Distribution Agreement

Signature:

______________________________ ____________

Samaneh Nasiri Ghosheh Bolagh Date

Generalizable Machine Learning Methods for Electrophysiology

By

Samaneh Nasiri Ghosheh Bolagh
Doctor of Philosophy

Computer Science and Informatics

Gari D. Clifford, DPhil.
Advisor

Qiao Li, Ph,D.
Committee Member

Rishi Kamaleswaran, Ph.D.
Committee Member

Matthew Reyna, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Generalizable Machine Learning Methods for Electrophysiology

By

Samaneh Nasiri Ghosheh Bolagh
B.Sc., Iran University of Science and Technology, Iran, 2013
M.Sc., Sharif University of Technology, 2015
M.Sc., Emory School of Medicine, GA, 2019

Advisor: Gari D. Clifford, DPhil.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2020

Abstract

Generalizable Machine Learning Methods for Electrophysiology
By Samaneh Nasiri Ghosheh Bolagh

Brain pathology is increasingly recognized as a crucial factor in many illnesses. As the availability of low-cost brain monitoring devices important, the volume of data continues to expand. The need for automated brain monitoring diagnostics is, therefore, more acute, particularly in low resource regions of the world. The ground truth for brain monitoring remains the multi-lead electroencephalogram (EEG) and standard practice is still focused on visual inspection of EEGs. However, this is a costly and time-consuming procedure. Moreover, the lack of significant public databases (of 10,000-100,000 patients) of heterogeneous populations has limited the development of verifiable algorithms that generalize well across populations. Due to the characteristics and complexities of EEG signals, accurate interpretation of EEG signals by human experts requires several years of training.

Developing accurate classifiers with high generalizability on other datasets is a challenging task in this area. Due to the non-stationary nature of the EEG signal, the statistical characteristics of the signal vary with time; therefore, a classifier that is trained on a temporally-limited amount of data from an individual may poorly generalize on EEG data recorded at a different time on the same subject. Another issue with the low generalizability issue in EEG data is related to high inherent inter-subject variability in the way an EEG manifests, which limits the usefulness of EEG applications. This phenomenon arises due to physiological differences (e.g. skull shape) between individuals, and 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 across people. Moreover, electrode sensor montages (the points at which the electrodes are attached and the references points) may differ and different 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' variety of training or even attentiveness on a given day. All these factors lead to significant variabilities in EEG signals, which lead to different joint distributions between the feature and label space of different recordings. Therefore, the transferability of the trained model on unseen subjects is degraded. The reason behind this problem is the assumption in machine learning techniques that training and test data should be drawn from the same distribution, an assumption that does not necessarily hold in large biomedical datasets.

This thesis has addressed this challenge via two approaches: 1) measuring the boosting effect of machine learning and deep learning methods in a non-Euclidean space to mitigate the effects of intra and inter-subject variability in seizure detection and sleep staging; 2) developing adversarial networks with attention mechanisms and importance weighting to learn both transferable and discriminative representations, and enhance the generalizability of the model for classifying sleep stages, and seizure detection.

Generalizable Machine Learning Methods for Electrophysiology

By

Samaneh Nasiri Ghosheh Bolagh
B.Sc., Iran University of Science and Technology, Iran, 2013
M.Sc., Sharif University of Technology, 2015
M.Sc., Emory School of Medicine, GA, 2019

Advisor: Gari D. Clifford, DPhil.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2020

Acknowledgments

There are no proper words to convey my sincere gratitude and respect for my thesis and research advisor, Prof Gari D. Clifford. I was always inspired by his in-depth knowledge of machine learning, cleverness, and visionary. His valuable feedback, advice, and encouragement helped me become a researcher and realize the power of critical reasoning. In addition to our academic collaboration, I have received his support and kindness in other aspects of life. Due to the travel ban, I have not seen my father for four years, but I can genuinely say Gari is my father here in the US.

I wish to thank the members of my dissertation committee: Prof. Li Qiao, Prof. Rishi Kamaleswaran, Prof. Matthew Reyna, for generously offering their time, support, insightful suggestions, and comments throughout the preparation and review of this thesis.

I thank all the wonderful people in the department of Biomedical Informatics for making the Ph.D. experience fruitful and pleasant. Special thanks go to Cathy, Julia, Robert, and Jim. I also thank my friends in the department Camilo, Pradyumna, Ayse, Zifan, Nick, Azade, Nasim, Parisa, Mohsen, Fereshteh, and Sahar, who made this journey joyful for me. I'm always thankful to my close friends, Parisa, Ramin, Ehsan, Soheil, Ali, and Bahare, for their support.

My gratitude goes out to my parents, without whom I would not have been able to reach my goals. I am thankful for their prayers, supports, and cares throughout my life. I also thank for heart-warming kindness from my siblings, Samira and Mojtaba. Without their encouragement from the beginning of my life, it would have been impossible to reach this stage. I would also like to thank Saeed for his unconditional kindness and support during these years.

At the end, I would like to express my sincere gratitude and appreciation to the true heroes in this world, all doctors and nurses, for the time and work they have put into helping others during this frightening time due to COVID-19.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis contributes to approaches for developing generalizable machine learning techniques for electrophysiology. This work focuses on multi-lead EEG signals [80], which are considered ground-truth signals for recognizing brain pathology. Automated brain monitoring diagnostics will play a crucial role in shortening the duration of clinical trials and improving patient experience and outcomes. There are abundant studies in the field; however, most of them are patient-specific and cannot generalize well across the population and datasets. Due to the non-stationary nature of the EEG signal [46] and different varaibilities in EEG signal, developing accurate classifiers with high generalizability on other datasets is a challenging task. This thesis proposes methods to improve robustness of machine learning approaches to simultaneously deal with intra- and inter-subject variation due to differences in patient brain morphology and clinical recording practices.

1.1 Motivation

EEG signals can capture electrical brain activities that can be used in diagnosing brain conditions, such as epilepsy, sleep abnormalities, Parkinson's, schizophrenia, and Alzheimer's disease. This work focuses on developing generalizable algorithms for detecting seizures in patients with epilepsy and sleep staging using EEG signals. It should be noted that the proposed techniques could generalize on different abnormalities as well as beyond EEG to other multilead modalities such as the electrocardiogram, electromyogram, and ultrasound arrays, for example, and could apply to areas such as fetal monitoring, cardiac arrhythmia diagnostics and blood flow imaging.

Epilepsy is the second most common neurological disorder where neurons produce abnormal signals and cause seizures, affecting at least 3.4 million in the US and 65 million people worldwide. Over a lifetime, about 1 in 26 US people will be diagnosed with epilepsy, according to the CDC's reports [1]. The physical consequences of uncontrolled epilepsy can be quite severe. Unfortunately, the patient cannot carry out any self-directed activities until the exhaustion of their energy during epilepsy seizures, which seriously affects the patients' life. Besides, patients with epilepsy carry a risk of sudden unexpected death (SUDEP) 20 times greater than the rest of the population. SUDEP often happens when a person is asleep. Based on three key features where the seizure begins, level of awareness during a seizure, and level of other symptoms such as movement or auras, seizures can be classified to focal, generalized, and unknown onset. Focal or focus seizures start in and affect part of one cerebral hemisphere. Generalized seizures affect both sides of the brain, where they can begin as focal and then become generalized. The consequence of this type of seizure can be severe; loss of consciousness, falls, severe muscle contractions. It should be noted that the presence of generalized seizures increases the risk factor for SUDEP. Therefore, an accurate and timely diagnosis for epileptic seizure plays a crucial role in improving the quality of life of epileptic patients and preventing it from death.

As mentioned, SUDEP often occurs during sleep; therefore, sleep staging plays a crucial role in saving lives in patients with epilepsy. Besides, sleep is associated with learning and memory processes. Humans spend a third of their life in sleep. A third of the US population experiences less than the recommended amount of sleep, which is linked to many chronic diseases and conditions, such as type 2 diabetes, heart disease, obesity, and depression [2]. As sleep pathologies are increasingly recognized as crucial factors in many illnesses, both as effects and causes, and the improved availability of low-cost sleep monitoring devices continues to accelerate the field, the volume of data continues to expand. Therefore, the need for automated sleep staging and diagnostics is more acute, particularly in low resource regions of the world.

EEG is an electrophysiological monitoring method to record the electrical activity of the brain. It is typically noninvasive, with the electrodes placed along the scalp, on specific areas based on 10-20 standard [36], as shown in Fig (2.1). The ground truth for EEG labeling for seizure detection or sleep staging remains the multi-lead EEG and manual labeling by an expert, which is costly and time-consuming. In addition to the time and cost involved in manual EEG labeling, the significant inter-expert variability remains an issue [108]. To build a large enough dataset to make health-AI models work, studies often combine data from multiple hospitals. Therefore, the condition or device used to capture the data can vary from hospital to hospital and even from department to department. Electrode mismatching, inherent inter-subject variability, and the non-stationary nature of EEG signals lead to different joint distributions, $P(X, Y)$ between different recording, where X and Y are feature and label spaces, respectively. Moreover, the generalizability of a model that is trained on a dataset is often low when it is tested on another dataset acquired in different environments with different acquisition hardware. Class imbalances between hospitals can then be associated with hardware differences (such as filter cut-offs). While this can be mitigated through careful inspection of the data, electrode placement difference, and patient physiological differences are harder to identify and reduce. Therefore, the transferability of the trained model for an application on unseen subjects is degraded. The reason behind this problem is that the primary assumption in machine learning techniques is that training and test data should be drawn from the same distribution. This assumption does not necessarily hold in large biomedical datasets. Data from two hospitals that are recorded with different devices and set-ups, but for the same task, cannot necessarily be leveraged directly in a machine learning approach. The main question raised is how to boost performance in the real-life application of EEGs through a generalized model across a large population. This issue could be interpreted as how one can diminish spatial and temporal shifts across individuals from different hospitals or recording environments to handle these different variabilities.

This thesis contributes to addressing these challenges and developing models with high generalizability, which are robust across the population.

1.2 Aim of this thesis

This thesis aims to provide generalizable methods for dealing with different variabilities in biomedical signals, such as intra- and inter-subject variabilities. To this end, signal processing and machine learning techniques were used to increase the generalizability of a model across the population. To achieve our final aim, the following novel research was performed:

- A method to cluster individuals and find groups of similar patients in a non-Euclidean space and measure the boosting effect of machine learning and deep learning methods to mitigate the impact of intra- and inter-subject variability in seizure detection and sleep staging tasks.

- A method to jointly learn patient-invariant representations and weight features to enhance the contribution of relevant features in the final model and decrease the impact of irrelevant features using adversarial networks. We evaluate the effectiveness of this method on the sleep staging task using the largest publicly available sleep dataset at this is time.

- A method based on adversarial training and attention mechanisms to extract transferable information across individuals from different datasets and pay attention to more important or relevant channels and transferable parts of data simultaneously. We evaluate the effectiveness of this method on the sleep staging task using two large datasets.

- A novel method to generate transferable features to fill in the gap between features from the training and test sets and adversarially trains the deep classifiers to make consistent predictions over the transferable features using adversarial networks. Experiments on EEG seizure databases show the effectiveness of the proposed method.

1.3 Thesis outline

The thesis comprises five chapters besides the introduction, all of which (except for the conclusion) have been published or are under review in key journals and conferences in the field (see section 1.4).

Chapter 2 presents our proposed method to learn relevant individuals based on their similarities effectively. The proposed method embeds all training patients into a shared and robust feature space. Individuals who share strong statistical relationships and are similarly based on their EEG signals are clustered in this feature space before being passed to a deep learning framework for sleep staging classification.

Chapters 3 proposes a generalizable method for cross-subject sleep staging based on adversarial training along with attention mechanisms to extract transferable information across individuals from different datasets and pay attention to more important or relevant channels and transferable parts of data, simultaneously. The method has been inspired by how clinicians manually label sleep stages.

Chapter 4 provides this concept that 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. Then, this chapter proposes a method to jointly learn patient-invariant representations and weight 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.

Chapter 5 provides a novel method to generates transferable features to fill in the gap between features from the training and test sets and adversarially trains the deep classifiers to make consistent predictions over the transferable features. By deceiving both classifier and domain discriminator, the proposed method generates transferable features that fill the gap between subjects' joint distributions.

Finally, Chapter 6 presents a summary of contributions, limitations, and possible future work.

1.4 List of publications

Work in this thesis has been published in the following journals and conference:

- S. Nasiri, G. D. Clifford, "Boosting Automated Sleep Stage Performance in Big Datasets using Population Sub-grouping", Sleep, Under review.
  (This publication appears in Chapter 2).

- S. Nasiri, G. D. Clifford, "Subject Selection on Riemannian Manifold for cross-subject classification", Neural Information Processing Systems (NeurIPS) - Machine Learning for Health, Long Beach, USA, 2017 Dec 1.
  (This publication appears in Chapter 2).

- S. Nasiri, G. D. Clifford, "Attentive Adversarial Network for Large-Scale Sleep

Staging", The Journal of Machine Learning Research (JMLR), 2020 June 20.
(This publication appears in Chapter 3).

- S. Nasiri, G. D. Clifford, "Importance Weighting with Adversarial Network for Large-Scale Sleep Staging", International Conference on Machine Learning (ICML) - Lifelong Learning, 2020 June 25.
  (This publication appears in its entirety in Chapter 4).

- S. Nasiri, G. D. Clifford, " Cross-Subject Seizure Detection using Generating Transferable Adversarial Features", Signal Processing Letter, Under review.
  (This publication appears in its entirety in Chapter 5).

Chapter 2

Boosting Automated Sleep Stage and Seizure detection Performance in Big Datasets using Population Sub-grouping

2.1 Abstract

Current approaches to automated sleep staging and seizure detection from the electroencephalogram (EEG) rely on constructing a large labeled training and test corpora by aggregating data from different individuals. However, many of the subjects in the training set may exhibit changes in the EEG that are very different from the subjects in the test set. Training an algorithm on such data without accounting for this diversity can cause underperformance. Moreover, test data may have unexpected sensor misplacement or different instrument noise and spectral responses. This work proposes a novel method to learn relevant individuals based on their similarities effectively. The proposed method embeds all training patients into a shared and robust

feature space. Individuals that share strong statistical relationships and are similar based on their EEG signals are clustered in this feature space before being passed to a deep learning framework for classification. For sleep staging task, using 994 patient EEGs from the 2018 PhysioNet Challenge (6,561 hours of recording), we demonstrate that the clustering approach significantly boosts performance compared to state-of-the-art deep learning approaches. The proposed method improves, on average, a precision score from 0.72 to 0.81, a sensitivity score from 0.74 to 0.82, and a Cohen's Kappa coefficient from 0.64 to 0.75 under 10-fold cross-validation. For seizure detection task, experiment on an EEG seizure database, CHB-MIT database [32] shows that the proposed method increases the accuracy over state-of-the-art from 86.83% to 89.84% and specificity from 87.38% to 89.64% while reducing the false positive rate/hour from 0.8/hour to 0.77/hour.

2.2 Introduction

As sleep pathologies are increasingly recognized as important factors in many illnesses, both as effects and causes, and the improved availability of low-cost sleep monitoring devices continues to accelerate the field, the volume of data continues to expand. The need for automated sleep staging and diagnostics is therefore more acute, particularly in low resource parts of the world. The standard 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') and manual labeling by a sleep expert into five stages: Wake (W), Non-REM 1 (N1), Non-REM 2 (N2), Non-REM 3 (N3) and REM [9]. In addition to the time and cost involved, the significant inter-expert variability remains an issue [108]. However, the lack of large public database with heterogenous populations, has limited the development of verifiable algorithms which generalize well across the populations. Recently, multiple authors have focused on developing an automated sleep scoring based on applying deep learning methods on different biomedical times series such as electrooculogram (EOG), and electromyogram (EMG), and electrocardiography (ECG) [10, 58, 89, 71]. Although, methods based on deep learning and signal process methods have shown promising results, the generalization of a method across different patients is still the main challenge in most clinical applications. It is well-known that brain signals are subject-specific, thus sleep stage algorithms are sometimes tuned to the individual: training and test data belong to the same subject [12, 109]. However, such an approach is inconvenient, expensive, and time-consuming to obtain a large number of training samples to train an associated classifier for every subject. To address this issue, subsets of past subjects can be used to initialize a classifier for training on a new subject [109]. This form of learning is referred to as cross-subject learning [109]. However, this method is limited, due to inter-subject variability and possibly significant differences between the current subject and past training data.

Traditionally, in the EEG community, studies tried to build subject-dependent frequency and spatial filters discriminating between EEG datasets corresponding to two different classes. Such spatial filters perform linear combinations of the EEG signals in order to create new signals with maximal variance in one condition and minimal variance in the other condition. Once these spatial filters have been designed, the (log-)variance of the spatially-filtered signals is used as features by a supervised classification algorithm. It is known that the covariance matrix can be a measure of how much the data is spread across the feature space. The spatial covariance matrices capture information on the spatial inter-dependencies of the brain regions. In the literature, it has been shown that Symmetric Positive Definite (SPD) matrices provide the strong ability to represent the brain signals. In other words, covariance matrices can capture spatial-temporal dynamics which is a measure of functional connectivity between different regions of the brain. Functional connectivity is defined in terms of statistical dependencies among neurophysiological measurements. If we assume these measurements conform to Gaussian assumptions then we need only characterize their correlations or covariance (correlations are normalized covariances). A set of all $C \times C$ SPD matrices form a non-Euclidean space, which is called a Riemannian manifold [5]. Covariance matrix is one of typical examples of SPD matrices, which is employed by several works [78, 75, 50] to present the second-order statistics of the set of signals to reduce inter-subject and intra-subject variabilities. Using tools from Riemannian geometry allows us to apply some methods on the manifold. In fact, using this approach aims to merge the spatial filtering and the classification procedure into one unique step. Jiang et al [45] used the spatial covariance matrices as features and showed that using this form of features improved the sleep staging performance on the MASS dataset. Clustering is an unsupervised learning method to group points, in such a way, points in a cluster are similar to each other, and less similar or dissimilar to points in other clusters. The number of clusters can be set manually or found by some criteria. The clustering technique tries to give an impression of data by identifying groups of similar behavior in the data. We use this concept to find groups of patients that EEG patterns of patients in a cluster behave similarly. Generally, there are two approaches for clustering; compactness and connectivity. In the compactness approach, points that are close to each other, based on the distance between them, lie in a cluster. K-means clustering belongs to this type of clustering approach. On the other hand, in a connectivity-based approach, which roots in graph theory, points are considered as nodes of a graph and connected points are put in the same cluster. Spectral clustering is a technique that follows this approach.

Traditionally, in the EEG community, studies tried to build subject-dependent frequency and spatial filters discriminating between EEG datasets corresponding to two different classes. Such spatial filters perform linear combinations of the EEG signals in order to create new signals with maximal variance in one condition and minimal variance in the other condition. Once these spatial filters have been designed, the (log-)variance of the spatially-filtered signals is used as features by a supervised classification algorithm. It is known that the covariance matrix can be a measure of how much the data is spread across the feature space. The spatial covariance matrices capture information on the spatial inter-dependencies of the brain regions. In the literature for BCI, it has been shown that EEG covariance matrix classification is much more robust than many conventional features [12, 109, 5, 20]]. In other words, covariance matrices can capture spatial-temporal dynamics which is a measure of functional connectivity between different regions of the brain. Functional connectivity is defined in terms of statistical dependencies among neurophysiological measurements. If we assume these measurements conform to Gaussian assumptions then we need only characterize their correlations or covariance (correlations are normalized covariances). A set of all C C SPD matrices form a non-Euclidean space, which is called a Riemannian manifold [9]. Covariance matrix is one of typical examples of SPD matrices, which is employed by several works [78, 75, 50] to present the second-order statistics of the set of signals to reduce inter-subject and intra-subject variabilities. Using tools from Riemannian geometry allows us to apply some methods on the manifold. In fact, using this approach aims to merge the spatial filtering and the classification procedure into one unique step. Jiang et al [45] used the spatial covariance matrices as features and showed that using this form of features improved the sleep staging performance on the MASS dataset.

