and using the entropy minimization principle, the optimization of this framework can be written as follows:

$$E_{G_y} = \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} w(\mathbf{x}_i^{tr}) L(G_y(G_f(\mathbf{x}_i^{tr}), y_i^{tr})) + \frac{\gamma}{n_{te}} \sum_{n=1}^{n_{te}} H(G_y(G_f(\mathbf{x}_j^{te}))) \quad (2)$$

$$E_{G_d} = -\frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} w(\mathbf{x}_i^{tr}) \log(G_d(G_f(\mathbf{x}_i^{tr}))) + \frac{1}{n_{te}} \sum_{n=1}^{n_{te}} \log(1 - G_d(G_f(\mathbf{x}_j^{te}))) \quad (3)$$

Where γ is a hyper-parameter to trade-off the labeled training and unlabeled test features. The transferability weighting framework can be trained end-to-end by a minimax optimization procedure as follows:

$$(\hat{\theta}_f, \hat{\theta}_y) = \arg \min_{\theta_f, \theta_y} E_{G_y} - E_{G_d}$$
$$(\hat{\theta}_d) = \arg \max_{\theta_d} E_{G_d} \quad (4)$$

An auxiliary discriminator $\tilde{G}_d$ is used to measure feature's transferability. This discriminator is not involved in adversarial training, i.e., the features G_f are not learned to confuse $\tilde{G}_d$. The output of the auxiliary discriminator $\tilde{G}_d$ is a probability, where having lower probability means the training features are similar to the test set. Besides, the labeled information from the training set is injected into the auxiliary discriminator $\tilde{G}_d$, to enhance the discriminability. Therefore, the output of the auxiliary discriminator can be written as follows:

$$\tilde{G}_d(G_f(\mathbf{x}_i)) = \sum_{c=1}^{5} \tilde{G}_y^c(G_f(\mathbf{x}_i)) \quad (5)$$

Where $\tilde{G}_y^c(G_f(\mathbf{x}_i))$ can be interpreted as the probability of each feature $\mathbf{x}_i$ belonging to class c. Therefore, the weight for measuring the transferability and discriminability is defined as:

$$w(\mathbf{x}_i^{tr}) = 1 - \tilde{G}_d(G_f(\mathbf{x}_i^{tr})) \quad (6)$$

The auxiliary label predictor $\tilde{G}_y$ is trained with the leaky-softmax by a multitask loss over 5 one-vs-rest binary classification tasks for the 5-stage sleep staging problem:

$$E_{\tilde{G}_y} = -\frac{\lambda}{n_{tr}} \sum_{i=1}^{n_{tr}} \sum_{c=1}^{5} [y_{i,c}^{tr} \log(\tilde{G}_y^c(G_f(\mathbf{x}_i^{tr}))) + (1 - y_{i,c}^{tr}) \log(1 - \tilde{G}_y^c(G_f(\mathbf{x}_i^{tr})))] \quad (7)$$

where $y_{i,c}^{tr}$ denotes whether class c is the ground-truth label for training feature $\mathbf{x}_i^s$, and λ is a hyper-parameter. There-

fore, training of the auxiliary discriminator is done as:

$$E_{\tilde{G}_d} = -\frac{1}{n_{tr}}\sum_{i=1}^{n_{tr}}\log(\tilde{G}_d(G_f(\mathbf{x}_i^{tr}))) - \frac{1}{n_{te}}\sum_{i=1}^{n_{te}}\log(\tilde{G}_d(G_f(\mathbf{x}_i^{te}))) \quad (8)$$

Weights in each mini-batch of batch size B are normalized as $w(\mathbf{x}) \leftarrow \frac{w(\mathbf{x})}{\frac{1}{B}\sum_{i=1}^{B} w(\mathbf{x}_i)}$ to make data from patients comparable. Thus, the overall optimization can be written as follow:

$$(\hat{\theta}_f, \hat{\theta}_y) = \arg\min_{\theta_f,\theta_y} E_{G_y} - E_{G_d}$$
$$(\hat{\theta}_d) = \arg\min_{\theta_d} E_{G_d}$$
$$(\hat{\theta}_{\tilde{y}}) = \arg\min_{\theta_{\tilde{y}}} E_{G_{\tilde{y}}} - E_{\tilde{G}_d} \quad (9)$$