Clustering is an unsupervised learning method to group points, in such a way, points in a cluster are similar to each other, and less similar or dissimilar to points in other clusters. The number of clusters can be set manually or found by some criteria. The clustering technique tries to give an impression of data by identifying groups of similar behavior in the data. We use this concept to find groups of patients that EEG patterns of patients in a cluster behave similarly. Generally, there are two approaches for clustering; compactness and connectivity. In the compactness approach, points that are close to each other, based on the distance between them, lie in a cluster. K-means clustering belongs to this type of clustering approach. On the other hand, in a connectivity-based approach, which roots in graph theory, points are considered as nodes of a graph and connected points are put in the same cluster. Spectral clustering is a technique that follows this approach.

To handle inter-subject variability problem, we use the spectral clustering method on a Riemannian manifold, which groups subjects based on the similarity of the features. We hypothesized that, by sub-grouping of the population on a robust feature space, the diversity between EEG patterns across subjects can be reduced and this leads to an improved sleep staging performance. Barachant et al. [5, 20] showed that appropriate forms of data covariance matrices provided superior cross-subject generalization capabilities compared to earlier works, as well as robustness to EEG artifacts, outliers, mislabeling, and sensors misplacement [109, 12]. In this work, the problem of boosting automated sleep stage performance in large dataset using population sub-grouping is investigated without assuming knowledge of labels from population. The main contribution of this work is to develop a population sub-grouping algorithm on a robust feature space. The inherent diversity between EEG signals across different individuals poses a challenge for the brain signal classification. Therefore, it is essential to identify subsets of subjects that most closely similar to each other. To obtain subsets of similar subjects in terms of EEG features, we apply spectral clustering on a Riemannian manifold. Subsequently, a Convolutional Neural Network (CNN) algorithm is applied in the tangent space at geometric mean of covariance matrices. Finally, the algorithm's performance is assessed on the largest open-access database – the 2018 PhysioNet Challenge database [30, 32]. All data used in the study are publicly available from PhysioNet.org, and therefore no ethics/institutional review board approval was required.

2.3 Method

2.3.1 Data

For evaluation of the proposed method, EEG data from the 2018 PhysioNet Challenge [30, 32] is used. Polysomnogram (PSG) recordings were collected in the Massachusetts General Hospital's (MGH) by the Computational Clinical Neurophysiology Laboratory and the Clinical Data Animation Laboratory. A total of 994 subjects are available, together with the demographic data and list of active treatment drugs (including antihypertensives, antidepressants, neuroactive compounds, insulin for diabetes, and sleep aids). Clinical characteristics of these patients are reported in Table 1. The goal of the 2018 PhysioNet Challenge was to develop an algorithm to detect non-apnea sleep arousals. However, the main focus of this work is to develop methods to boost the performance of sleep stages classification tasks by sub-grouping of populations. Based on the American Academy of Sleep Medicine (AASM) standard [9], every 30-second window of data is annotated as wakefulness, rapid eye movement (REM), stage 1, stage 2 or stage 3. PSG recording includes thirteen signals including electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), airflow and oxygen saturation (SaO2), and ECG signals, where all data were sampled at 200 Hz. In this work, we just use EEG data from PSG measurement which are obtained from six channels, 'F3-M2', 'F4-M1', 'C3-M2', 'C4-M1', 'O1-M2', 'O2-M1' which is shown in Fig (2.1).

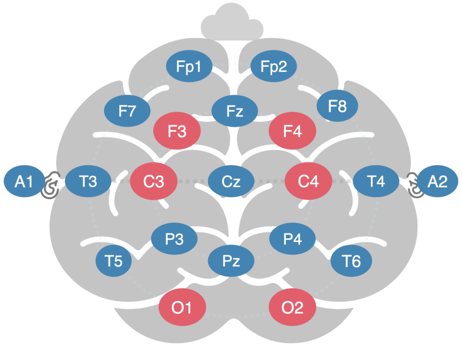


Figure 2.1: 10-20 EEG Placement: Used electrodes are shown in red. “CC BY 4.0”

The input to the proposed method is a sequence of 30-sec EEG window. In order to extract windows from a continuous EEG signal, the following steps are implemented:

- Remove the 60 Hz component form the signal using a notch filter with quality factor $Q = 30$

- For each record, normalize channels by removing the mean and scaling by standard deviation.

- Apply a 5^{th}-order Butterworth 0.5-40 Hz band-pass filter.

- Segment the continuous raw EEG to a sequence of 30-s windows and assigning a label to each window (i.e., sleep stage) based on the annotation file.

- Only consider N1, N2, N3, REM and wake label for sleep staging (ignore arousals).

Most biomedical applications deal with class imbalances, where observations are not equally distributed through the classes. In the sleep stage classification task, the number of wake and N2 stages is greater than for other stages. The distribution of sleep stages in the 2018 PhysioNet Challenge database is not uniform - N2 represents 42% of total observations. To handle this imbalance issue in the dataset, two approaches are used. The first is to give weights to the classes by multiplying the loss of each example. The second is that in the training step, training samples are over-sampled to provide approximately an equal number of sleep stages in each class. To achieve this, we use the synthetic minority over-sampling technique [17] in the training process to generate the synthetic data points by considering the similarities between existing minority samples.

Different metrics are considered in the literature to evaluate the performance of sleep classifiers. Due to imbalanced nature of the dataset, the F1-score, and recall (sensitivity) are reported in addition to overall precision for the epoch-by-epoch sleep stage classification task. These metrics are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TN + TP + FP + FN} \tag{2.1}$$

$$\text{Percision} = \frac{TP}{TP + FP} \tag{2.2}$$

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN} \tag{2.3}$$

$$\text{F1-score} = 2 * \frac{\text{Percision} \times \text{Recall}}{\text{Percision} + \text{Recall}} \tag{2.4}$$

Where TP =True Positive, TN=True Negative, FP=False Positive and FN=False Negative.

The other primary metric that we have used for performance evaluation of our proposed method is Cohen's Kappa coefficient, which measures the agreement between the class obtained by the algorithm and the expert labels.

2.3.2 Sub-grouping of population

An important assumption of machine learning techniques is that the training and test data should be drawn from the same distribution. However, in real-world applications, this assumption rarely holds. Having individuals with different biological demographics such as gender, age, medical conditions, body mass index, and different brain topographies results in different distributions of the features of interest. As a result, although the activity in the brain from different individuals can be similar, the activity in the electrodes may not similar. In fact, neural activity is not propagating in similar way in different subjects. Moreover, cortical folding, tissue conductivity and tissue shapes of brains are different across people [29]. All these factors lead to significant inter-subject variability in EEG signals. In other word, inter-subject variability leads to different joint distribution, P(X,Y) between different individuals (where X and Y are feature and label space, respectively). Therefore, the transferability of the trained model for an application on unseen subjects is degraded. The main challenge in this problem is to diminish spatial and temporal shifts across individuals and thereby develop a generalized model across a large population. The spatial shift in data can be caused by the variation of sensors' location on the brain, which can be partly solved by finding a robust feature space. As mentioned earlier, it has been shown that Symmetric Positive Definite (SPD) matrices provide a strong ability to representations the brain signals [20, 4]. The covariance matrix is a typical example of SPD matrices, which has been employed in several studies [78, 75, 50]. These studies showed that using second-order statistics of multi-channel signals reduce inter-subject and intra-subject variabilities between EEG signals. Spatial covariance matrix can well separate useful information about brain functional connectivity structure [4] and create a feature space which is comparable across subjects. Moreover, it has been shown that SPD matrices have an excellent robustness to the considerable variability of real-world environmental conditions such as instrument noise [20, 76, 75]. To describe the extraction of spatial covariance matrices, we let $X_i \in \mathbb{R}^{C \times N}$ denote a 30-sec window indexed by i, with C as the number of channels, and the number of samples is given by $N = 30 \times f_s$, where f_s is the sampling frequency. The feature covariance matrices are obtained simply by using a Sample Covariance Matrix estimator [4], such as:

$$\Sigma = \frac{1}{N-1} \, X_i \, X_i^T \tag{2.5}$$

where X_i^T is the transpose of X_i. Note that each epoch is zero-mean after the preprocessing step.

SPD matrices belong to a non-Euclidean space, which is called a Riemannian manifold. This manifold has useful mathematical properties, which we hypothesize are beneficial for EEG classification [5, 26, 76, 75].

Riemannian Distance: For any two covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, the Riemannian distance is defined according to the Riemannian metric as [6]

$$\delta_R(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) = \|\log\left(\mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1/2}\right)\|_F = \left[\sum_{c=1}^{C} \log^2 \lambda_c\right]^{1/2} \tag{2.6}$$

where $\lambda_c, c = 1 \ldots C$ are the real eigenvalues of $\mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1/2}$ and C the number of channels. This distance is *Affine-invariant* [6], i.e, it is invariant with respect to similar and congruent transformations, and inversion.

Riemannian Mean: The Riemannian geometric mean of I covariance matrices, also called the Fréchet or Karcher mean, is the point on the manifold minimizing the dispersion given by [6]:

$$\mathfrak{G}\left(\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_I\right) = \arg\min_{\mathbf{\Sigma}} \sum_{i=1}^{I} \delta_R^2\left(\mathbf{\Sigma}, \mathbf{\Sigma}_i\right). \tag{2.7}$$

There is no closed form expression for the mean of $I > 2$. However a gradient descent procedure in the manifold can be used in order to find the solution [7].

Zianai et al [109] have proposed to transform the covariance matrices of every subject in order to center them with respect to a reference covarince matrix, making the data from different subjects comparable. We use this transformation in our work., which is given by

$$\Sigma_i^{(j)} \Rightarrow (\bar{\Sigma}^{(j)})^{-\frac{1}{2}}\Sigma_i^{(j)}(\bar{\Sigma}^{(j)})^{-\frac{1}{2}} \tag{2.8}$$

where $\bar{\Sigma}^{(j)}$ is the center of mass of covarinace matrices for subject j.

In order to find sub-groups of the population in big data, we assume that the dataset comprises several sub-distributions that correspond to different groups in the population. Moreover, we assume that clustering patients will allow identification of sub-groups. We hypothesis that clustering patients in the robust feature space could identify subsets of relevant individuals and diminish inter-subject physiological variabilities. In other words, we assume by clustering the patients, those individuals in a same cluster have similar joint distribution, $P\left(X, Y\right)$. To sub-grouping of population, a spectral clustering is applied by using Riemannian distance and geometric mean on the manifold to find sub-groups of relevant subjects [11]. The clustering is an unsupervised method, i.e. no labeled data is used in this step. We assumed that unlabeled data from all subjects are available. In real-life scenario, if a new subject comes to the dataset, we compute the covariance matrices from his/her dataset, and after mapping to the manifold, we try to find the closest cluster to that patient and use the model associated to that cluster. The framework of clustering is shown in Fig (2.2).

Spectral clustering constructs a similarity graph, $G = (V,\ E), where V = \Sigma_i, \cdots, \Sigma_n$ which represent covariance matrices on the manifold as vertices. $E = \left(w_{ij}\right)|_{i,j=1,\cdots,n}$ is a corresponding set of edges between each pair of Σ_i and Σ_j, where the edge between two vertices is defined as a similarity metric [97, 65]. In this work, we create a fully connected graph and weight edges by the output of Gaussian similarity function, which is defined as follows:

$$w(x_i, x_j) = exp(-\frac{\delta_R^2(\bar{\boldsymbol{\Sigma}}_i, \bar{\boldsymbol{\Sigma}}_j)}{\sigma^2}) \tag{2.9}$$

where the parameter controls the width of the neighborhoods.

Spectral clustering uses an important concept from graph theory to cluster the points; graph Laplacian matrix. Laplacian matrix is a way to represent the graph which has several important properties. Laplacian matrix is constructed by degree matrix and adjacency matrix. Adjacency matrix, $A_{ij} = w_{ij}$, is represent a graph by a matrix, where rows and columns represent the nodes and the entries represent the weight of the edge between row i and column j by w_{ij}. The degree matrix is a diagonal matrix where each element on the diagonal represents the degree of a vertex v_i, the total amount of weights of edges connected to it [97, 65].

$$\begin{aligned} d_i =& \sum_{j=1}^{n} w_{ij} \\ d_i =& \sum_{j=1}^{n} w_{ij} D \ = \begin{bmatrix} d_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & d_n \end{bmatrix} \end{aligned}$$

Therefore, the graph Laplacian matrix is defined as follow:

$$\begin{aligned} L =& D - A \\ f(x) =& \begin{cases} d_i, & \text{if } i = j \\ -w_{ij}, & \text{if } (i,j) \in E \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

In order to cluster the graph, data points should be embedded into a low-dimensional space. To do that, eigenvalues and eigenvectors of normalized Laplacian matrix, $L = A^{(-1/2)}DA^{(1/2)}$, is computed. One of the important properties of the Laplacian matrix is that the number of eigenvalues equal to 0 represent the number of connected components in the graph.

In order to determine the number of clusters in spectral clustering, an eigengap heuristic approach is used. Let $\lambda_1, \cdots, \lambda_n$ be the eigenvalues of the Laplacian, the goal is to choose k such that $\lambda_1, \cdots, \lambda_k$ are relatively small, but $\lambda_{(k+1)}$ is relatively large. After defining the number of clusters (K), the first K eigenvectors $u_1, \cdots, u_k$

of Laplacian matrix are stacked vertically to create a matrix $U \in \Re^{(n \times k)}$. K-means clustering is applied to each row of matrix U to cluster the data to K different clusters [97].

Therefore, after making data from different subjects comparable by Eq (2.8), we use the defined Riemannian distance and geometric mean as the metric for measuring distance and mean in spectral clustering method, respectively.

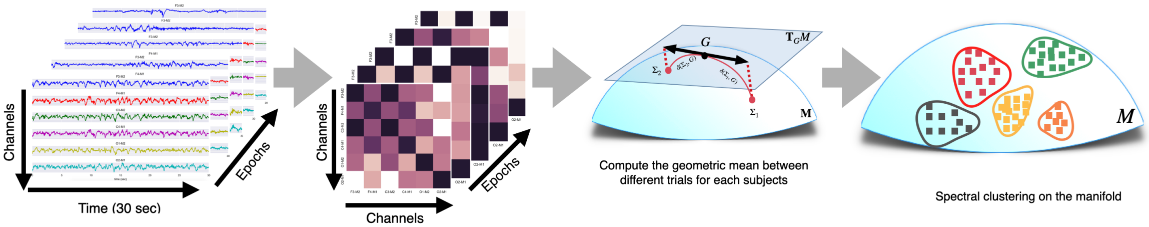

Channels
Time (30 sec)
Epochs
Channels
Channels
Epochs
$T_G M$
G
Σ_2
Σ_1
M
Compute the geometric mean between different trials for each subjects
M
Spectral clustering on the manifold

Figure 2.2: Framework for clustering of covariance matrices on a Riemannian manifold: For each epoch of the EEG signal, the covariance is computed using Eq (2.5). Due to the special structure of covariance matrix (Symmetric Positive Definite), covariance matrices lie on a non-Euclidean space, Riemannian manifold. For each individual, the geometric mean of covariance matrices obtained from different epochs is computed. In order to compute the geometric mean, exponential and logarithmic mapping is performed to map covariance matrices to tangent space, T_M. To take account of their geometry, a spectral clustering is applied on that space using a Riemannian distance and geometric mean. “CC BY 4.0”

2.3.3 Deep Learning Model on a Riemannian Manifold

Due to Euclidean assumptions in many machine learning and deep learning methods, one cannot directly apply these methods to manipulate SPD matrices. Furthermore, ignoring the structure of SPD matrices in the learning framework often leads to detrimental effects. For example, the Euclidean average of SPD matrices suffers from a swelling effect, i.e., the determinant of the average is larger than any of the original determinants [41, 40]. Therefore, it is necessary to use tools from Riemannian manifold to preserve the geometry information of SPD matrices, which can convey valuable information about functional connectivity between different regions of the brain. There are different techniques to manipulating SPD matrices, which can be found in Congedo et al. [20]. One popular method is to map data to local Euclidean space and apply conventional machine learning methods to learn patterns of SPD matrices in this new space. In the manifold, a tangent space, $T_G M$, is a local approximation of the manifold at point G. Each point on the manifold is flattened and mapped to the tangent space. In this work, for each cluster, covariance matrices are flattened at the center of the mass of the data in the cluster. Vectorizing of covariance matrices is achieved by projecting every element from the manifold to the tangent space at the Riemannian mean of that cluster. Deep neural networks have been successfully applied for learning general and transferable features in big data areas. Deep learning discovers the hidden structures and high-level abstract concepts in large data without any prior handcraft feature selection. Deep networks have delivered promising results in several machine learning problems and applications, such as image and speech recognition. Convolutional Neural Networks (CNNs) have received more attention in the field of brain signal processing, such as for seizure detection [113], sleep staging [62], Parkinson's disease [85]. It can be shown that these networks extract first-order information from input data by performing linear combinations and element-wise nonlinear operations. Moreover, second-order statistics computed from handcrafted features, e.g., covariances, have proven highly effective in diverse brain signal classification tasks. Therefore, in order to leverage these two properties, we train a deep neural network for each cluster on the tangent space at the geometric mean of that cluster. Fig (2.3) illustrates the framework of the proposed method on the tangent space. The architecture of the CNN is the same of the architecture in [25]. The network is trained on the cross-entropy objective function for 1600 epochs using Adam optimizer with a batch size of 21 samples [47], and learning rate $\eta = 10^{-6}$. A major channel in training a deep neural network is the trade-off between training time and its performance. It is known that a little number of iterations for training a deep neural network leads to underfitting of the network, and many iterations of the training may lead to overfitting of the training model, then the generalization error will increase and model will have poor performance on the validation data. A simple and effective method to train a deep neural network and prevent overfitting is the early stopping, which is considered as a regularization method.

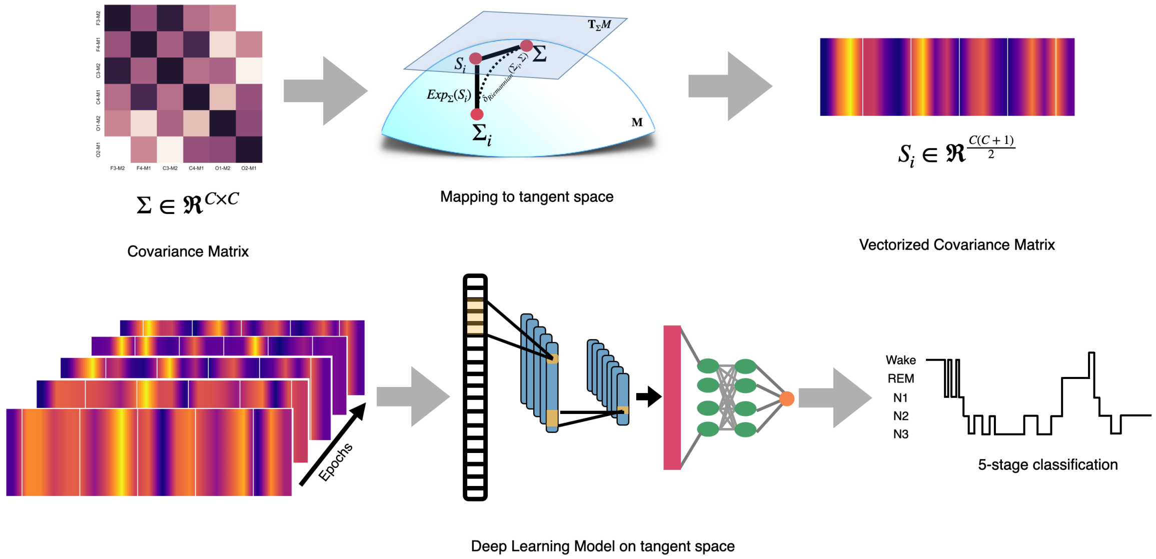


Figure 2.3: Framework of the learning method: After segmenting the data, extracting covariance matrices, sub-grouping the patients by clustering them, covariance matrices in each cluster are mapped to the tangent space at the geometric mean obtained from the covariance matrices. This mapping is performed, since the tangent space is considered as a Euclidean space and therefore Euclidean-based machine learning can be used. The vectorized covariance matrix on tangent space is shown in such way that the x-axis shows the element in the vector between $[x_{min}, x_{max}]$, and the y-axis shows the number of elements in each range. Then, for each cluster, a CNN model is trained on the tangent space to classify sleep stages into 5 classes (Wake, N1, N2, N3, REM). “CC BY 4.0”

In this work, we use this technique after a reasonable number of training iterations (where we set 1500 iterations). A major challenge in training a deep neural network is the trade-off between training time and its performance. It is known that a little number of iterations for training a deep neural network leads to underfitting of the network, and many iterations of the training may lead to overfitting of the training model, then the generalization error will increase and model will have poor performance on the validation data. A simple and effective method to train a deep neural network and prevent overfitting is the early stopping, which is considered as a regularization method. In this work, we use this technique after a reasonable number of training iterations (where we set 1500 iterations). The procedure was repeated through 10-fold cross-validation. The trained model was tested on patients that were not included in the training step. Therefore, we can evaluate how the proposed model generalized across patients in a big dataset. It is known that healthcare data suffer from noisy labeling issues due to the inherent difficulty in manually labeling healthcare data, including sleep dataset. Manually annotated epochs according to AASM criteria may contain a large section of another sleep stage (up to 14 seconds), which makes it difficult to interpret the label of an epoch with certainty. One approach to handling this issue is to assume transition epochs (one which does not have neighboring sleep stages of the same class) are more likely to be partially composed of another sleep stage. If we remove these transition epochs, it is possible to avoid this noisy labeling. Therefore, before doing clustering and training a model, we remove the transition epochs and run the experiment again.

To demonstrate how covariance matrices are robust to electrode misplacement, 497 patients (50% of the total population) are chosen and the average between EEG signals of C3 and O1 and between O2 and C4 are taken to create two new channels, as shown in Fig (2.4). This simulates channel misplacement at the geometric mean of the two electrode locations - instead of having the actual signal from O1 and O2, we have the signals from somewhere between (O1, C3) and (O2, C4) for half of the population.

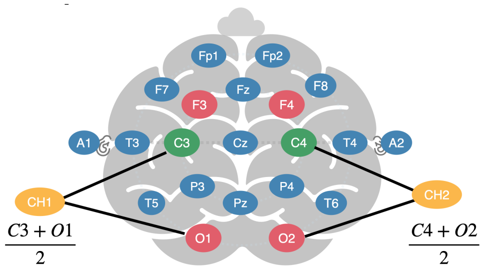


Figure 2.4: Defining two new channels by taking the average between two neighbors' channel (O2, C4) and (O1 and C3) to show how the proposed method are robust to electrode's misplacement. We did that for randomly selected 497 patients (50% of the total population). "CC BY 4.0"

2.4 Results

The goal of this part is to show the comparison of the CNN performance on a Riemannian manifold before and after sub-grouping of the population. As discussed in the supplementary material, the optimal number of clusters was estimated via Eigengap heuristic [97] method. In our dataset, the number of clusters was found to be five. The result of the clustering of subjects is shown in Fig (2.5), derived through a mapping into two dimensions using the t-Distributed Stochastic Neighbor Embedding (t-SNE) approach [57] with a Riemannian distance. t-SNE is a common visualization technique for reducing the dimensionality of data or model. Since it preserves local structure and points which are close to one another in the high-dimensional data set will tend to be close to one another in the projection, the separation of data we can observe is likely to be true in the original (higher) dimensional representation. For each subject, the covariance matrix is computed from each 30-sec window using Eq (2.5), then the geometric mean is computed across all covariance matrices using Riemannian geometric mean (Eq (2.7)). Therefore, for each subject we have a geometric mean covariance matrix $\bar{\Sigma}_j$ with $C \times C$ dimension. t-SNE projection is done on geometric mean covariance matrices from all 994 subjects. Each point in the t-SNE project represents the geometric mean of a subject.

The biomedical factors such as age, gender, BMI, AHI for each subject are not publicly available. Therefore, we cannot claim that these subjects are clustered in one group due to these factors. However, it is possible patients in the same group have similar dynamical functional connectivity. In addition, it is known that functional connectivity is linked to structural connectivity which captures information on anatomical pathways [38]. In other words, the brain network can be modeled as a graph, where nodes are channels on the scalp, and edges between nodes should be defined by functional, structural, and causality. Having different groups of patients based on their spatial covariance matrices can be interpreted that patients in the

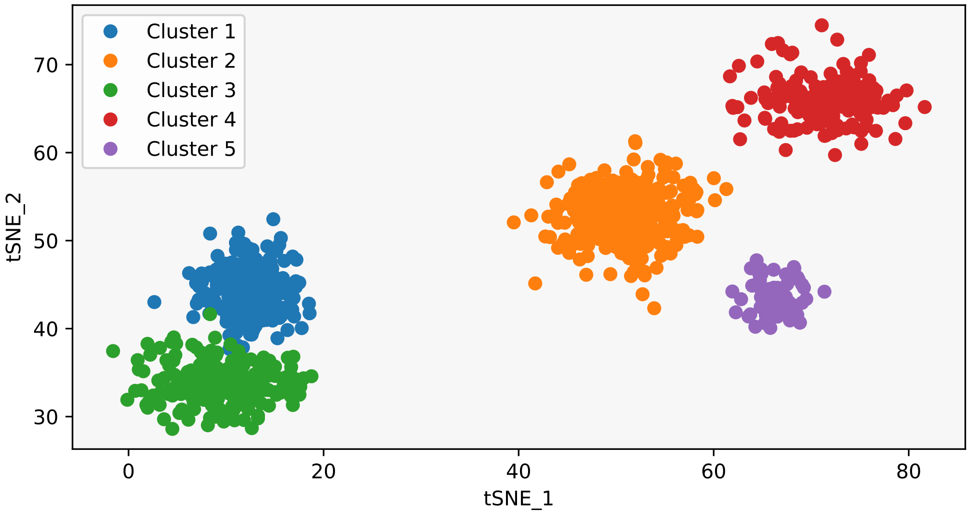


Figure 2.5: t-Distributed Stochastic Neighbor Embedding (t-SNE) visualization of patient clustering on a Riemannian manifold using Riemannian distance and mean. Each point is associated to a subject. “CC BY 4.0”

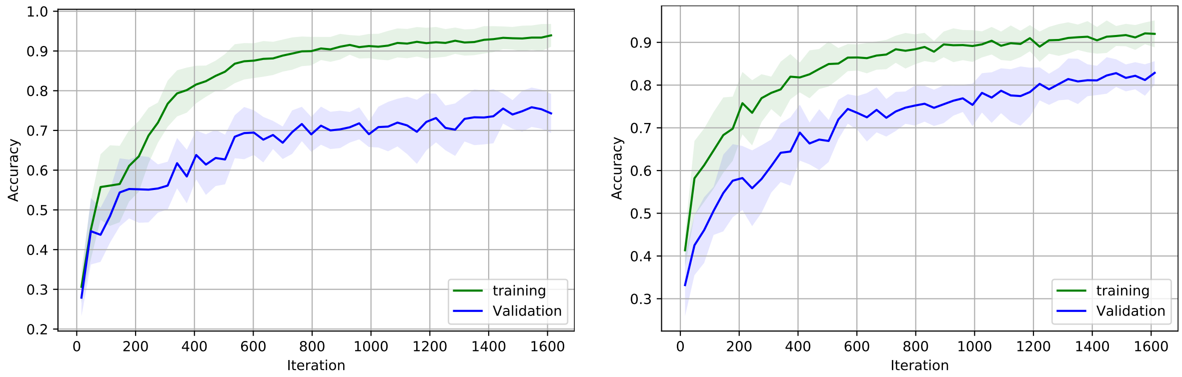
same group have similar brain anatomy and functional connectivity.

In order to increase the generalization capability of the algorithm, the learning model is evaluated using cross-validation. After the data are clustered, 10-fold cross-validation is performed to split data into training, validation for each cluster. For each cluster, a CNN model was trained using data from the training set, evaluated on the validation set to prevent overfitting, and tested on the test set, therefore in this experiment, five CNN models are trained. Any given subject's recordings belong exclusively to training, validation or test sets. Similar to approach of Howe-Patterson et al. [37], in total, 794 subjects are used for training, 100 subjects for validation, and 100 subjects for testing. Table 4.1 provides the number of training, validation, test, and the number of samples in each class for clusters.

Table 2.1: Number of subjects and samples per class for each cluster

Cluster ID	# Subjects	# training	# validation	# test	# Wake	# N1	# N2	# N3	# REM
1	248	198	25	25	37237	36151	94900	22926	27833
2	306	216	45	45	45931	40298	112736	33214	35598
3	211	181	15	15	29940	29096	76732	21092	24487
4	164	144	10	10	23203	20622	63992	17137	18918
5	65	55	5	5	9247	9248	23897	7309	7036

Fig (2.6) shows how the trained model performs over epochs and how it generalizes on the validation set. The figure illustrates the average and standard deviation across 10 folds. Fig 2.6 (a) shows the learning curve of training a deep neural network on a tangent space of a Riemannian manifold without the clustering procedure. As mentioned earlier, in order to prevent overfitting, early stopping is used in this work. We stopped the training of the model at 1600 iteration, where the generalizability of the model stops, and the performance on validation set starts to degrade. To have a fair comparison regarding the generalization error, we stopped the training of the model at 1600 iteration after the clustering procedure was done as well. Fig 2.6 (b) shows the learning curve of training a model after clustering the subjects where it shows that the models have more capacity due to the homogeneity of data.

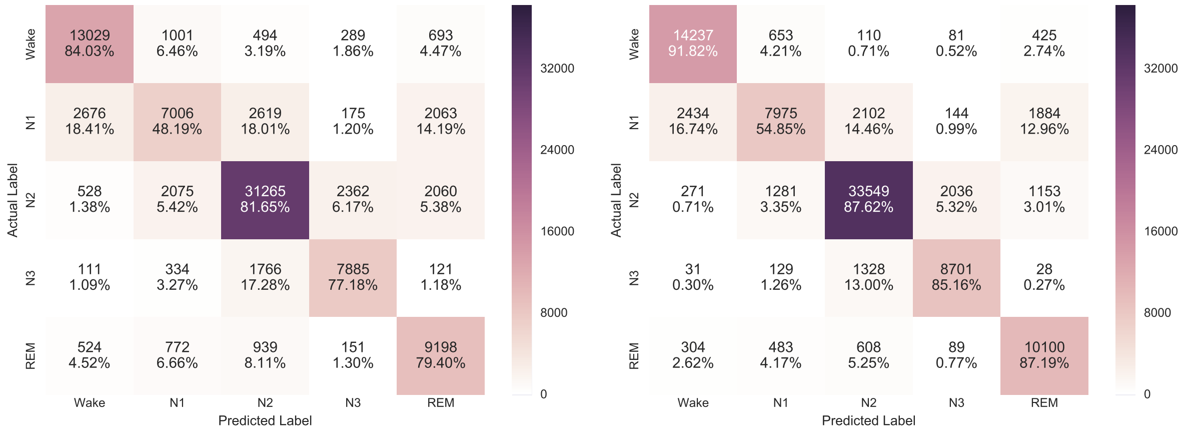


(a) Before Clustering

(b) After Clustering : (+ average over 5-clusters)

Figure 2.6: Learning curve across 10 folds: it shows how the trained model performs over epochs and how it generalizes on the validation set. It is important to show how trained model works on validation set to avoid from overfitting. The figure illustrates the average and standard deviation of accuracy across 10 folds. “CC BY 4.0”

Fig (2.7) shows the confusion matrix obtained before/after applying the trained model on test subjects in each cluster, and then taking the average across clusters. Table 2.3 and 2.2 present the performance of sleep staging before and after applying the proposed method on each class, respectively. The mean 5-class precision, sensitivity and kappa are 0.72, 0.74 and 0.65, respectively. Table 3 presents per-class performance achieved by sub-grouping the population. The mean 5-class precision, sensitivity, and kappa are 0.81, 0.82, and 0.74.



(a) Before Clustering

(b) After Clustering : (+ average over 5-clusters)

Figure 2.7: Five-stage classification confusion matrices: comparing actual label which is obtained by sleep technicians vs. the proposed method on held-out 100 test patients. "CC BY 4.0"

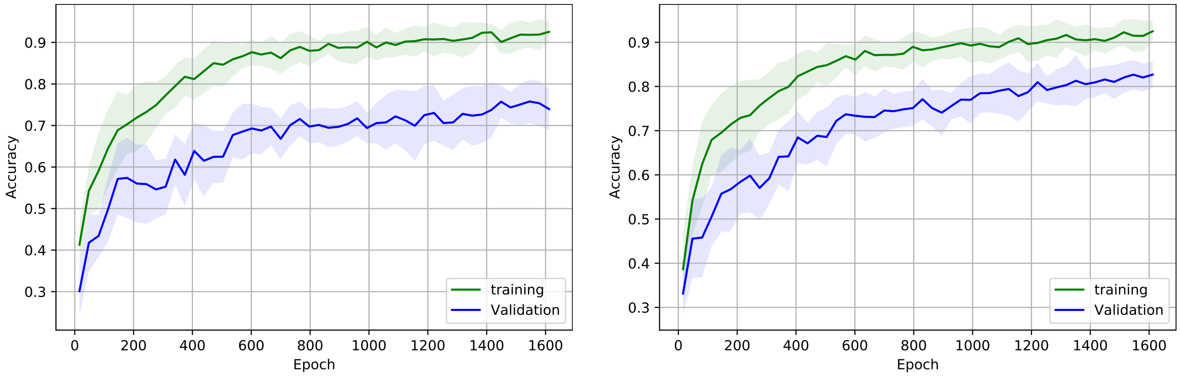
Table 2.2: Per-class performance achieved **before** sub-grouping of population

Sleep Stages	Precision	Sensitivity	F1-Score	Kappa	Number of Samples
Wake	0.77	0.84	0.80	0.75	15506
N1	0.62	0.48	0.54	0.45	14539
N2	0.84	0.81	0.82	0.65	38290
N3	0.72	0.77	0.74	0.70	10217
REM	0.65	0.79	0.71	0.65	11584

Table 2.3: Per-class performance achieved **after** sub-grouping of population

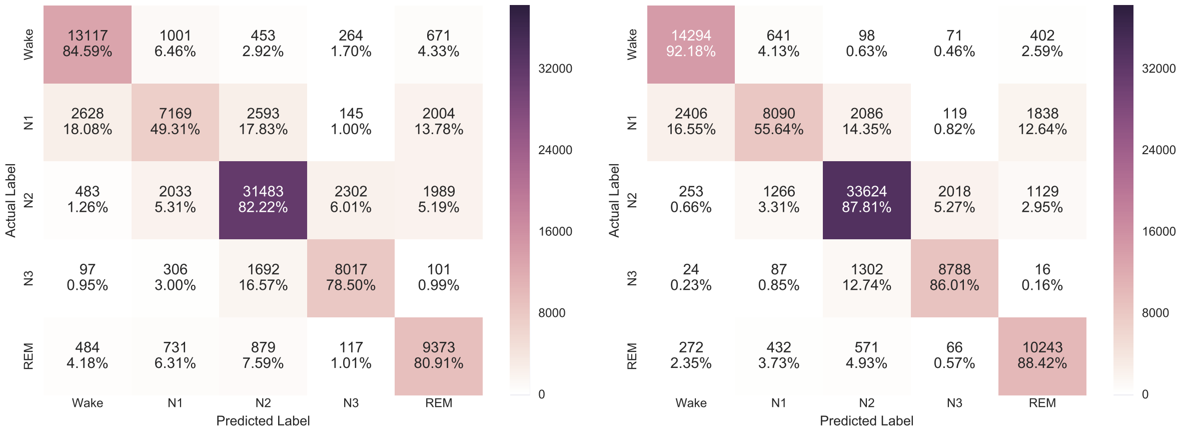
Sleep Stages	Precision	Sensitivity	F1-Score	Kappa	Number of Samples
Wake	0.82	0.92	0.78	0.83	15506
N1	0.76	0.55	0.64	0.58	14539
N2	0.89	0.88	0.88	0.78	38290
N3	0.79	0.85	0.81	0.79	10217
REM	0.74	0.87	0.80	0.76	11584

We note that the model performance for wake, N2, N3 and REM stages is better than for the N1 stage. This issue is related to low human inter-scorer agreement of N1 stage. This may be because N1 is a transition stage between being awake and fully asleep, thus the EEG pattern in this stage is similar to wakefulness and sleep. Colten et al. [3] 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 Fig (2.7). In addition, Basner et al [35] discussed that there is a positive correlation between kappa values and the amount of time spent in a sleep stage. N1 stage is scored when EEG theta activity predominates over alpha. Stage N1 has low-voltage, mixed-frequency background, possibly slow eye movements, and vertex sharp waves. Difficulties of scoring a sleep stage arise when the EEGs contain both awake and N1 patterns and it is difficult to decide which pattern occupies ¿ 15 sec of a 30-sec epoch. Moreover, distinction between N1 and N2 is based on identification of spindles and K complexes. In fact, the reliability of sleep staging may be improved for N1 stage by improving the reliability of spindle scoring for discriminating N1 and N2 stages [21]. However, definition of K complexes is qualitative with no criteria for minimum amplitude or for the duration of the complex's different phases [38]. In addition, manual scoring of spindles is subject to much inconsistency among scorers. Based on AASM standard, the presence of low amplitude, mixed frequency EEG is characteristic of both N1 and REM stages, which discriminating N1 from REM makes it challenging. Furthermore, REM sleep may be blunted in patients with apnea, since the interrupted breathing patterns lead to sleep fragmentation. Therefore, N1 stages will be often be confused with REM. In addition, the number of N1 stage is much smaller than other stages, leading to poorer generalization. Consequently, we observe a lower true positive rate (sensitivity) for N1 classification. In terms of evaluation metrics common to sleep staging, the proposed method outperforms the state-of-art algorithms on the 2018 PhysioNet Challenge dataset with F1=0.80 and κ = 0.75. Specifically, Perslev et al. [71] obtained F1=0.77 under 5-fold cross-validation on the same dataset, which is inferior to the proposed method with population sub-grouping. As explained earlier, healthcare datasets suffer from noisy labeling issues. In this work, we removed transition epochs (one which do not have neighboring sleep stages of the same class) to combat this problem. The learning curves and clustering results do not change noticeable, which shows that the clustering step and training based on the covariance matrices are more robust to noisy labeling problems. Moreover, since we took average covariance matrices for each subject and then mapped each covariance matrix on a tangent space on the geometric mean of each cluster, the overall performance does not change with removing the transition epochs. Fig (2.8) shows the learning curve before/after clustering step when we removed the transition epochs which are more vulnerable to noisy labeling. The performance on test sets are presented in Fig (2.9) and Table 2.4 and 2.5, which show a slight improvement (1-2% improvement) in comparison to the results of the proposed method with transition epochs.