4. Experiments

4.1. Data

Sleep Heart Health Study: The SHHS database consists of two rounds of polysomnographic recordings (SHHS-1 and SHHS-2) sampled at 125 Hz in a sleep center environment. The data used in the study are de-identified, and therefore an ethics/institutional review board waiver was provided for this research. Following (Duggal et al., 2020), we use only the first round (SHHS-1) containing polysomnographic records from participants included 52.9% women and 47.1% men, over two channels (C4-A1 and C3-A2). Recordings were manually classified into one of six classes (W, REM, N1, N2, N3, and N4). As suggested in (Berry et al., 2012a), we merge N3 and N4 stages into a single N3 stage. Table 1 shows number of sleep stages per class.

Table 1. Number of subjects and epochs per class for each dataset

Dataset	# Subjects	# Wake	# N1	# N2	# N3	# REM
SHHS	5,792	1,690,997	217,535	2,397,062	739,230	817,330
train	4,054	1,183,252	152,744	1,678,666	515,730	5,725,780
test	1,738	507,745	64,791	718,396	223,500	244,752

4.2. Preprocessing

Before presenting the signal to the network, preprocessing is performed to reduce the negative effects of signal artifacts. Two filters were applied to the EEG channels: a notch filter to remove 60 Hz power line interference, and a band-pass filter to allow a frequency range of 0.5-180 Hz through. Normalization of EEG amplitude is then carried out as the last step to minimize the difference in EEG amplitudes using min-max normalization across different subjects. After the preprocessing steps, spectrograms are generated for each EEG channel to transform data to the time-frequency domain. Each 30-second epoch is transformed into log-power spectra via a short-time Fourier transform (STFT) with a window size of two seconds and a 50 % overlap, followed by logarithmic scaling. A Hamming window and 256-point Fast Fourier Transform (FFT) are used on each epoch. This results in an image $\mathbf{S} \in R^{F\times T}$ where $F = 129$ (the number of frequency bins), and $T = 29$ (the number of spectral columns).

4.3. Network Implementation

For extracting features for the adversarial neural network, we use the same architecture of Biswal et al. (Biswal et al., 2018). It includes a 3-layer of 1-D CNN (kernel size = 3), which was applied to each EEG channel, followed by batch normalization (BatchNorm), rectified linear (ReLU) units, and max pooling units, we called it as SpectNet here. A cross-entropy loss function is used as a discriminator $\mathcal{L}_d$ and classification $\mathcal{L}_y$. We apply back-propagation to train the classifier layer and all domain discriminators. Mini-batch stochastic gradient descent (SGD) is employed with the momentum of 0.95 using the learning rate and progressive training strategies as in (Ganin et al., 2016) to learn the weights of a deep neural network and hyper-parameter are optimized with importance weighted cross-validation (Sugiyama et al., 2007). To address the class imbalance, we balance each batch for positive and negative examples, which leads to oversampling the positive class. The proposed methods were implemented with PyTorch 1.0 and Python3.6.

4.4. Results

The training data were randomly selected from 4054 patients ($\approx$ 70% of the total population) of the SHHS. Classification results are based on the test set ($\approx$ 30% of the total population = 1738 patients), which not included in the training set shown in Table (2). We also compare previous methods for sleep staging on this dataset. The proposed method outperforms all other methods with respect to average accuracy, sensitivity and F1-score, and Kappa, showing that SSA performs well with different base networks for sleep staging tasks. To evaluate the proposed approach performance and see how adversarial domain adaption network helps to develop a model with high generalizability, we initially conduct simple experiments. Similar to the literature on sleep stage assessment, to evaluate model performance, accuracy, specificity, sensitivity, and F1-score per class are reported. The other primary metric that we have used for performance evaluation of our proposed method is Cohen's Kappa coefficient (κ). This metric measures the agreement between the labels obtained by the algorithm and the ground truth annotations.

Table 2. Wide single-column table in a twocolumn document.