(a) Before Clustering (b) After Clustering : (+ average over 5-clusters)

Figure 2.8: Learning curve across 10 folds after removing transition epochs which are more vulnerable to noisy labeling issue: it shows how the trained model performs over epochs and how it generalizes on the validation set. It is important to show how trained model works on validation set to avoid from overfitting. The figure illustrates the average and standard deviation of accuracy across 10 folds. “CC BY 4.0”



(a) Before Clustering

(b) After Clustering : (+ average over 5-clusters)

Figure 2.9: Five-stage classification confusion matrices: comparing actual label which is obtained by sleep technicians vs. the proposed method on held-out 100 test patients after removing the transition epochs in training step. “CC BY 4.0”

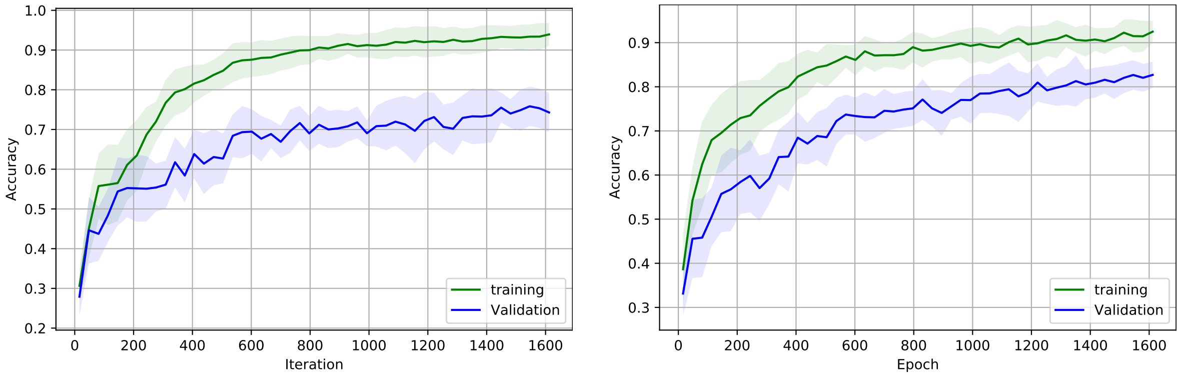
Table 2.4: Per-class performance achieved **before** sub-grouping of population and after removing transition epochs

Sleep Stages	Precision	Sensitivity	F1-Score	Kappa	Number of Samples
Wake	0.78	0.84	0.81	0.76	15506
N1	0.63	0.49	0.55	0.47	14539
N2	0.84	0.82	0.83	0.69	38290
N3	0.73	0.78	0.76	0.72	10217
REM	0.66	0.80	0.72	0.67	11584

Table 2.5: Per-class performance achieved **after** sub-grouping of population and removing transition epochs

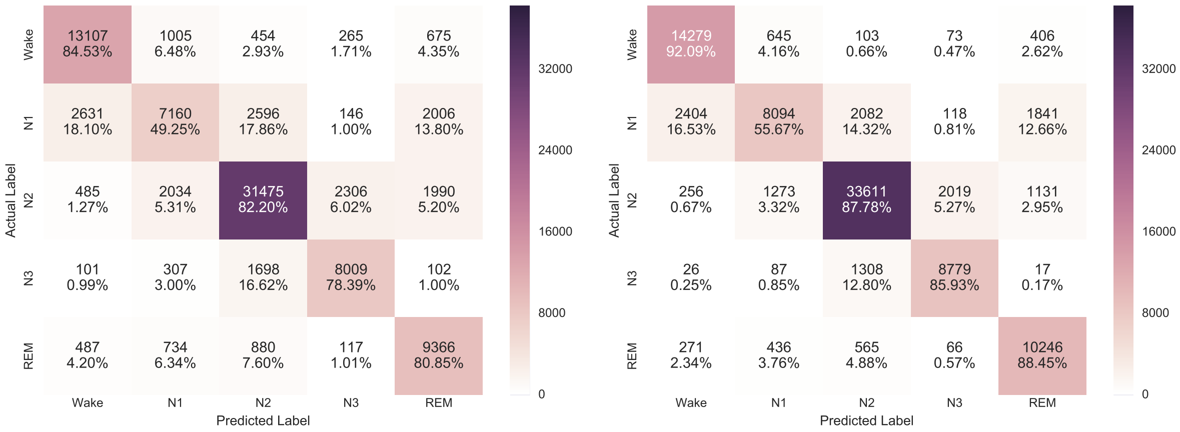
Sleep Stages	Precision	Sensitivity	F1-Score	Kappa	Number of Samples
Wake	0.82	0.92	0.78	0.83	15506
N1	0.76	0.55	0.64	0.58	14539
N2	0.89	0.88	0.88	0.78	38290
N3	0.79	0.86	0.82	0.79	10217
REM	0.75	0.88	0.81	0.77	11584

As mentioned earlier, to show the robustness of covariance features to electrodes' misplacement, for half of the population we replaced O1 and O2 channels with the average signals between (O1,C3) and (O2,C4), respectively. For this case, the learning curves and confusion matrices before and after applying the proposed method are shown in Fig (2.10) and (2.11), respectively. Table 2.6 and 2.6 show that the proposed method still improves the classification performance for sleep staging by 7-8%.



(a) Before Clustering (b) After Clustering : (+ average over 5-clusters)

Figure 2.10: Learning curve across 10 folds after removing transition epochs and defining two new channels instead of O1 and O2: it shows how the trained model performs over epochs and how it generalizes on the validation set. It is important to show how trained model works on validation set to avoid from overfitting. The figure illustrates the average and standard deviation of accuracy across 10 folds. “CC BY 4.0”



Actual Label \ Predicted Label	Wake	N1	N2	N3	REM
Wake	13107 84.53%	1005 6.48%	454 2.93%	265 1.71%	675 4.35%
N1	2631 18.10%	7160 49.25%	2596 17.86%	146 1.00%	2006 13.80%
N2	485 1.27%	2034 5.31%	31475 82.20%	2306 6.02%	1990 5.20%
N3	101 0.99%	307 3.00%	1698 16.62%	8009 78.39%	102 1.00%
REM	487 4.20%	734 6.34%	880 7.60%	117 1.01%	9366 80.85%

(a) Before Clustering

Actual Label \ Predicted Label	Wake	N1	N2	N3	REM
Wake	14279 92.09%	645 4.16%	103 0.66%	73 0.47%	406 2.62%
N1	2404 16.53%	8094 55.67%	2082 14.32%	118 0.81%	1841 12.66%
N2	256 0.67%	1273 3.32%	33611 87.78%	2019 5.27%	1131 2.95%
N3	26 0.25%	87 0.85%	1308 12.80%	8779 85.93%	17 0.17%
REM	271 2.34%	436 3.76%	565 4.88%	66 0.57%	10246 88.45%

(b) After Clustering : (+ average over 5-clusters)

Figure 2.11: Five-stage classification confusion matrices: comparing actual label which is obtained by sleep technicians vs. the proposed method on held-out 100 test patients after removing the transition epochs in training step, and defining two new channels for half of the population. “CC BY 4.0”

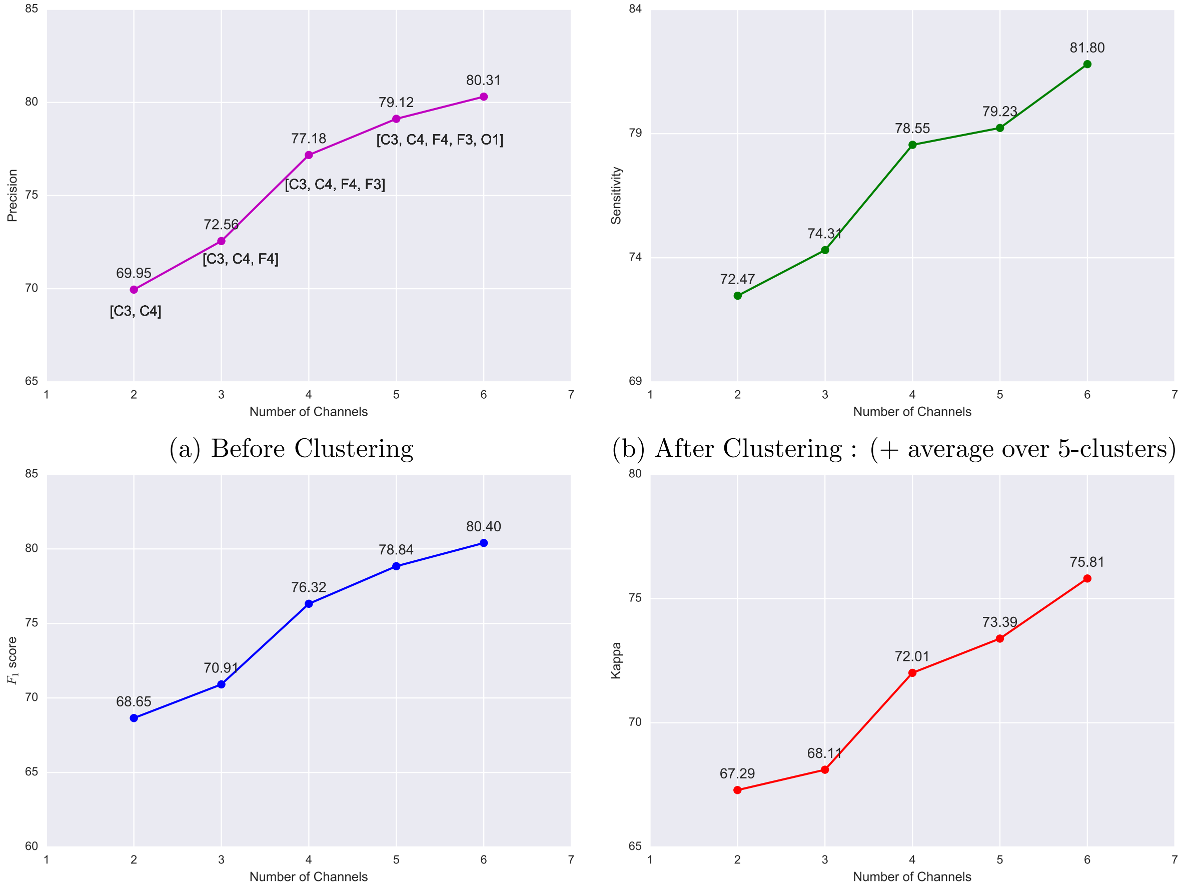
Table 2.6: Per-class performance achieved **before** clustering and after removing transition epochs and replacing O1 and O2 channels with the average of (O1,C3) and (O2, C4)

Sleep Stages	Precision	Sensitivity	F1-Score	Kappa	Number of Samples
Wake	0.77	0.84	0.81	0.75	15506
N1	0.63	0.49	0.55	0.47	14539
N2	0.84	0.82	0.83	0.69	38290
N3	0.73	0.78	0.76	0.72	10217
REM	0.66	0.80	0.72	0.67	11584

Table 2.7: Per-class performance achieved **after** clustering and removing transition epochs and replacing O1 and O2 channels with the average of (O1,C3) and (O2, C4)

Sleep Stages	Precision	Sensitivity	F1-Score	Kappa	Number of Samples
Wake	0.82	0.92	0.78	0.83	15506
N1	0.76	0.55	0.64	0.58	14539
N2	0.89	0.88	0.88	0.78	38290
N3	0.79	0.85	0.82	0.79	10217
REM	0.75	0.88	0.81	0.77	11584

In this work, spatial covariance matrices are used as input features, where the dimension of the covariance matrix is $C \times C$ where C is the number of channels on the scalp. When they are mapped to tangent space, the dimension of the feature is $\frac{C(C+1)}{2}$. Therefore, we run the proposed method with varying numbers of channels to find important channels for the sleep staging task. We plot the 5-stage classification performance w.r.t. the different number of channels and the best combination of them in Fig (2.12). If one is to use two channels for extracting features, C3 and C4 channels are the best options based on the classification performance. For the cases of three, four, and five channels [C3, C4, F4], [C3, C4, F4, F3], [C3, C4, F3, F4, O1] provide the best classification performance, respectively. In this work, the best performance is obtained using all available channels.



(a) Before Clustering

(b) After Clustering : (+ average over 5-clusters)

(c) Before Clustering

(d) After Clustering : (+ average over 5-clusters)

Figure 2.12: Classification performance w.r.t. different numbers of channels: The best-obtained performance based on channels combination has been shown. “CC BY 4.0”

2.5 Sub-grouping of Population for Seizure Detection

To demonstrate the merit of the proposed approach, we used a public EEG database, the PhysioNet CHB-MIT database. This database contains EEG data with 23 channels from 23 patients divided among 24 cases (one patient has 2 recordings, 1.5 years apart) [83] (www.physionet.org/physiobank/database/chbmit/). The goal in this database is to detect whether a 10 second segment of signal contains a seizure or not with high sensitivity and specificity and low false negative rate, as annotated in the database.

At the first step of pre-processing, a 5^{th}-order Butterworth 0.5-30 Hz band-pass filter was applied. Each recording was divided into 10 sec epochs and classified as either dominantly seizure or non-seizure (using expert labels). Then, the FFT coefficients were extracted in three standard bands: theta (4-7 Hz), alpha (8-13 Hz) and beta (13-30 Hz). With a bin size of 0.1 Hz, this resulted in 250 Fourier coefficients for each of the 23 channels. These coefficients were then concatenated and covariance matrices extracted. Then to increase the similarity of the data between subjects, each covariance matrix was transformed per Eq (2.8). After subject selection, a SVM classifier was trained on labeled data from the subjects that were located in the same cluster and then tested on the withheld patient (i.e via a leave-one-subject-out cross validation (LOSO-CV) procedure). In order to use many popular and efficient classifiers, most of the literature focuses on mapping the covariance matrices into a tangent space of Riemannian manifolds to extend Euclidean-based algorithms to the Riemannian manifold of the SPD matrices [6, 109]. The SVM classifier can be applied on the tangent space located at the geometric mean of the whole set of trials from relevant subjects to a given test subject as follows: $\mathbf{\Sigma}_{\mathfrak{G}} = \mathfrak{G}(\mathbf{\Sigma}_i, i = 1, \cdots, I)$. Each SCM, $\mathbf{\Sigma}_i$, is then mapped into this tangent space, to yield the set of $m = \frac{n(n+1)}{2}$ dimensional vectors [6]:

$$s_i = \mathbf{\Sigma}_{\mathfrak{G}}{}^{-\frac{1}{2}} log_{\mathbf{\Sigma}_{\mathfrak{G}}}(\mathbf{\Sigma}_i) \mathbf{\Sigma}_{\mathfrak{G}}{}^{-\frac{1}{2}} \tag{2.10}$$

In the experiments detailed here, the LIBSVM toolbox [16] was used.

Table 2.8 provides the per-patient (LOSO-CV) results and table 5.2 summarizes the average results and compares them to the state-of-the-art. The methods proposed previously by Chen *et al.* [18] and Thodoroff *et al.* [92] are based on the wavelet transformation and deep learning, respectively. Table 5.2 shows an increase over previous works in accuracy and specificity by 2-3%. (Subject-specific works are not included in this comparison, since training and testing on the same subject is less useful and inflates statistics.) We also note that we improve the false positive rate from 1.7/hour to 0.77/hour over Shoeb's original work [83]. To the best of our knowledge, the method described in this article is the first work to propose a subject selection on a Riemannian manifold for unsupervised cross-subject seizure detection.

Table 2.8: Performance on the CHB-MIT database

Subject ID	**Accuracy (%)**	**Sensitivity (%)**	**False Positive rate (seizures/h)**	**Latency (sec)**
1	93.33	98.28	0.194	5.10
2	84.47	100.00	0.43	7.52
3	93.51	100.00	0.26	2.63
4	91.41	84.14	0.22	7.42
5	94.25	100.00	0.34	4.46
6	82.79	46.63	1.74	3.01
7	86.56	98.22	1.48	5.62
8	88.89	100.00	0.35	4.02
9	95.41	100.00	1.28	8.23
10	92.82	100.00	1.20	2.87
11	94.39	85.16	0.46	2.52
12	84.13	62.40	2.34	5.63
13	90.62	83.53	2.86	8.12
14	84.73	56.33	0.55	3.78
15	86.13	78.49	0.24	5.85
16	86.40	58.42	1.66	3.34
17	90.84	81.19	0.82	6.21
18	85.40	97.93	0.41	5.13
19	91.84	100.00	0.21	9.89
20	93.76	69.59	0.58	2.84
21	93.62	100.00	0.46	2.78
22	87.28	100.00	0.52	12.44
23	92.76	68.79	0.14	1.36
24	90.76	89.38	0.10	5.01
mean±(std)	89.84 ± (3.90)	85.77±(16.96)	0.77±(0.75)	5.24±(2.65)

Table 2.9: Performance comparison of works on the CHB-MIT database

Method	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**
Chen et al [18]	86.83 %	85.29 %	87.38 %
Thodoroff et al [92]	84.18%	85.16 %	83.21 %
Proposed Method	**89.84** %	**85.77** %	**89.64** %

2.6 Discussion

We presented a method that relies on population sub-grouping on a robust feature space for the scoring of sleep stages from EEG data. The proposed method includes an unsupervised clustering approach on a Riemannian manifold, a non-Euclidean space.

Our hypothesis is, that by sub-grouping of patients, the diversity in EEG data is diminished, and the transferability of the trained model is increased. In cross-subject classification task, inter-subject variability leads to different probability distributions between individuals, and consequently poor generalization across subjects. Potentially, individuals with different biological phenotypes would provide enough diversity in the dataset, but achieving this would require vastly more high-quality labelled data than is currently available to a single researcher, and a single network cannot therefore be robust to such variabilities. By sub-grouping the population, we attempted to diminish the diversity in EEG signals and create homogeneous cohorts in dataset. Evaluation of the proposed method on the 2018 PhysioNet Challenge data, for 5-class sleep staging demonstrates that population sub-grouping significantly improved the results compared to no sub-grouping. This was particularly marked for challenging sleep stages such as N1 and REM. We have compared the performance of a CNN model on the Riemannian manifold before/after clustering. Therefore, we claim that the sub-grouping of the population improves performance. Perslev et al. [6] used the EEG signals and a temporal fully convolutional network for sleep staging on the same dataset, where their model's performance is F1=0.77 under 5-fold cross-validation on this dataset, which is inferior to the proposed method with population sub-grouping. To reduce the effect of noisy labeling, we removed transition epochs (one which does not have neighboring sleep stages of the same class), which improved the total performance. Another potential approach to dealing with the noisy labels would be to relabel the transition epochs by an ensemble learning approach (multiple independent classifiers) and using a consensus or weighted voting approach, particularly on shorter windows to identify the transition points. However, this approach is left for future work. Besides, to show the robustness of covariance matrices to the electrode's misplacement, we replaced O1 and O2 channels with the average signal between (C3, O1) and (C4, O1), respectively. By rerunning the experiment, the proposed method improved the classification performance by (7-8%). The effect of the number of channels and the best combination of channels for the proposed method were investigated, which shows that central electrodes (C3 and C4) carry transferable and discriminable information about sleep stages, which is consistent with recommended or standard EEG placements based on RK standard [43]. We have compared the performance of a CNN model on the Riemannian manifold before/after clustering. Results provide evidence that the sub-grouping of the population improves performance. Notably, Perslev et al. [6] used the EEG signals and a temporal fully convolutional network for sleep staging on the same dataset, where their model's performance is F1=0.77 under 5-fold cross-validation on this dataset, which is inferior to the method with population sub-grouping proposed here (F1 =0.8). We note that this work focused only on the EEG, and the addition of other signals, such as the EMG, EOG and ECG is likely to boost performance further. Moreover, we are interested in applying the proposed method on other EEG acquisition protocols having different sensor configurations and also other sleep labeling protocols, e.g. Rechtschaffen and Kales (RK) protocol, in future work.

Chapter 3

Attentive Adversarial Network for Large-Scale Sleep Staging

3.1 Abstract

Current approaches to developing a generalized automated sleep staging method rely on constructing a large labeled training and test corpora by leveraging electroencephalograms (EEGs) from different individuals. However, data in the training set may exhibit changes in the EEG pattern that are very different from the data in the test set due to inherent inter-subject variability, heterogeneity of acquisition hardware, different montage choices and different recording environments. Training an algorithm on such data without accounting for this diversity can lead to underperformance. In order to solve this issue, different methods are investigated for learning an invariant representation across all individuals in datasets. However, all parts of the corpora are not equally transferable. Therefore, forcefully aligning the nontransferable data may lead to a negative impact on the overall performance. Inspired by how clinicians manually label sleep stages, this paper proposes a method based on adversarial training along with attention mechanisms to extract transferable information

across individuals from different datasets and pay attention to more important or relevant channels and transferable parts of data, simultaneously. Using two large public EEG databases - 994 patient EEGs (6,561 hours of data) from the PhysioNet 2018 Challenge (P18C) database and 5,793 patients (42,560 hours) EEGs from Sleep Heart Health Study (SHHS) - we demonstrate that adversarially learning a network with attention mechanism, significantly boosts performance compared to state-of-the-art deep learning approaches in the cross-dataset scenario. By considering the SHHS as the training set, the proposed method improves, on average, precision from 0.72 to 0.84, sensitivity from 0.74 to 0.85, and Cohen's Kappa coefficient from 0.64 to 0.80 for the P18C database.

3.2 Introduction

A third of the US population experiences less than the recommended amount of sleep, which is linked to many chronic diseases and conditions, such as type 2 diabetes, heart disease, obesity, and depression [2]. As sleep pathologies are increasingly recognized as crucial factors in many illnesses, both as effects and causes, and the improved availability of low-cost sleep monitoring devices continues to accelerate the field, the volume of data continues to expand. The need for automated sleep staging and diagnostics is, therefore, more acute, 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') and 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) [9]. In addition to the time and cost involved in manual sleep staging, the significant inter-expert variability remains an issue ([108]). 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 is a challenging task in this area. Due to the non-stationary nature of the EEG signal [46], the changes in statistical characteristics of the signal with time, a classifier that is trained on a temporally-limited amount of data from an individual may poorly generalize on EEG data recorded at a different time on the same subject. Another issue with the low generalizability issue in EEG data is related to high inherent inter-subject variability in the way an EEG manifests, which limits the usefulness of EEG applications. This phenomenon arises due to physiological differences (e.g. skull shape) between individuals, and 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 across people [29]. Moreover, electrode sensor montages (the points at which the electrodes are attached and the references points) may differ and different 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' variety of training or even attentiveness on a given day. All these factors lead to significant variabilities in EEG signals.

In this paper, a multi-adversarial neural network with an attention mechanism is proposed to tackle these challenges in order to develop a generalized model for automated EEG sleep staging.

Technical Significance: The proposed method is the first work to combine multi-adversarial networks with attention mechanisms for sleep staging with two large datasets. The proposed method can operate in an unsupervised manner to highlight the critical channels contributing to the class estimate and pay attention to the more transferable part of EEG patterns across subjects, which contribute more to the classification task. Using two large EEG databases, 994 patient EEGs from the PhysioNet 2018 Challenge database (≈ 6,561 hours of data) and 5,793 patients (≈ 42,560 hours) EEGs from Sleep Heart Health Study (SHHS), we demonstrate that adversarially learning a network with an attention mechanism significantly boosts performance compared to state-of-the-art deep learning approaches in the cross-dataset scenario. The proposed method improves, on average, precision from 0.72 to 0.84, sensitivity from 0.74 to 0.85, and a Cohen's Kappa coefficient from 0.64 to 0.80 for the PhysioNet 2018 Challenge database.

Clinical Relevance: Automated sleep staging from the EEG has been previously proposed to incorporate a particular carefully engineered feature extraction part and calibrating the data for each subject or dataset. These methods are time-consuming and costly and do not generalize well to other datasets or even subjects. In general, previous studies have analyzed data from fewer than 100 individuals, and most of them are on homogeneous and/or non-public datasets. Disregarding heterogeneity between individuals and insufficient sample size leads to essential limitations of clinical usage in real-life problems. Therefore, the proposed method attempts to solve this problem by training a network on a large dataset and testing it on another large dataset by learning transferable features, only paying attention to the essential part of data. The proposed method finds the important channels in the dataset, which can provide explanatory information for the clinician.

3.3 Related Work

As mentioned before, in order to build a data set large enough to make health-AI models work, studies often combine data from multiple hospitals. Therefore, the condition or device used to capture the data can vary from hospital to hospital and even department to department. Electrode mismatching, inherent inter-subject variability and the non-stationary nature of EEG signals, lead to different joint distribution, $P(X, Y)$ between different recording, where X and Y are feature and label space, respectively. Moreover, the generalizability of a model that is trained on a dataset is low when it is going to be tested on another dataset acquired in different environments with different acquisition hardware. Class imbalances between hospitals can then be associated with hardware differences (such as filter cut-offs). While this can be mitigated through careful inspection of the data, electrode placement differences and patient physiological differences are harder to identify and mitigate. Therefore, the transferability of the trained model for an application on unseen subjects is degraded. The reason behind this problem is that the primary assumption in machine learning techniques is that training and test data should be drawn from the same distribution, an assumption that does not necessarily hold in large biomedical datasets. In other words, data from two hospitals that are recorded with different devices and set-ups, but for the same task, can not necessarily be leveraged directly in a machine learning approach. The main question raised is that of how to boost performance in the real-life application of EEG through the development of a generalized model across a large population. This issue could be interpreted as how one can diminish spatial and temporal shifts across individuals from different hospitals or recording environments to handle these different variabilities.

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 [10]. In the literature, it has been shown that Symmetric Positive Definite (SPD) matrices provide a strong ability to representations the brain signals [20, 4]. The covariance matrix is a typical example of SPD matrices, which has been employed in several studies [78, 75, 50]. These studies showed that using second-order statistics of multi-channel signals reduce inter-subject and intra-subject variabilities between

EEG signals. The spatial covariance matrix can well separate useful information about brain functional connectivity structure ([4]) and create 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 [20].

Other studies [49, 56, 90] 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 [54]. 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 instance importance using labeled training samples and unlabeled test samples [56, 49, 44].

Other methods to increase the generalization ability of a model involve transfer learning - finding subsets of past subjects to initialize a classifier for training on a new subject [109]. Bolagh et al. [12, 11] proposed subject-selection and subject clustering to select relevant individuals based on the similarity between the EEG pattern of different individuals. Raza et al. [74] proposed bagging methods to handle mismatching between training and test distributions. Chai et al. [15] proposed an adaptive subspace feature matching (ASFM) 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.

Recently, multiple authors have focused on developing an automated sleep scoring based on applying deep learning (DL) methods [10, 58, 89, 71]. Due to the nature of EEGs, which consist of spatial and temporal information, most convolutional and recurrent processing methodologies are suitable for EEG processing. Biswal et al. [10] proposed to use a combination of deep recurrent and convolutional neural networks to classify sleep stages as well as sleep abnormalities events. Spectrograms from EEG channels were fed to the CNN module as input, and then the CNN output was fed into a bidirectional recurrent neural network. Zhang et al. [111] also used the same approach for assessing the generalization capability of their model by testing their model on two different datasets. These methods have gained attention these days 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 [52, 35], 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 non-linearity, which further reduces their real-life application in a clinical setting.

Recently, the use of generative adversarial networks (GANs) [34] to handle temporal and spatial shifts has received more attention [94, 81, 51]. In fact, similar to GANs, a two-player minimax game is constructed, in which the first player discriminator between training and test sets and the second player is adversarially trained to deceive the discriminator and extract transferable features [28]. 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, forcefully aligned the irrelevant channels, and EEG patterns may have a large impact on overall performance. An attention mechanism [96] is an effective method to focus on essential regions of data, with numerous successes in deep learning tasks such as classification, segmentation, and detection.

In this work, we use the Sleep Heart Health Study (SHHS) database (a private database that is available on request from the study investigators) to develop an attention mechanism to highlight relevant channels and transferable part of EEG pattern across datasets. The attention mechanisms explore the part of data that are more similar across different subjects and contribute more to the classification task. More specifically, a multi-adversarial neural network with an attention mechanism is proposed to tackle the challenges detailed above in creating a generalized model for EEG processing. Finally, the algorithm's performance is assessed on the largest open-access EEG database – the PhysioNet 2018 Challenge (P18C) database [30, 32]. All data used in the study are de-identified, and therefore an ethics/institutional review board waiver was provided for this research.

3.4 Method

In this paper, we focus on the cross-dataset scenario, where two datasets from different hospitals, with two different individuals, acquisition hardware, environment, are leveraged to develop a generalized sleep stage algorithm. The goal is to create a classifier (network) on a training dataset (labeled source domain $\mathcal{D}_s$), which generalizes well on the test dataset (unlabeled target domain $\mathcal{D}_t$).

In the following, we describe the proposed method based on a multi-adversarial neural network with an attention mechanism for the cross-dataset sleep staging task. At first, a high-level overview of the adversarial domain adaptation method, which is proposed in ([28]), is given, and then the proposed method for automatically classify sleep stages with attention mechanisms is presented.

Ganin et al. ([28]) inspired the idea of GANs and used the same idea for the domain adaptation problem, where adaptation behavior is achieved via adversarial training. The feature extractor, similar to the generator in GANs, tries to perform some transformation on data from two domains such that the transformed samples have the same distribution. The second network, (a discriminator network), similar to GANs, should be able to classify the domains as source and target. 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:

$$C(\theta_f, \theta_y, \theta_d) = \frac{1}{n_{tr}} \sum_{x_i \in \mathcal{D}_{tr}} L_y(G_y(G_f(\mathbf{x}_i)), y_i) - \frac{\lambda}{n} \sum_{\mathbf{x}_i \in \mathcal{D}_{tr} \cup \mathcal{D}_{te}} L_d(G_d(G_f(\mathbf{x}_i)), d_i) \quad (3.1)$$

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 representations extracted from all EEG channels across subjects. Since it is obvious that some parts of the brain are more involved in a given task, all channels are not equally transferable. Moreover, some parts of the EEG pattern are significantly dissimilar across subjects. Therefore, to highlight the important channels in the task, we split the discriminator G_d in Equation (4.1) into K channel-wise discriminators G_k^d; $k = 1, 2, \ldots, K$, and each is responsible for matching the training and test datasets corresponding to channel k, as shown in Fig (3.1). Applying this to all K discriminators G_k^d; $k = 1, 2, \ldots, K$ yields

$$\mathcal{L}_{ch} = \frac{1}{Kn} \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in \mathcal{D}_{tr} \cup \mathcal{D}_{te}} L_d(G_d^k(\mathbf{f}_i^k), d_i) \tag{3.2}$$

where $\mathbf{f}_i^k = (G_f^k(\mathbf{x}_i))$ is the feature representation in channel k, d_i is the domain label of point $\mathbf{x}_i$, L_d is the cross-entropy loss.

the output $\hat{d}_i^k = G_d^k(\mathbf{f_i^k})$ of each channel-wise discriminator G_d^k is the probability of the region k in windows i belonging to the training set. When the probability approaches 1, it indicates that the channel k belongs to the training set, and 0 represents that it belongs to the test set.

Following [100], the entropy functional is used as criteria to generate the attention value. The output of each discriminator G_d^k is in the range $[0-1]$, which is the probability of signal from the channel $\mathbf{x}_i^k$ belonging to the training dataset. When the probability approaches 1, it indicates that the channel k belongs to the training dataset, and 0 represents that it belongs to the test dataset. Using entropy helps the algorithm to increase certainty about the signal of channel k and quantify its transferability.

$$w_i^k = 1 - H(G_d^k(\mathbf{f}_i^k)) \tag{3.3}$$

where $H(p) = -\sum_j p_j log(p_j)$. We also add a residual connection for more stable optimization. Therefore, the final channel feature representation $\mathbf{h}_i$ generated from attended channel features can be expressed as:

$$\mathbf{h}_i = \sum_{k=1}^{K}(1 + w_i^k) \cdot \mathbf{f}_i^k \tag{3.4}$$

Finally, the minimum entropy regularization is utilized to refine the classifier adaptation. However, the entropy for the windows that are similar across datasets should be minimized. Therefore, the network should attend more to the windows, which have low domain discrepancy, which equals to minimizing the entropy for these windows.

The attentive entropy loss L_a can be expressed as follows:

$$\mathcal{L}_a = -\frac{1}{n} \sum_{\mathbf{x}_i \in \mathcal{D}_{tr} \cup \mathcal{D}_{te}} \sum_{j=1}^{c}(1 + H(\hat{d}_i)) \cdot \mathbf{p}_{i,j} \cdot \log(\mathbf{p}_{i,j}) \tag{3.5}$$

where c is the number of classes, and $\mathbf{p}_{i,j}$ is the probability of predicting point $\mathbf{x}_i$ to class j. Therefore, the end-to-end optimization framework can be express as follows:

$$\begin{aligned} C(\theta_f, \theta_y, \theta_d, \theta_d^k|_{k=1}^K) =& \frac{1}{n_{tr}} \sum_{\mathbf{x}_i \in \mathcal{D}_{tr}} L_y(G_y(G_f(\mathbf{x}_i)), y_i) + \frac{\gamma}{n} \sum_{\mathbf{x}_i \in \mathcal{D}_{tr} \cup \mathcal{D}_{te}} \sum_{j=1}^{c}(1 + H(\hat{d}_i)) \cdot \mathbf{p}_{i,j} \cdot \log(\mathbf{p}_{i,j}) \\ & - \frac{\lambda}{n} \left[\sum_{\mathbf{x}_i \in \mathcal{D}_{tr} \cup \mathcal{D}_{te}} \mathcal{L}_d(G_d(\mathbf{h}_i, d_i))] + \frac{1}{K} \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in \mathcal{D}_{tr} \cup \mathcal{D}_{te}} \mathcal{L}_d(G_d^k((G_f(\mathbf{x}_i)))) \right] \end{aligned} \tag{3.6}$$

Where λ_s, and γ, chosen via grid search, are hyper-parameter that trade-off between domain discriminator and attentive entropy loss, respectively. The network parameters can be learned end-to-end by a minimax optimization procedure as follows:

$$
\begin{aligned}
(\hat{\theta}_f, \hat{\theta}_y) &= \underset{\theta_f, \theta_y}{\arg\min}\, C(\theta_f, \theta_y, \theta_d, \theta_d^k|_{k=1}^K) \\
(\hat{\theta}_d, \hat{\theta}_d^i, \cdots, \hat{\theta}_d^K) &= \underset{\theta_d, \theta_d^i, \cdots, \theta_d^K}{\arg\max}\, C(\theta_f, \theta_y, \theta_d, \theta_d^k|_{k=1}^K)
\end{aligned} \tag{3.7}
$$

Fig (3.1) shows the overall framework of the proposed method.

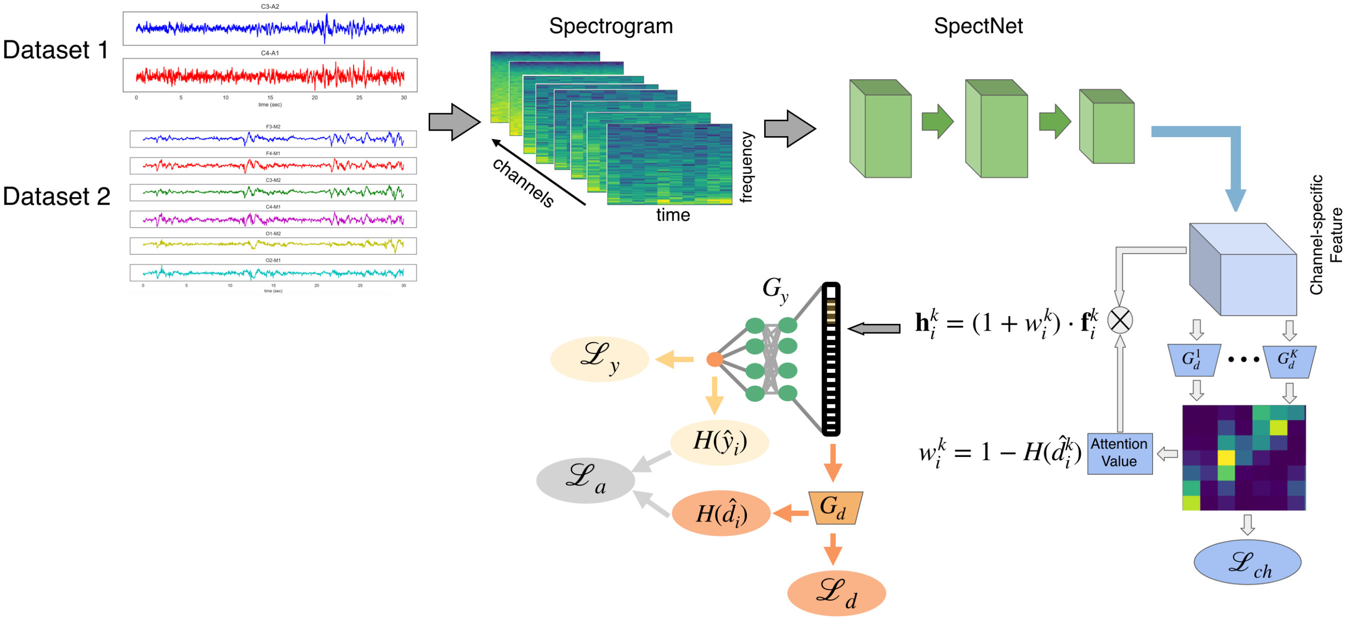


Figure 3.1: Framework of proposed method: After extracting spectrogram from all EEG channels, we put those to feature extractor (we called SpectNet). Then, a multi-adversarial network (blue) is developed for highlighting important channels across datasets attention, and a adversarial network (orange) is used to boost the certainty of output for similar signals in the feature space. “CC BY 4.0”

3.5 Experimental Set-Up

3.5.1 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. Following ([23]), 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 ([8]), we merge N3 and N4 stages into a single N3 stage. Table 4.1 shows number of sleep stages per class.

PhysioNet 2018 Challenge: The P18C database includes PSG data from 1,985 subjects included 65% male and 35% women, which were monitored at the MGH sleep laboratory for the diagnosis of sleep disorders. The data were partitioned two-part: public dataset (n = 994) and hidden dataset (n = 989). The sleep stage labels for 994 of the recordings were made available for the public dataset, where includes Wake, REM N1, N2, and N3 stages. It includes multiple physiological signals that are all sampled at 200 Hz and were manually scored by certified sleep technicians at MGH sleep laboratory according to the AASM guidelines into 30 second 'epochs'. In this work, we use the EEG channels, which include 'F3-M2', 'F4-M1', 'C3-M2', 'C4-M1', 'O1-M2', and 'O2-M1' channels.

Table 3.1: Number of subjects and samples 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
P18C 2018	994	145,558	135,409	372,257	101,678	113,872

Fig (3.2) illustrates the electrode position on the scalp (looking from top down, with the nose at the top of the diagram). Note that the green electrodes (C3 and C4) are common to both databases and were used in this study.

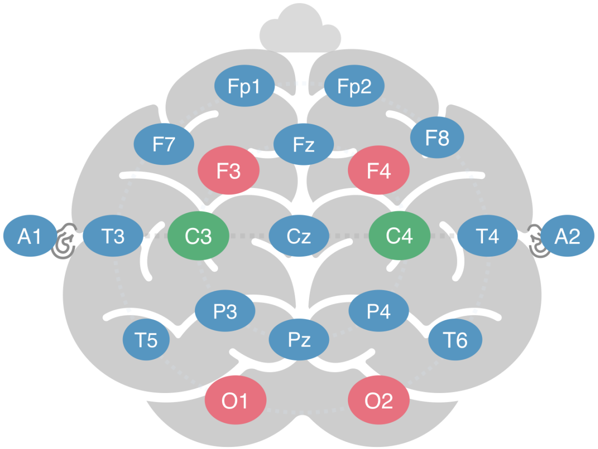


Figure 3.2: 10-20 EEG Placement: Red electrodes were used in the P18C database and blue electrodes were used in the SHHS database. Green electrodes (C3 and C4) are common to both databases. “CC BY 4.0”

3.5.2 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 $\mathbf{S} \in \mathbb{R}^{F \times T}$ where $F = 129$ (the number of frequency bins), and $T = 29$ (the number of spectral columns).

3.5.3 Network Implementation

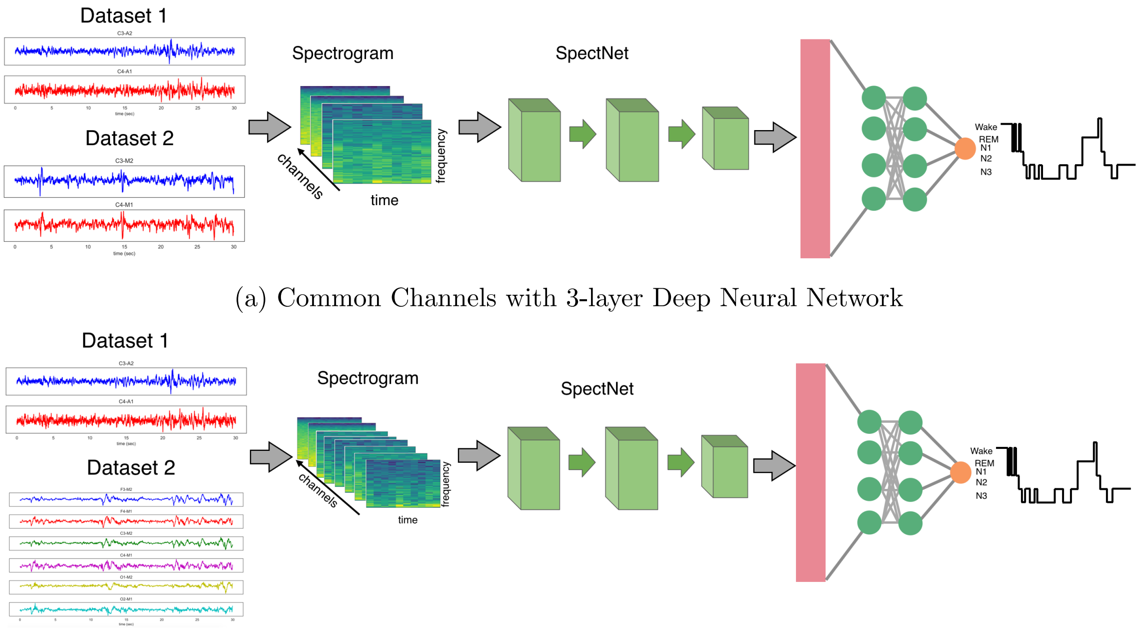
For extracting features for the adversarial neural network, we use the same architecture of Biswal et al. ([10]). It includes a 3-layer of 1-D CNN (kernel size = 3), which was applied to each EEG channel, followed by batch normalization (BatchNorm), rectified linear (ReLU) units, and max pooling units, we called it as SpectNet here. A cross-entropy loss function is used as a domain 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 [28] to learn the weights of a deep neural network. 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 0.4 and Python3. See Figure 3.3.