Sleep Stages	Precision	Sensitivity	F1-Score	Kappa	Imbalanced Acc	Number of Epochs
Wake	0.96	0.93	0.95	0.92	0.97	507745
REM	0.86	0.91	0.88	0.86	0.96	244752
N1	0.45	0.66	0.53	0.52	0.95	64791
N2	0.95	0.91	0.93	0.88	0.94	718396
N3	0.91	0.90	0.90	0.89	0.97	223500
avg	0.82	0.86	0.84	0.82	0.96	-

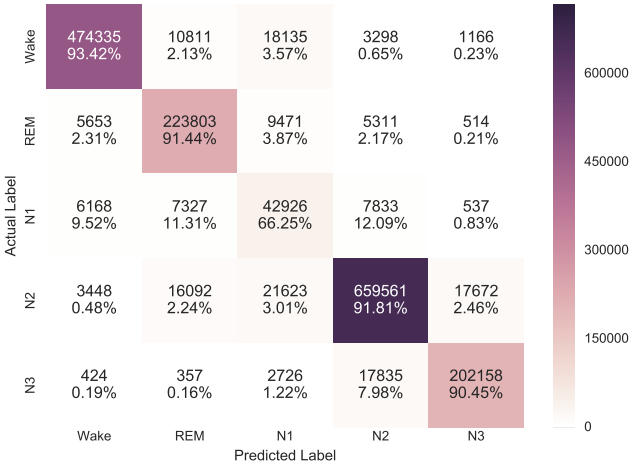
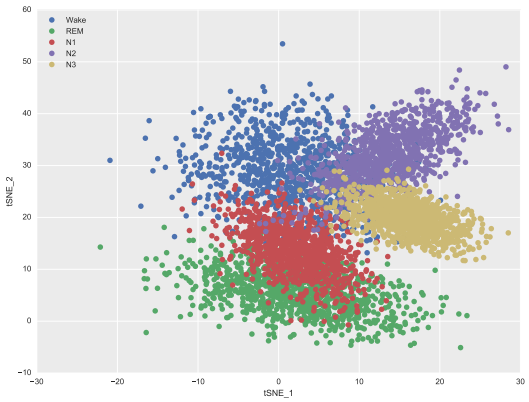


Figure 2. Confusion Matrix for test set, which includes 1738 patients from SHHS dataset. The model is trainined on training set, EEGs from 4054 patients. Note: training and test set do not overlap in patients; i.e. cross-subject scenario

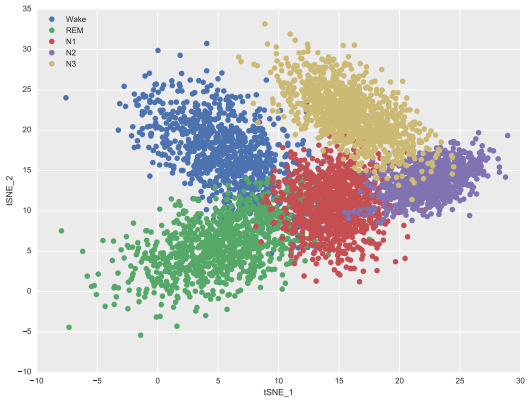
Figures (3) show the t-SNE embedded of the features learned by CNN, the proposed method methods on the SHHS dataset. It shows that features determined by the proposed practice can better discriminate test features compared to previous methods, specifically reduce the confusion between the N1 stage with other stages.

Table (3) summarizes the average results and compares them to the state-of-the-art. The methods proposed previously by Biswal et al. (Biswal et al., 2018), Zhange et al. (Zhang et al., 2019), and Sors et al. (Sors et al., 2018) which evaluated their method on the SHHS dataset imply that the knowledge from irrelevant features from patients lead to a negative impact on the overall performance on the test set. The proposed method down-weights dissimilar features to enhances the generalizability. Moreover, the proposed method pays more attention to relevant features to test set by assigning suitable weights. Injecting label information into the discriminator improves the discriminability of the model.

Based on this experiment, the proposed method boosts the



((a)) Without adversarial training



((b)) With adversarial training

Figure 3. t-SNE visualization of the last hidden layer representations in the feature extraction network without/with adversarial training. Colored points represent the different stages, showing how the algorithm discriminate classes. Wake (blue), REM (green), N1 (red), N2 (purple) and N3 (flax).

performance by 5% on average. It shows that using adversarial network with importance weighting framework boosts the N1 class performance. The N1 stage is often confused for wake and N2, and it is considered a transition period from being awake to falling asleep. Colten et al. ((Altevogt et al., 2006)) defined the N1 stage as "active sleep", which means N1 may also occur between other stages of sleep, such as between N3 and REM. Therefore, it is often confused with many other stages, as we can see in confusion matrices in Figure (2).