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. Due to a large number of patients, the SHHS database and the P18C database are considered as the training (labeled/source) and test (unlabeled/target) sets, respectively.

3.6 Results

As a baseline, we start with simple experiments:

- Extract spectrograms of common EEG channels, C3 and C4, and use a 3-layer SpectNet, where it is followed by a fully connected neural network and softmax to classify sleep stages.
- Repeat the above experiment with all EEG channels.

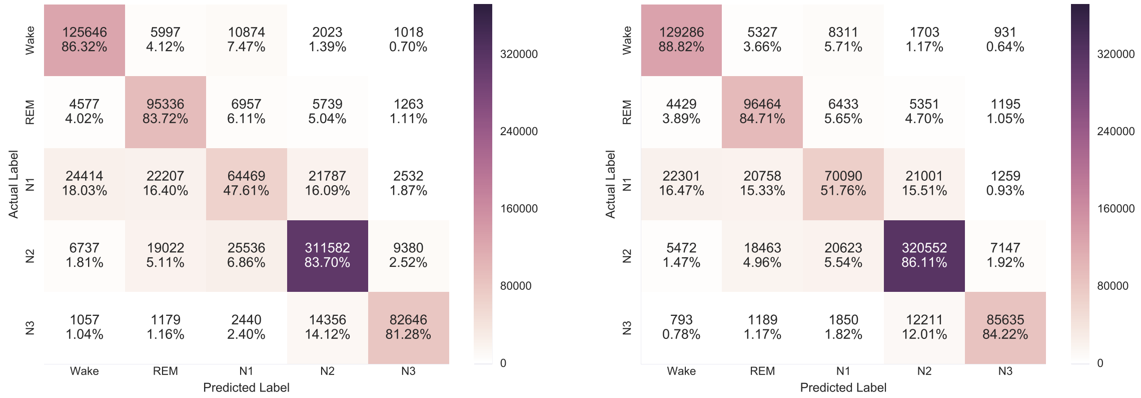


(a) Common Channels with 3-layer Deep Neural Network

(b) All available channels with 3-layer Deep Neural Network DNN

Figure 3.3: Baseline experiment. “CC BY 4.0”

Figure (3.4) provides the confusion matrix for sleep staging, which shows the SpectNet agreement with expert scores. Sleep experts score each 30 second EEG epoch as wake, REM, non-REM stage 1, 2, or 3. Table (3.2) and (3.3) present the performance of each class achieved by using a simple three-layer, SpectNet, with two common channels and all channels of P18C dataset, respectively. Based on this experiment, using all available channels of P18C database boosts the performance by 3% on average. It seems that when the algorithm exploits all channels, the N1 class can be better distinguished than with fewer channels.



(a) Using common channels (C3 and C4)

(b) Using all channels

Figure 3.4: Confusion matrix for sleep staging for using all data from P18C dataset, showing SpectNet agreement with expert scores. The SpectNet is trained on SHHS dataset and tested on two common channels and all channels. “CC BY 4.0”

Table 3.2: Per-class performance achieved using **two common channels** (C3 and C4) by a 3-layer CNN network and a softmax layer

Sleep Stages	Precision	Sensitivity	F1-Score	Kappa	Acc	Number of Samples
Wake	0.77	0.86	0.81	0.76	0.92	145558
REM	0.66	0.83	0.74	0.68	0.91	113872
N1	0.58	0.47	0.52	0.43	0.85	135409
N2	0.87	0.83	0.85	0.73	0.86	372257
N3	0.85	0.81	0.83	0.80	0.95	101678
avg	0.75	0.76	0.75	0.68	0.90	-

Based on this experiment, using all available channels of P18C database boosts the performance by 3% on average. It seems that when the algorithm exploits all channels, the N1 class can be better distinguished than with fewer channels. 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. [3] 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 (3.4).

Table 3.3: Per-class performance achieved using **all available channels** by a 3-layer CNN network and a softmax layer

Sleep Stages	Precision	Sensitivity	F1-Score	Kappa	Acc	Number of Samples
Wake	0.79	0.88	0.83	0.79	0.93	145558
REM	0.67	0.84	0.75	0.70	0.91	113872
N1	0.65	0.51	0.57	0.50	0.87	135409
N2	0.88	0.86	0.87	0.76	0.88	372257
N3	0.89	0.84	0.86	0.84	0.96	101678
avg	0.78	0.79	0.78	0.72	0.91	-

Using all channels from two datasets and using adversarial domain adaptation (ADA) [28], with the SHHS database as the training set and the P18C database as the test set, as shown in Figure (3.5) we see an improved performance with all metrics. The performance on the test set is presented in Figure (3.6). Table (3.4) presents the performance of each class achieved with this method. One can conclude that adversarially learning transferable features across subjects boosts the performance of N1 class significantly.

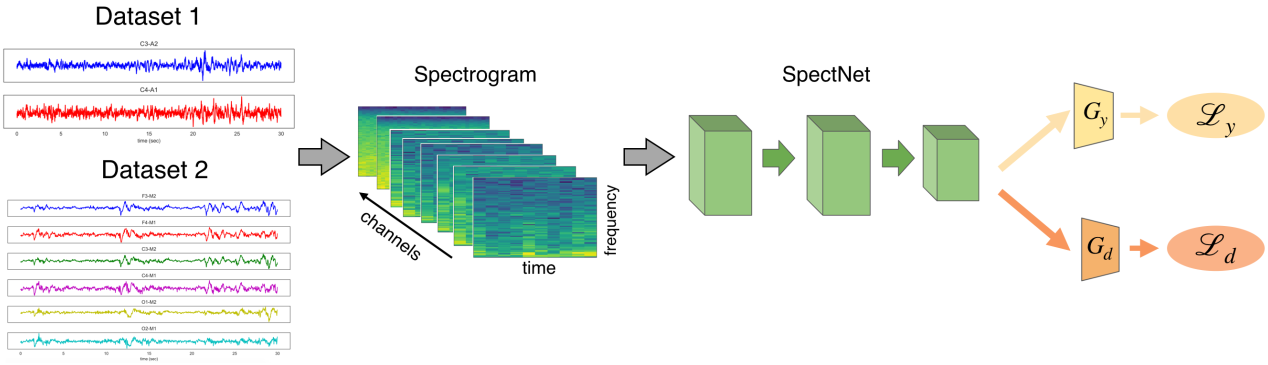
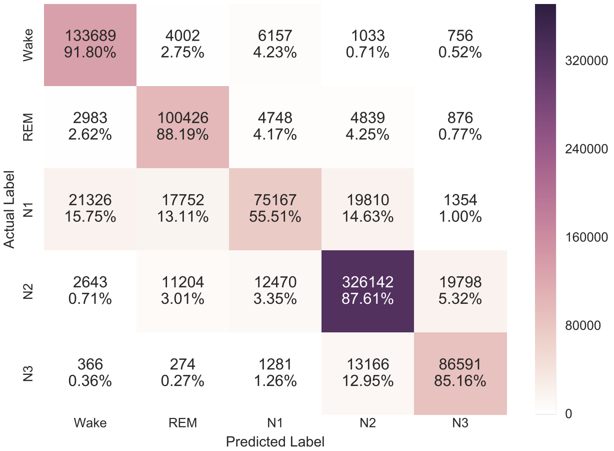


Figure 3.5: Using all channels of two datasets and adversarial domain adaptation network (ADA) without attention mechanism “CC BY 4.0”



Actual Label \ Predicted Label	Wake	REM	N1	N2	N3
Wake	133689 91.80%	4002 2.75%	6157 4.23%	1033 0.71%	756 0.52%
REM	2983 2.62%	100426 88.19%	4748 4.17%	4839 4.25%	876 0.77%
N1	21326 15.75%	17752 13.11%	75167 55.51%	19810 14.63%	1354 1.00%
N2	2643 0.71%	11204 3.01%	12470 3.35%	326142 87.61%	19798 5.32%
N3	366 0.36%	274 0.27%	1281 1.26%	13166 12.95%	86591 85.16%

Figure 3.6: Confusion Matrix using all channels and ADA, using SHHS database as training (source) set and P18C database as test (target) set) “CC BY 4.0”

Table 3.4: Per-class performance achieved using all channels and ADA with SHHS database as training (source) set and P18C database as test (target set)

Sleep Stages	Precision	Sensitivity	F1-Score	Kappa	Acc	Number of Samples
Wake	0.83	0.91	0.87	0.83	0.94	145558
REM	0.75	0.88	0.81	0.77	0.93	113872
N1	0.75	0.55	0.63	0.57	0.89	135409
N2	0.89	0.87	0.88	0.78	0.98	372257
N3	0.80	0.85	0.82	0.79	0.95	101678
avg	0.81	0.81	0.80	0.75	0.92	-

Finally, the performance of multi-stage classification using multi-adversarial neural network with attention mechanism is reported in Figure (3.7) and per-class performance is given in Table (3.5).

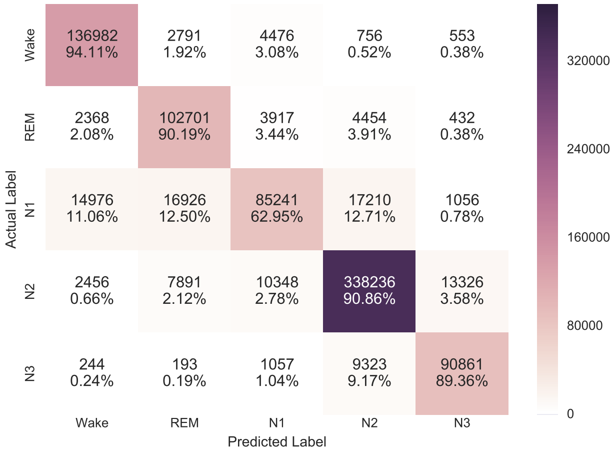


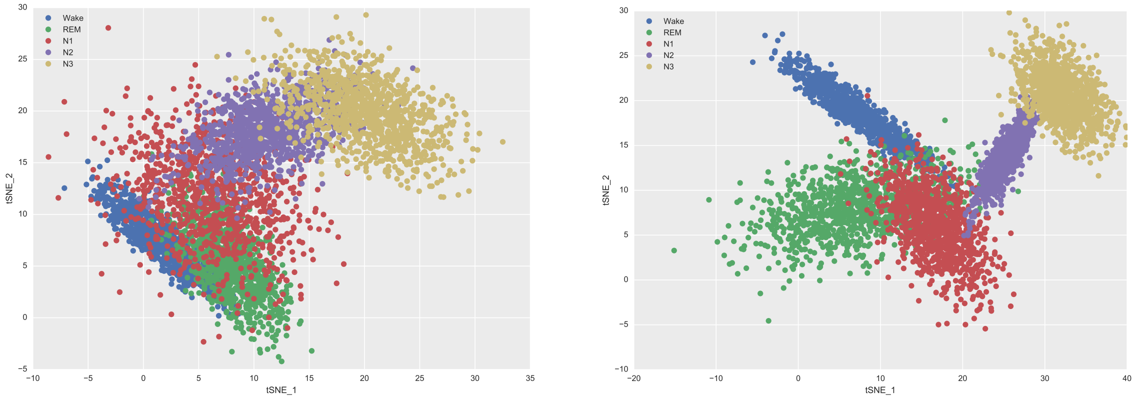
Figure 3.7: Proposed method (SHHS database → P18C database). “CC BY 4.0”

Table 3.5: Per-class performance achieved using all channels and multi-adversarial network with attention mechanisms

Sleep Stages	Precision	Sensitivity	F1-Score	Kappa	Acc	Number of Samples
Wake	0.87	0.94	0.90	0.88	0.96	145558
REM	0.78	0.90	0.84	0.81	0.95	113872
N1	0.81	0.62	0.70	0.66	0.91	135409
N2	0.91	0.90	0.91	0.83	0.91	372257
N3	0.85	0.89	0.87	0.85	0.96	101678
avg	0.84	0.85	0.84	0.80	0.94	-

In terms of evaluation metrics, which are mostly used in the sleep staging task, the proposed method outperforms the state-of-art algorithms on the 2018 P18C database. For instance, Perslev et al. ([71]) obtained 0.77 F1-score under 5-fold cross-validation on the same dataset. The average accuracy of proposed method is 0.94 on the unseen (P18C) database, which is significantly higher than other state-of-the-art methods ([10]). Our proposed method significantly beats their results after using adversarial training with an attention mechanism.

Feature Visualization: Figure (4.3) illustrates the discriminability of learned features in the deep neural network without/with adversarial training for scoring sleep



(a) Without domain adapatation

(b) With adversial domain adapatation

Figure 3.8: 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). "CC BY 4.0"

stages, using t-distributed Stochastic Neighbor Embedding ([57]). The figure visualizes the network activations of the last hidden layer of DNN for each segment from 1000 samples for each class; 500 samples from the SHHS database and 500 samples from PhysioNet Challenge 2018 database. Figure (4.3(a)) shows the representations generated by a conventional deep neural network, where classes are less easily distinguished; there is significant confusion between N1 and wake, and between REM and N2. However, an adversarial neural network with an attention mechanism learns features with high transferability and discriminability. (see Figure (4.3(b).) As mentioned earlier, this figure shows that the N1 stage may not be like any stage, and it is considered as a transition stage between other stages (Wake, REM and N2).

Attention Mechanism: To investigate the key channels (sensors) on the scalp, we show the attention weights across channels for a randomly selected sample from PhysioNet 2018 Challenge database. Figure 3.9 illustrates this - the hotter the color, the larger the attention value. It can be seen that the network pays more attention to features extracted from channel C4 rather than channel C3. Moreover, it seems that the C4 channel is a more transferable channel across databases and subjects. These results intuitively show which channel can be used for a wearable devices to capture

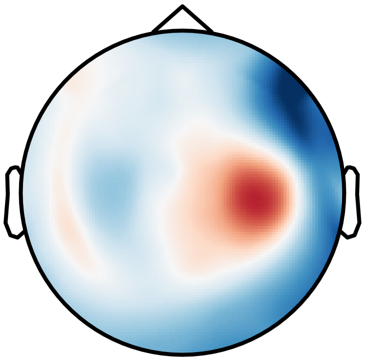


Figure 3.9: Attention visualization of sensors on the brain. “CC BY 4.0”

sleep stages.

3.7 Conclusion

In this work, adversarial training with an attention mechanism was proposed for the sleep staging task across two large and heterogeneous datasets. In the cross-dataset classification task, inherent inter-subject variability, hardware acquisition heterogeneity, and recording environment differences lead to different probability distributions between individuals, and hence poor generalization across subjects/dataset. Potentially, individuals with different biomedical demographics and phenotypes would provide enough diversity in the dataset, in which a conventional network cannot be robust to such variabilities, although given the need to factor in differences in montages, electrode placement errors and hardware systems, the dataset would likely be prohibitively large. The proposed method uses a multi-adversarial network to attend to relevant channels across datasets and highlight the important part of a segment of signal, and extract transferable features across the dataset, which achieves state-of-the-art performance (without prior knowledge) on a large public dataset, the PhysioNet 2018 Challenge database. The proposed method identified the important channel (C4), which suggests single-channel sleep staging with acceptable performance is possible. The method developed in this work can be applied to other biomedical signals (e.g. the electrocardiogram (ECG), 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. Therefore, in addition to integrating adversarial training with attention mechanism, there are two main directions we would like to pursue for future work: 1) to apply the method in the cross-modality scenario, where we combine different modality such as EEG, ECG, and PPG which are recorded simultaneously in sleep; 2) to extend this method to leverage a dataset with different labels, i.e., partial domain adaptation, where the label sets are not equal across the dataset. This is a is much more challenging, but closer to real-world scenarios.

Chapter 4

Importance Weighting with Adversarial Network for Large-Scale Sleep Staging

4.1 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.

4.2 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 [2]. 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) [9]. In addition to the time and cost involved in manual sleep staging, the significant inter-expert variability remains an issue [108]. 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 [46] 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 [29]. 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 [10, 58, 89, 71]. 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.

4.3 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 [10]. In the literature, it has been shown that Symmetric Positive Definite (SPD) matrices provide a strong ability to provide useful representations of brain signals [20, 4]. The covariance matrix is a typical example of an SPD matrix, which has been employed in several studies [78, 75, 50]. 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 [4] 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 [20].

Other studies [49, 56, 90] 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 adaptation, 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 [54]. 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 [56, 49, 44].

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 [109]. Bolagh et al. [12, 11] proposed subject-selection and subject clustering to select relevant individuals based on the similarity between the EEG pattern of different individuals. Raza et al. [74] proposed bagging methods to handle mismatching between training and test distributions. Chai et al. [15] 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 [86] 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 [72] 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. [10] 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. [111] 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 [52, 35], 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) [34] to handle temporal and spatial shifts has received more attention [94, 81, 51]. Notably, Ganin et al. constructed a two-player minimax game (rather like the appriach of GANs), in

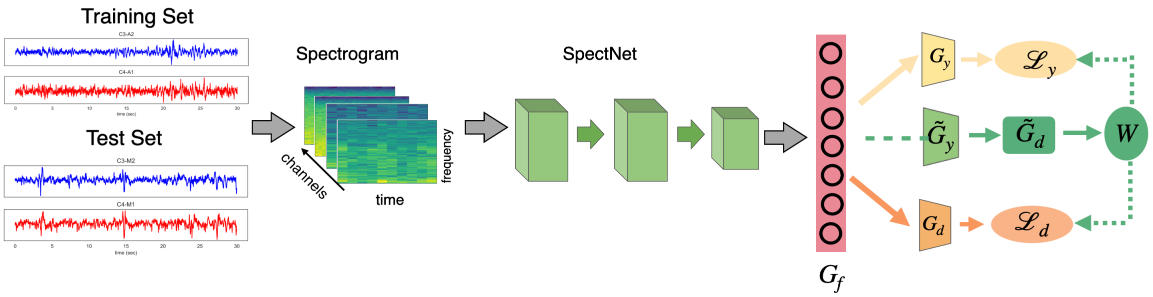


Figure 4.1: The proposed method for generalized sleep staging problem, where G_f is the feature extractor, G_y is the training classifier, G_d is domain discriminator (involved in adversarial training) for alignment features from traing and test; $\tilde{G}_d$ is the auxiliary domain discriminator (uninvolved 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. “CC BY 4.0”

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 [28]. 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 [96] is an effective method to focus on essential regions of data, with numerous successes in deep learning tasks such as classification, segmentation, and detection.

4.4 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 [28]. 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:

$$\begin{aligned} C(\theta_f, \theta_y, \theta_d) = & \frac{1}{n_{tr}} \sum_{\mathbf{x}_i \in \mathcal{D}_{tr}} L_y(G_y(G_f(\mathbf{x}_i)), y_i) \\ & - \frac{\lambda}{n} \sum_{\mathbf{x}_i \in \mathcal{D}_{tr} \cup \mathcal{D}_{te}} L_d(G_d(G_f(\mathbf{x}_i)), d_i) \end{aligned} \tag{4.1}$$

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 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:

$$\begin{aligned} E_{G_y} &= \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} w(\mathbf{x}_i^{tr}) L(G_y(G_f(\mathbf{x}_i^{tr}), y_i^{tr})) \\ &+ \frac{\gamma}{n_{te}} \sum_{n=1}^{n_{te}} H(G_y(G_f(\mathbf{x}_j^{te}))) \qquad (4.2) \\ E_{G_d} &= -\frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} w(\mathbf{x}_i^{tr}) \log(G_d(G_f(\mathbf{x}_i^{tr}))) \\ &+ \frac{1}{n_{te}} \sum_{n=1}^{n_{te}} \log(1 - G_d(G_f(\mathbf{x}_j^{te}))) \qquad (4.3) \end{aligned}$$

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:

$$
\begin{aligned}
(\hat{\theta}_f, \hat{\theta}_y) &= \arg\min_{\theta_f, \theta_y} E_{G_y} - E_{G_d} \\
(\hat{\theta}_d) &= \arg\max_{\theta_d} E_{G_d}
\end{aligned}
\tag{4.4}
$$

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:

$$
\tilde{G}_d(G_f(\mathbf{x}_i)) = \sum_{c=1}^{5} \tilde{G}_y^c(G_f(\mathbf{x}_i)) \tag{4.5}
$$

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:

$$
w(\mathbf{x}_i^{tr}) = 1 - \tilde{G}_d(G_f(\mathbf{x}_i^{tr})) \tag{4.6}
$$

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:

$$
\begin{aligned}
E_{\tilde{G}_y} = &-\frac{\lambda}{n_{tr}} \sum_{i=1}^{n_{tr}} \sum_{c=1}^{5} [y_{i,c}^{tr} \log(\tilde{G}_y^c(G_f(\mathbf{x}_i^{tr}))) \\
&+ (1 - y_{i,c}^{tr}) \log(1 - \tilde{G}_y^c(G_f(\mathbf{x}_i^{tr})))]
\end{aligned}
\tag{4.7}
$$

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. Therefore, training of the auxiliary discriminator is done as:

$$
\begin{aligned}
E_{\tilde{G}_d} = &- \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \log(\tilde{G}_d(G_f(\mathbf{x}_i^{tr}))) \\
&- \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \log(\tilde{G}_d(G_f(\mathbf{x}_i^{te})))
\end{aligned} \tag{4.8}
$$

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:

$$
\begin{aligned}
(\hat{\theta}_f, \hat{\theta}_y) &= \arg\min_{\theta_f, \theta_y} E_{G_y} - E_{G_d} \\
(\hat{\theta}_d) &= \arg\min_{\theta_d} E_{G_d} \\
(\hat{\theta}_{\tilde{y}}) &= \arg\min_{\theta_{\tilde{y}}} E_{G_{\tilde{y}}} - E_{\tilde{G}_d}
\end{aligned} \tag{4.9}
$$

4.5 Experiments

4.5.1 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 [23], 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 [8], we merge N3 and N4 stages into a single N3 stage. Table (4.1) shows number of sleep stages per class.

Table 4.1: 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

4.5.2 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 $\mathbf{S} \in \mathbb{R}^{F \times T}$ where $F = 129$ (the number of frequency bins), and $T = 29$ (the number of spectral columns).

4.5.3 Network Implementation

For extracting features for the adversarial neural network, we use the same architecture of Biswal et al. [10]. 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 [28] to learn the weights of a deep neural network and hyper-parameter are optimized with importance weighted cross-validation [87]. 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.

4.5.4 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 (4.2). 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.

Figures (4.3) 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.

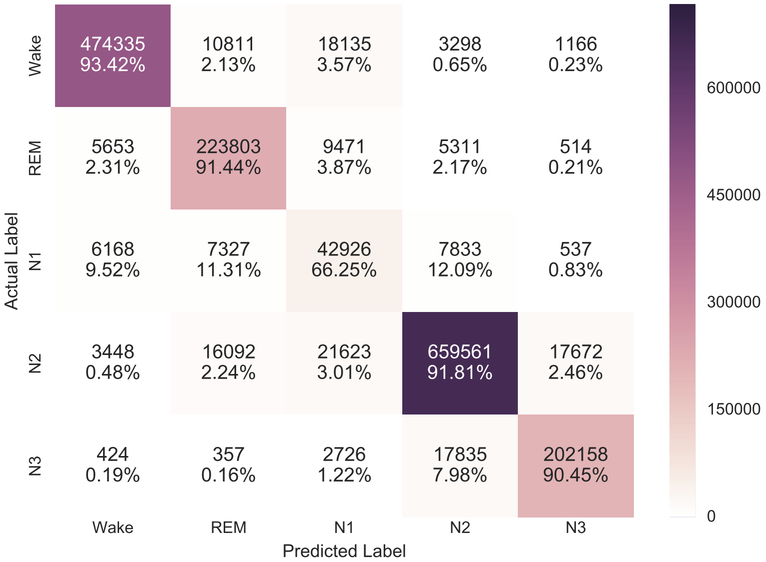
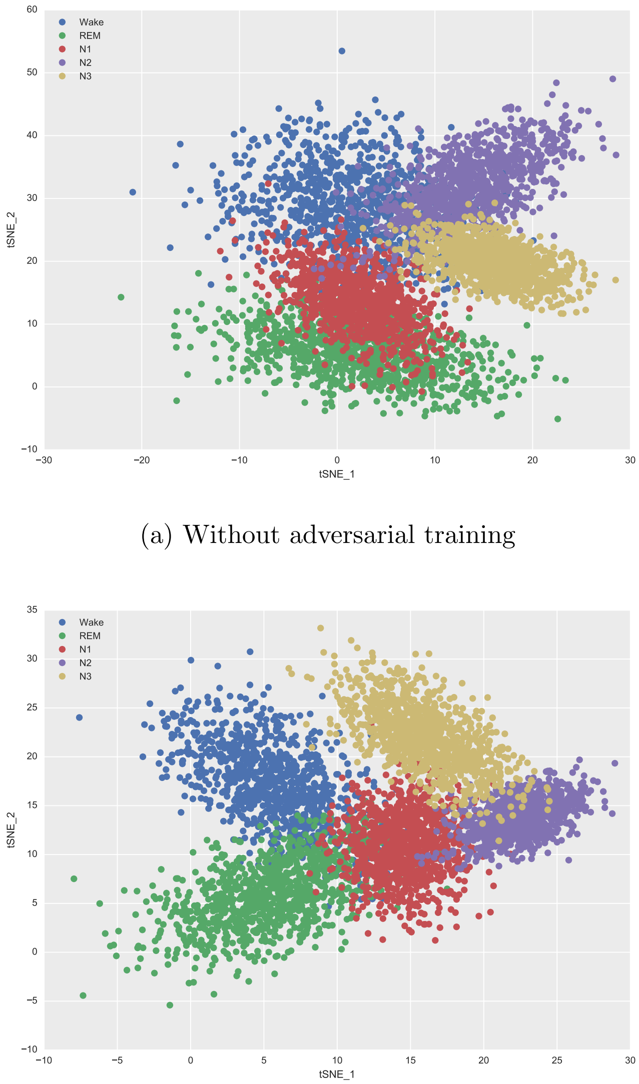


Figure 4.2: Confusion Matrix for test set, which includes 1738 patients from SHHS dataset. The model is trainined on training set, EEGs from 4054 patients. Note: training and test set do not overlap in patients; i.e. cross-subject scenario. “CC BY 4.0”

Table 4.2: 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	-

Table (4.3) summarizes the average results and compares them to the state-of-the-art. The methods proposed previously by Biswal et al. [10], Zhange et al. [111], and Sors et al. [86] 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.



(a) Without adversarial training

(b) With adversarial training

Figure 4.3: 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). “CC BY 4.0”

Based on this experiment, the proposed method boosts the 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. [3] 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 (4.2).

Table 4.3: Performance of class imbalanced model compared to other studies

Method	# patients	precision-Wake	precision-N1	precision-N2	precision-N3	precision-REM	Overall precision	Kappa
[10]	10000	84.5%	56.2%	88.4%	85.4%	92%	81.3%	0.79
[86]	5793	91%	35%	89%	85%	86%	77.2%	0.81
[111]	5793	92%	37%	91%	77%	88%	77%	0.82
Proposed method	5792	97%	45%	95%	91%	86%	82%	0.82

4.6 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.

Chapter 5

Cross-Subject Seizure Detection using Generating Transferable Adversarial Features

5.1 Abstract

Epilepsy is a second common neurological disorder, where neurons produce abnormal signals and cause seizures, affecting 65 million people around the world. Current approaches to developing a generalized automated seizure detection rely on constructing a large labeled training and test corpora by leveraging electroencephalograms (EEGs) from different individuals. However, due to inherent inter-subject variability, heterogeneity of acquisition hardware, different montage choices, and various recording environments, the EEG pattern may have a different distribution. Therefore, training an algorithm on such data without accounting for this diversity can lead to underperformance. Learning a robust representation across the population can be diminished these variabilities. However, these types of methods extract common knowledge from all individuals and suppressing the specific information from each patient, which may

potentially deteriorate the model's adaptability. To this end, this work proposes a novel method to generates transferable features to fill in the gap between features from the training and test sets and adversarially trains the deep classifiers to make consistent predictions over the transferable features. Experiments on an EEG seizure databases show that the proposed method increases the accuracy over state-of-the-art from 86.83% to 91.71% and specificity from 87.38% to 94.73% while reducing the false positive rate/hour from 0.8/hour to 0.58/hour.

5.2 Introduction

Epilepsy is a second common neurological disorder, where neurons produce abnormal signals and cause seizures, affecting 65 million people around the world. An accurate and timely diagnosis for epileptic seizure plays a crucial role to improve the quality of life of epileptic patients. The ground truth for seizure detection remains the multi-lead electroencephalogram (EEG) and manual labeling by an expert, which is costly and time-consuming. In addition to the time and cost involved in manual identifying the onset and end point of each seizure, the significant inter-expert variability remains an issue ([108]). However, the lack of a sizeable public database with heterogeneous populations has limited the development of verifiable algorithms that generalize well across the population. Over the last decade, various methods for seizure detection from EEG signals have been proposed [92, 112, 11]. Traditionally, seizure detection algorithms were deigned patient-specific; training and test data belong to a same subject. Though, patient-specific models can achieve high accuracy, they may have poor performance on the unseen patient or even unseen session data from the same subject. Those methods rely on extracting hand-crafted features from EEG signals, and feeding them to tradition machine learning algorithms, such as support the vector machine (SVM). Recent studies have been developed deep learning (DL) methods to reduce the cost of hand-engineered techniques and enhance the generalization of a model across patients. Developing a model with high generalizability is a challenging task due to the characteristics and complexities of EEG signals. EEG signals have non-stationary nature, which lead to the changes in statistical characteristics of the signal with time. Therefore, a classifier that is trained on a temporally-limited amount of data from an individual may poorly generalize on EEG data recorded at a different time on the same subject. Another issue with the low generalizability issue in EEG data is related to high inherent inter-subject variability in the way an EEG manifests, which limits the usefulness of EEG applications. This phenomenon arises due to physiological differences (e.g. skull shape) between individuals, and 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 across people ([29]). Moreover, electrode sensor montages (the points at which the electrodes are attached and the references points) may differ and different 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' variety of training or even attentiveness on a given day. All these factors lead to significant variabilities in EEG signals, and reduce the usefulness of traditional method in real-life scenario. The primary assumption in machine learning techniques is that training and test data should be drawn from the same distribution, an assumption that does not necessarily hold in large biomedical datasets. In terms of probability theory, one can conclude that these varibilities lead to different joint distribution, $P(X, Y)$ between different recording, where X and Y are feature and label space, respectively. Thus, the transferability of the trained model for an application on unseen subjects is degraded due to having different varibilities.

It has been shown that Symmetric Positive Definite (SPD) matrices provide a strong ability to representations the brain signals ([20, 4]). The covariance matrix is a typical example of SPD matrices, which has been employed in several studies ([78, 75, 50]). These studies showed that using second-order statistics of multi-channel signals reduce inter-subject and intra-subject variabilities between EEG signals. The spatial covariance matrix can well separate useful information about brain functional connectivity structure ([4]) and create 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 ([20]).

Other methods to increase the generalization ability of a model involve transfer learning - finding subsets of past subjects to initialize a classifier for training on a new subject ([109]). Bolagh et al. ([12, 11]) proposed subject-selection and subject clustering to select relevant individuals based on the similarity between the EEG pattern of different individuals. Raza et al. ([74]) proposed bagging methods to handle mismatching between training and test distributions. Chai et al. ([15]) proposed an adaptive subspace feature matching (ASFM) 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.

In the literature, this issue can be partly solved by finding an invariant representation across subjects [92, 24, 42]. DL methods have gained attention these days since they simplify processing pipelines through end-to-end learning, removing the need for domain-specific knowledge for feature engineering. Due to the nature of EEGs, which consist of spatial and temporal information, most convolutional and recurrent processing methodologies are suitable for EEG processing. Besides, their capability to extract robust representations across the population makes them a promising candidate for seizure detection tasks. Besides, the use of generative adversarial networks (GANs) ([34]) to handle inter-subject varibility between training and test sets has received more attention [93, 107, 69, 101] has been gained increasing attention. These methods use adversarial networks for data augmentation purpose, i.e. generate the synthetic features to increase samples' diversity and leverage synthetic and read features to improve the seizure prediction accuracy. Similar to GANs, a two-player minimax game is constructed, in which the first player discriminator between training and test sets and the second player is adversarially trained to deceive the discriminator and extract transferable features ([28]). In these networks, the feature extractor learns patient-invariant feature representations in such way the distance between EEG distributions of training and test is minimized, while the classifier is simultaneously trained on the training set to minimize the training error. However, these types of methods extract common knowledge from all individuals and suppressing the specific information from each patient and distorting the original feature distributions, which may potentially deteriorate the model's adaptability.

This work address this problem and develop a generalized model an adversarial neural network generates transferable features as adversaries to both the classifier and the discriminator between training and test sets. The proposed method fill the gap between EEG patterns from individuals by generating transferable features without performing the adversarial feature adaptation to learn domain-invariant representations. Experimental results on public seizure detection dataset, CHB-MIT [32], significantly boosts performance compared to state-of-the-art deep learning approaches in the cross-subject scenario.

5.3 Method

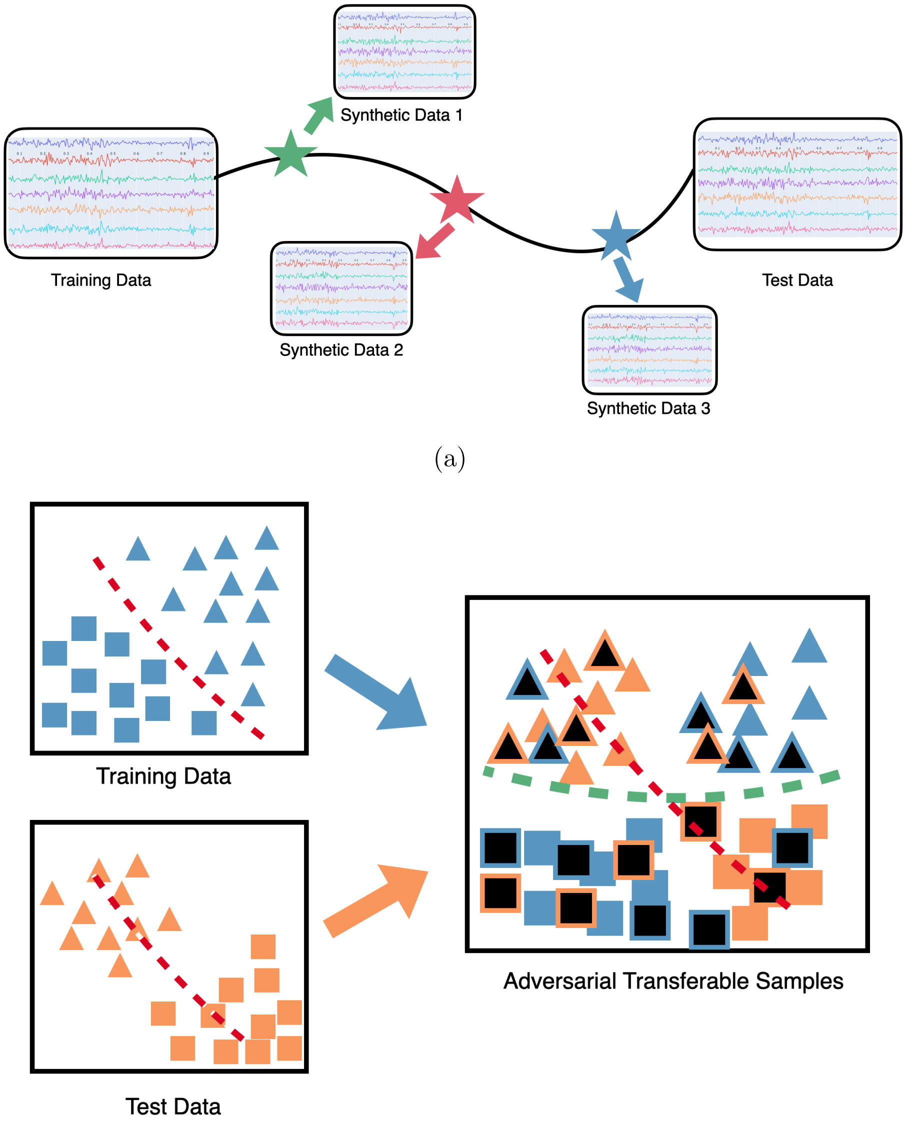
In this paper, we focus on the seizure detection across the population, where the developed model generalize well on unseen patients. The goal is to develop a classifier (network) on a training dataset (labeled domain $\mathcal{D}_{tr}$), which generalizes well on the test dataset from other subjects (unlabeled domain $\mathcal{D}_{te}$) which is not presented in the training. As mentioned, existing adversarial network methods diminish patient-specific variations by learning patient-invariant representations. In adversarial network, the feature extractor and the domain discriminator are defined as $\mathbf{f} = F(\mathbf{x})$, $\mathbf{d} = D(\mathbf{f})$, where D and F form a two-player minimax game. The first player, D, discriminator between training and test sets and the second player, F, is adversarially trained to deceive the discriminator and extract transferable features.

$$
\begin{aligned}
l_d(\theta_D, \mathbf{f}) = &- \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \log[D(\mathbf{f}_{tr}^{(i)})] \\
&- \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \log[D(\mathbf{f}_{te}^{(i)})]
\end{aligned}
\tag{5.1}
$$

By minimizing the cross-entropy loss over training features, the deep classifier is also trained.

$$
l_c(\theta_C, \mathbf{f}) = \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} l_{ce}[C(\mathbf{f}_{tr}^{(i)}, \mathbf{y}_{tr}^{(i)})] \tag{5.2}
$$

To diminish the mismatching of distribution between training and test sets, the transferable features is generated. To generate the transferable features, the gradients of the loss functions l_d and l_c w.r.t. each features ($\{\mathbf{f}_{tr}$, $\mathbf{f}_{te}\}$) are computed. To fill the gap between training and test distributions and push the decision boundary away from real features, the transferable example should confuse the discriminator, D, and the classifier C. Therefore, the transferable features are generated adversarially through a joint loss of l_d and l_c:



(a)

(b)

Figure 5.1: An overview of our approach (a) Showing the idea: fill the gap between training and test set by generating synthetic EEG dataset (b) In the left side, it shows that a model trained only on the training samples is not adaptive to the test samples due to mismatching between distributions and in the right side, using adversarial network, transferable instance is generated between training and test sets and the decision boundary is adapted to both training and test sets through training with transferable examples. "CC BY 4.0"

$$\begin{aligned}\mathbf{f}_{te^{k+1}} \leftarrow \mathbf{f}_{te^k} &+ \beta \nabla_{\mathbf{f}_{te^k}} l_d(\theta_D, \mathbf{f}_{te^k}) \\ &- \gamma \nabla_{\mathbf{f}_{te^k}} l_2(\mathbf{f}_{te^k}, \mathbf{f}_{te^0}) \end{aligned} \tag{5.3}$$

$$\begin{aligned}\mathbf{f}_{tr^{k+1}} \leftarrow \mathbf{f}_{tr^k} &+ \beta \nabla_{\mathbf{f}_{tr^k}} l_d(\theta_D, \mathbf{f}_{tr^k}) \\ &- \gamma \nabla_{\mathbf{f}_{tr^k}} l_2(\mathbf{f}_{tr^k}, \mathbf{f}_{tr^0}) \\ &+ \beta \nabla_{\mathbf{f}_{tr^k}} l_c(\theta_C, \mathbf{f}_{tr^k}) \end{aligned} \tag{5.4}$$

Where K is the number of iterations for generating each transferable feature, and $k = 0, \cdots, K-1$ is the current iteration. In this work, $K = 10$ is a sufficient number of iterations. l_2-distance between the generated features and the original features are controlled to avoid divergence of the generated features.