5. Conclusion

In this work, adversarial training with a weighting scheme was proposed for the sleep staging task across a heterogeneous dataset, which includes EEGs from 5792 patients. Inherent inter-subject variability, electrode misplacement, and heterogeneity in the medical history of patients in a large dataset may lead to an algorithm having poor generalization across subjects/dataset. Potentially, individuals with different biomedical demographics and phenotypes would provide enough diversity in the dataset. However, a conventional network cannot be robust to such variabilities, given the need to factor in differences in montages, electrode placement errors, the dataset would likely be prohibitively large. The proposed method uses an adversarial network with an importance weighting framework to assign a weight for each sample based on its transferability and discriminability. Features from patients with higher weight contribute more to the final model, and irrelevant features are down-weighted to mitigate their negative impact. The proposed method achieves state-of-the-art performance (without prior knowledge) on 30% ($\approx$ 1738 patient) of the total patients. The method developed in this work can be applied to other biomedical signals (e.g. the electrocardiogram, electromyogram EMG and photoplethysmogram (PPG), where multiple datasets from different hospitals are recorded for the same task. The ultimate goal of the research presented here, however, is to solve real-world automate sleep stage classification problems.

References

Recommended amount of sleep for a healthy adult: a joint consensus statement of the american academy of sleep medicine and sleep research society. *Sleep*, 38(6):843–844, 2015.

Altevogt, B. M., Colten, H. R., et al. *Sleep disorders and sleep deprivation: an unmet public health problem*. National Academies Press, 2006.

Barachant, A., Bonnet, S., Congedo, M., and Jutten, C. Common spatial pattern revisited by riemannian geometry. In *2010 IEEE International Workshop on Multimedia Signal Processing*, pp. 472–476. IEEE, 2010.

Berry, R. B., Brooks, R., Gamaldo, C. E., Harding, S. M., Marcus, C., Vaughn, B. V., et al. The aasm manual for the scoring of sleep and associated events. *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine*, 176:2012, 2012a.

Berry, R. B., Budhiraja, R., Gottlieb, D. J., Gozal, D., Iber, C., Kapur, V. K., Marcus, C. L., Mehra, R., Parthasarathy, S., Quan, S. F., et al. Rules for scoring respiratory events in sleep: update of the 2007 aasm manual for the scoring of sleep and associated events. *Journal of clinical sleep medicine*, 8(05):597–619, 2012b.

Biswal, S., Sun, H., Goparaju, B., Westover, M. B., Sun, J., and Bianchi, M. T. Expert-level sleep scoring with deep neural networks. *Journal of the American Medical Informatics Association*, 25(12):1643–1650, 2018.

Bolagh, S. N. G., Shamsollahi, M. B., Jutten, C., and Congedo, M. Unsupervised cross-subject BCI learning and classification using riemannian geometry. In *24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2016)*, 2016.

Bolagh, S. N. G., Clifford, G., et al. Subject selection on a riemannian manifold for unsupervised cross-subject seizure detection. *arXiv preprint arXiv:1712.00465*, 2017.

Chai, X., Wang, Q., Zhao, Y., Li, Y., Liu, D., Liu, X., and Bai, O. A fast, efficient domain adaptation technique for cross-domain electroencephalography (eeg)-based emotion recognition. *Sensors*, 17(5):1014, 2017.

Congedo, M., Barachant, A., and Bhatia, R. Riemannian geometry for eeg-based brain-computer interfaces; a primer and a review. *Brain-Computer Interfaces*, 4(3):155–174, 2017.

Duggal, R., Freitas, S., Xiao, C., Chau, D. H., and Sun, J. Rest: Robust and efficient neural networks for sleep monitoring in the wild. *arXiv preprint arXiv:2001.11363*, 2020.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

Gayraud, N. T., Rakotomamonjy, A., and Clerc, M. Optimal transport applied to transfer learning for p300 detection. 2017.

Table 3. Performance of class imbalanced model compared to other studies

Method	# patients	precision-Wake	precision-N1	precision-N2	precision-N3	precision-REM	Overall precision	Kappa
(Biswal et al., 2018)	10000	84.5%	56.2%	88.4%	85.4%	92%	81.3%	0.79
(Sors et al., 2018)	5793	91%	35%	89%	85%	86%	77.2%	0.81
(Zhang et al., 2019)	5793	92%	37%	91%	77%	88%	77%	0.82
Proposed method	5792	97%	45%	95%	91%	86%	82%	0.82

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 131–135. IEEE, 2017.

Jayaram, V., Alamgir, M., Altun, Y., Scholkopf, B., and Grosse-Wentrup, M. Transfer learning in brain-computer interfaces. *IEEE Computational Intelligence Magazine*, 11(1):20–31, 2016.

Kaplan, A. Y., Fingelkurts, A. A., Fingelkurts, A. A., Borisov, S. V., and Darkhovsky, B. S. Nonstationary nature of the brain activity as revealed by eeg/meg: methodological, practical and conceptual challenges. *Signal processing*, 85(11):2190–2212, 2005.

Li, J., Qiu, S., Du, C., Wang, Y., and He, H. Domain adaptation for eeg emotion recognition based on latent representation similarity. *IEEE Transactions on Cognitive and Developmental Systems*, 2019.

Li, Y., Wong, K. M., and de Bruin, H. Electroencephalogram signals classification for sleep-state decision–a riemannian geometry approach. *IET signal processing*, 6(4): 288–299, 2012.

Liu, H., Long, M., Wang, J., and Jordan, M. Transferable adversarial training: A general approach to adapting deep classifiers. In *International Conference on Machine Learning*, pp. 4013–4022, 2019.

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., and Alsaadi, F. E. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26, 2017.

Long, M., Cao, Y., Wang, J., and Jordan, M. I. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015.

Ma, B.-Q., Li, H., Zheng, W.-L., and Lu, B.-L. Reducing the subject variability of eeg signals with adversarial domain generalization. In *International Conference on Neural Information Processing*, pp. 30–42. Springer, 2019.

Malafeev, A., Laptev, D., Bauer, S., Omlin, X., Wierzbicka, A., Wichniak, A., Jernajczyk, W., Riener, R., Buhmann, J., and Achermann, P. Automatic human sleep stage scoring using deep neural networks. *Frontiers in neuroscience*, 12:781, 2018.

Perslev, M., Jensen, M., Darkner, S., Jennum, P. J., and Igel, C. U-time: A fully convolutional network for time series segmentation applied to sleep staging. In *Advances in Neural Information Processing Systems*, pp. 4417–4428, 2019.

Phan, H., Andreotti, F., Cooray, N., Chén, O. Y., and De Vos, M. Seqsleepnet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(3):400–410, 2019.

Raza, H. and Samothrakis, S. Bagging adversarial neural networks for domain adaptation in non-stationary eeg. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, 2019.

Rodrigues, P., Congedo, M., and Jutten, C. A data imputation method for matrices in the symmetric positive definite manifold. 2019.

Saifutdinova, E., Congedo, M., Dudysova, D., Lhotska, L., Koprivova, J., and Gerla, V. An unsupervised multichannel artifact detection method for sleep eeg based on riemannian geometry. *Sensors*, 19(3):602, 2019.

Sankaranarayanan, S., Balaji, Y., Castillo, C. D., and Chellappa, R. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8503–8512, 2018.

Sors, A., Bonnet, S., Mirek, S., Vercueil, L., and Payen, J.-F. A convolutional neural network for sleep stage scoring from raw single-channel eeg. *Biomedical Signal Processing and Control*, 42:107–114, 2018.

Sugiyama, M., Krauledat, M., and MÃžller, K.-R. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(May): 985–1005, 2007.

Tagluk, M. E., Sezgin, N., and Akin, M. Estimation of sleep stages by an artificial neural network employing

eeg, emg and eog. *Journal of medical systems*, 34(4): 717–725, 2010.

Tang, X. and Zhang, X. Conditional adversarial domain adaptation neural network for motor imagery eeg decoding. *Entropy*, 22(1):96, 2020.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7167–7176, 2017.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Younes, M. and Hanly, P. J. Minimizing interrater variability in staging sleep by use of computer-derived features. *Journal of Clinical Sleep Medicine*, 12(10):1347–1356, 2016.

Zanini, P., Congedo, M., Jutten, C., Said, S., and Berthoumieu, Y. Transfer learning: a riemannian geometry framework with applications to brain-computer interfaces. *IEEE Transactions on Biomedical Engineering*, 2017.

Zhang, L., Fabbri, D., Upender, R., and Kent, D. Automated sleep stage scoring of the sleep heart health study using deep neural networks. *Sleep*, 42(11):zsz159, 2019.