It is shown that adversarial training enhances local smoothness of the output distributions [61]. Therefore, training classifier with the generated features improves the robustness of the classifier against distribution mismatching, since it should make accurate predictions for the transferable features $\mathbf{f}_{tr}$ in the training set. Moreover, the classifier must make consistent predictions for the transferable features $\mathbf{f}_{te}$ and their original counterparts $\mathbf{f}_{te}$ in the test set. Thus, the loss function for adversarial training of the classifier C can be formulated as follows:

$$\begin{aligned} l_{c,adv}(\theta_C, \mathbf{f}_*) = &\frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} l_{ce}(C(\mathbf{f}_{tr*}^{(i)}), \mathbf{y}_{tr*}^{(i)}) \\ &+ \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} |C((\mathbf{f}_{te*}^{(i)})) - ((\mathbf{f}_{te*}^{(i)}))| \end{aligned} \tag{5.5}$$

To avoid from divergence of the generated features, the domain discriminator is also trained with generated transferable features. Besides, training the classifier and domain discriminator with the generated features can bridge the discrepancy between training and test sets. Therefore, the adversarially training the domain discriminator is done using the following loss:

$$\begin{aligned} l_{d,adv}(\theta_D, \mathbf{f}_*) = &- \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \log[D(\mathbf{f}_{tr*}^{(i)})] \\ &- \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \log[D(\mathbf{f}_{te*}^{(i)})] \end{aligned} \tag{5.6}$$

By jointly minimizing the error in (5.1) , (5.6) with respect to domain discriminator, D and error () and (5.5) with respect to C, the transferable features are generated. The optimization problem for the the proposed method can be formulated as:

$$\begin{aligned} \min_{\theta_D, \theta_C} \quad & l_d(\theta_D, \mathbf{f}) + l_c(\theta_C, \mathbf{f}) \\ & l_{d,adv}(\theta_D, \mathbf{f}_*) + l_{c,adv}(\theta_C, \mathbf{f}_*) \end{aligned} \tag{5.7}$$

The proposed method generates transferable features and adversillay train adapting deep classifiers, as shown in Fig (4.1). The proposed method runs over the features $\mathbf{f}$ and propagates only through the deep classifier C.

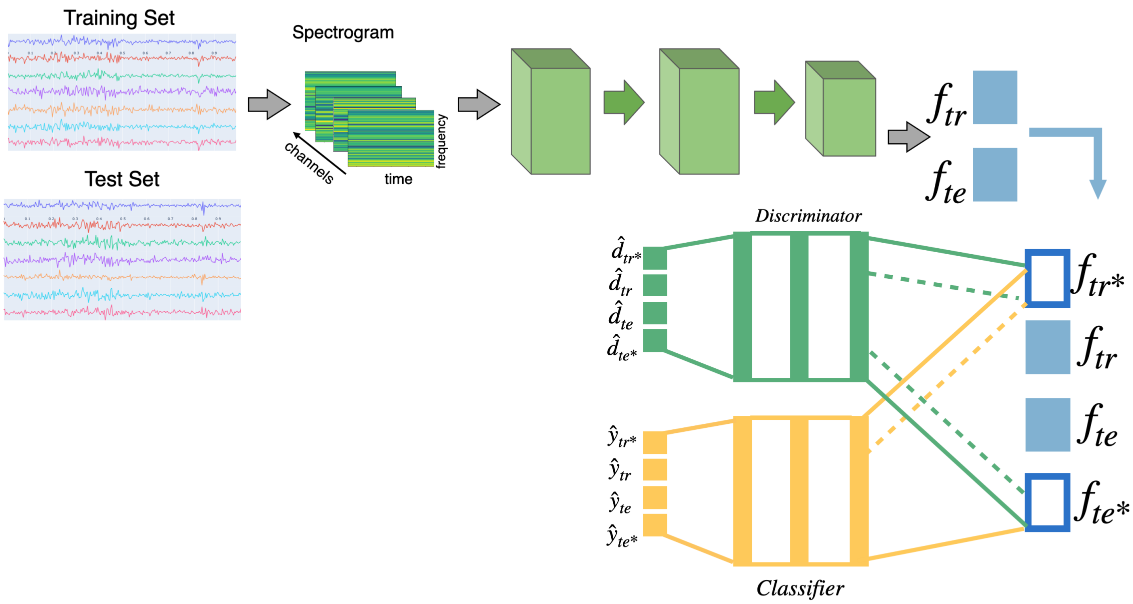


Figure 5.2: The feature extractor F yields representations f_{tr} and f_{te} of the training and test data, which are fixed in the training process to guarantee adaptability. The dashed lines indicate the adversarial generation of transferable examples f_{tr} and f_{te} through maximizing the errors of the category classifier C and domain discriminator D. We adversarially train the classifiers with transferable examples: C to minimize the source error and D to distinguish training from test. “CC BY 4.0”

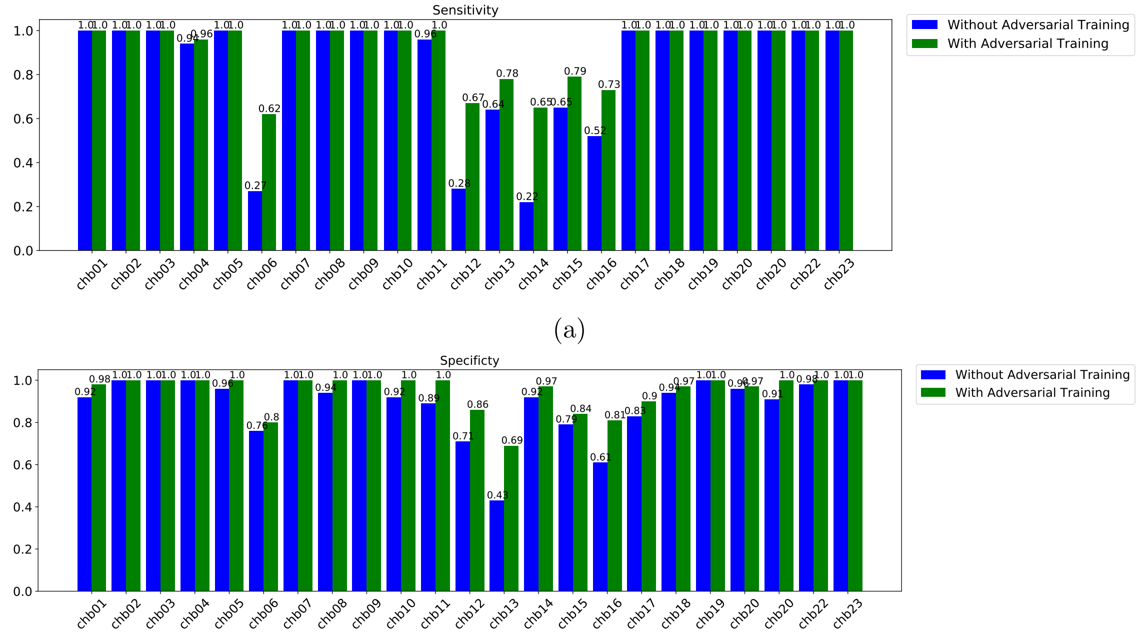
5.4 Experiments

5.4.1 Dataset

For evaulation of the proposed method, EEGs from the Children's Hospital of Boston-Massachusetts Institute of Technology dataset (CHB-MIT) [83], which is a large publicly available data set from the PhysioNet database [32], is used. The database contains 686 sclap EEGs sampled at 256 Hz at 16-bit resolution and stored in EDF file. The database includes EEGs from 23 patients (5 males, 17 females), aged between 1.5 and 22, over 23 channels (the bipolar montage) [83], where 18 channels $\{FP1-F7, F7-T7, T7-P7, P7-O1, FP1-F3, F3-C3, C3-P3, P3-O1, FP2-F4, F4-C4, C4-P4, P4-O2, FP2-F8, F8-T8, T8-P8, P8-O2, FZ-CZ, CZ-PZ\}$ are common among patinets. The length of most of recording are one hour, however recordings for case 10 are two hours long and those belonging to cases 4, 6, 7, 9, and 23 are four hours long. Seizures onset and offset are defined in second in recordings which contain seizures. From the 686 records, 198 records contain seizures. After preprocessing EEG signals with filtering signals with a band-pass filter (0.5-180 Hz), normalizing EEG amplitudes using min-max normalization across subjects, each recording was divided into four second epochs (1024 time samples). Therefore the dimension of data after preprocessing part is (18×1024) and each segment is classified as either dominantly seizure or non-seizure (using expert labels). Similar to [73], spectrograms are generated for each EEG channel to transform data to the time-frequency domain. Each 4-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. To extract original feature representations, a CNN network similar to [92], which contains three convolutional layers with BatchNorm, is used. Adam optimization with initail learning rate $\eta = 10^{-4}$ is used for training the network. The learning rate changes by $\eta_p = \frac{\eta_0}{\eta_0(1+wp)^{\phi}}$, $w = 10$, $\phi = 0.80$, and p is the progress ranging from 0 to 1.

Table (5.1) provides the per-patient (LOSO-CV) results, which shows that the proposed method exceeds the performance in cross-subject scenario. It shows that augmenting transferable features fill the gap between the training and test distributions, where mitigate the imbalance issue in the datasets.

Table (5.2) summarizes the average results and compares them to the state-of-the-art, where Chen *et al.* [18] and Thodoroff *et al.* used [92] the wavelet transformation and deep learning (CNN followd by RNN), respectively. It shows an increase over previous works in accuracy and specificity by 4-5% in terms of sensitivity. It should note that subject-specific works are not included in this comparison, since when the main focus of this work is developing a generalized model across the population. We also note that we improve the false positive rate from 1.7/hour to 0.58/hour over Shoeb's original work [83]. To the best of our knowledge, the method described in this article is the first work to propose filling the gap between individuals using generating transferable adversarial samples for cross-subject seizure detection. Adversarial training has some advantages which helps us to create a generalized model: training a network with generated transferable features can be seen as a regularizing the network to avoid over-fitting. Besides, Sinha et al. (2018) proved robustness guarantees with adversarial training.



(a)

(b)

Figure 5.3: Classification Performance with/without Adversarial Training (a) Sensitivity (b) Specificity; where the blue bars shows the classification performance using feeding spectogram to CNN network followed by softmax layer, and green bars shows the classification performance with the proposed method shown in Fig (5.2). "CC BY 4.0"

Table 5.1: Performance on the CHB-MIT dataset

Subject ID	**Accuracy (%)**	**Sensitivity (%)**	**False Positive rate (seizures/h)**	**Latency (sec)**
1	95.33	100.00	0.17	4.70
2	86.47	100.00	0.23	7.13
3	96.50	100.00	0.10	2.11
4	94.41	96.15	0.12	6.82
5	96.25	100.00	0.18	4.12
6	83.79	62.31	1.12	2.64
7	88.56	100.00	1.07	5.17
8	90.89	100.00	0.30	3.42
9	98.41	100.00	0.86	7.88
10	94.80	100.00	0.92	2.34
11	96.39	100.00	0.24	2.29
12	86.13	67.92	1.91	5.12
13	92.62	78.19	2.58	7.88
14	86.73	65.23	0.35	3.39
15	89.13	79.81	0.10	5.69
16	87.40	73.39	1.37	2.93
17	93.84	100.00	0.62	5.87
18	87.40	100.00	0.36	4.42
19	93.84	100.00	0.12	9.78
20	95.76	100.00	0.22	2.37
21	95.62	100.00	0.13	2.59
22	89.28	100.00	0.19	11.82
23	94.76	100.00	0.06	1.04
mean±(std)	91.71	86.09±(17.98)	0.58±(0.66)	4.85±(2.71)

Table 5.2: Performance comparison of works on the CHB-MIT database

Method	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**FPR**	**Latency(sec)**
Chen et al [18]	86.83%	85.29%	87.38%	–	–
Thodoroff et al [92]	84.18%	85.16%	83.21%	0.88	–
Bolagh et al [11]	89.84%	85.77%	89.64%	0.77	5.24
Proposed method	**91.71%**	**91.09%**	**94.73%**	**0.58**	**4.45**

5.5 Conclusion

In this work, adversarial training with generated transferable features and original features was proposed for the cross-subject seizure detection task across a heterogeneous dataset, which includes EEGs from 23 patients. Inherent inter-subject variability, electrode misplacement, and heterogeneity in the medical history of patients in the population may lead to an algorithm having poor generalization across subjects. 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 a framework for generating transferable features to adapting deep classifiers. By deceiving both classifier and domain discriminator, the proposed method generates transferable features which fill the gap between subjects' joint distributions. The proposed method achieves state-of-the-art performance (without prior knowledge) on CHB-MIT dataset. 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 different biomedical task; i.g, sleep staging using EEGs.

Chapter 6

Conclusion

6.1 Summary and contributions

The work presented in this thesis addressed the issue of fusing multichannel and multimodal physiological data for the classification and prediction of critical events in brain health. Theoretically, this thesis focuses on methods to improve the robustness of machine learning approaches to deal simultaneously with intra- and inter-subject variation in data due to differences in patient brain morphology and clinical recording practices. In particular, the work in tihs thesis focuses on multi-lead EEG signals, which are considered the most useful signals for recognizing neural brain pathology. Developing automated EEG analysis algorithms with high generalizability to other patients is a challenge due to the nature of EEG signals and different variations across datasets and populations. The main issue is high inherent inter-subject variability in the way EEG manifests at surface electrodes. This variability arises due to physiological differences (e.g. skull shape and brain morphology) between individuals. The fact that neural activity propagation is specific to individuals is because cortical folding, tissue conductivity, and tissue shapes of brains differ [53, 82]. Additionally, when electrodes are applied, small differences in the locations on the skull are inevitable, even when the same individual places the same electrodes on the same subject [109, 105, 39]. Differences in hardware (or software) filter choices, and properties or quality of electrodes, introduce further differences. Finally, electrode sensor montages (standardized points at which the electrodes are attached, and the reference locations) often deliberately differ between studies. this can be because of the hardware used (with an under-complete set of channels), because of different training practices, or because of experiential preferences. All these factors lead to significant variability in EEG signals for individuals over time, between individuals, and between databases.

The aim of this work was to develop generalizable machine learning methods for electrophysiology. Chapters 2 proposed a technique for measuring the boosting effect of machine learning and deep learning methods in a non-Euclidean space, specifically a Riemannian manifold, to mitigate the effects of intra- and inter-subject variability in seizure detection and sleep staging. Given the abundance of labeled data from other subjects, it is tempting to use them for training a generalized classifier. Nevertheless, not all such training subjects may improve the performance on a given test subject because of inherent inter-subject variability [12]. We hypothesized that by sub-grouping patients, the diversity of the EEG data is diminished, and the trained model's transferability is increased. In the cross-subject classification task, inter-subject variability leads to different probability distributions between individuals, and consequently poor generalization across subjects. Potentially, individuals with different biological phenotypes would provide enough diversity in the dataset, but achieving this would require vastly more high-quality labelled data than is currently available to a single researcher, and a single network cannot therefore be robust to such variabilities. By sub-grouping the population, we attempted to diminish the diversity in EEG signals and create homogeneous cohorts in the dataset. Therefore, a clustering technique on the feature space, robust to intra- and inter-subject variability, was proposed to find groups of similar individuals. Evaluation of the proposed method on the 2018 PhysioNet Challenge data [32, 30], for five-class sleep staging, demonstrates that population sub-grouping significantly improved the results compared to no sub-grouping. Results on the CHB-MIT database [32, 83], for seizure detection, also confirmed that population sub-grouping significantly improved the results compared to the same classification approach with no sub-grouping.

Recently, it has been shown that adversarial networks have a high capability to extract robust representation from large images database. Chapter 3 investigates the use of adversarial networks for the largest sleep staging databases. Then, an adversarial network with two attention mechanisms was developed to replicate the manner in which clinicians perform sleep staging. The first attention mechanism attends more to channels that are important across datasets and individuals. The second attention mechanism pays more attention to segments of data that are more transferable across the population. Using two large public EEG databases, the SHHS and the PhysioNet 2018 Challenge [32, 30], comprising over 50,000 hours of multichanel data from over 7,700 individuals, we demonstrate that adversarially learning a network with attention mechanisms significantly boosts performance compared to state-of-the-art deep learning approaches on the cross-dataset scenario.

Chapter 4 describes the development of a method for using an adversarial network with an importance weighting framework to assign a weight for each feature 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. Again, using the SHHS data, 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 for the five-class sleep staging task.

Finally, chapter 5 presents a novel method for seizure classification, which generates transferable features to fill in the gap between features from the training and test sets and adversarially trains the deep classifiers to make consistent predictions over the transferable features. Experiments on an EEG seizure database, the CHB-MIT database [32], show that the proposed method increases the accuracy over state-of-the-art from 86.83% to 91.71% and specificity from 87.38% to 94.73% while reducing the false positive rate from 0.8/hour to 0.58/hour for the seizure detection task.

6.2 Limitations

The work presented in this thesis has some limitations. The methods presented in Chapters 2 to 5 were developed and tested offline. The methods were not tested online at the point of care because this work focused on developing generalizable machine learning approach, which tackles different variabilities in a large dataset.

The functionalities presented in this work are based on high-quality input data, i.e., high signal to noise ratio (SNR). However, EEGs are sometimes contaminated with different types of noises; therefore, different denoising methods need to be evaluated in the context of the proposed algorithms.

Moreover, noisy labeling, which may cause by intra- and inter-expert variability, is not considered in this work. However, this issue is common in biomedical data and may lead to the underperformance of automatic models.

The use of deep networks that are complex and require large quantities and varieties of labeled data/individuals and time for training. Therefore, they may not be useful for mobile applications at this moment in time. Although computational cost was considered as a factor in the development, for resources are not infinite, it will be necessary to transfer the algorithms to a portable devices to assess the real contribution to sleep staging and seizure-detection tasks in real-life applications beyond the high intensity clinical environment.

Finally, it should be noted that in this thesis, the proposed methods were evaluated only on EEGs and for two tasks; sleep staging and seizure detection. Although the applicability of the methods beyond these domains is likely (e.g. to cardiac arrhythmia classification from multilead electrocardiograms, for example), this has yet to be demonstrated.

6.3 Future work

This work helps support the development of generalizable approaches to enable adaptation of population-trained machine learning models to patient-specific or situation-specific models and may play a pivotal role in catalyzing precision health through personalized machine learning models that account for the topography of the individual's brain, and the way electrodes are placed. Individual brain activity can then be rapidly diagnosed using low-cost wearables. The techniques could generalize beyond EEG to other multi-lead modalities such as the electrocardiogram, electromyogram, and ultrasound arrays, for example. They could apply to areas such as fetal monitoring, cardiac arrhythmia diagnostics, and blood flow imaging. Therefore, as future work, we will extend our methods on different modalities and tasks to evaluate how they generalize them across tasks and modalities, beginning with electrocardiograms.

Moreover, it has been shown that using graphical analysis to capture the structural and functional connectives would be useful for predicting outcomes of antidepressant treatment. Using graphical analysis allows us to capture the non-Euclidean information in the EEGs and spectral information, important nodes (channels), and causality between different regions of the brain simultaneously. One of our future works will focus on developing a method for capturing temporal-spectral-spatial-causality information in a unified framework within, and between modalities.

Additionally, we will extend our methods on the multi-task learning scenario of classifying classify sleep, sedation, and pathologies (e.g., seizures) simultaneously. It is known that deep networks trained on large-scale data can learn transferable features to promote learning multiple tasks. Since deep features eventually transition from general to specific along with deep networks, a fundamental problem of multi-task learning is how to exploit the task relatedness underlying parameter tensors and improve feature transferability in the multiple task-specific layers. One approach is to discover the task relationships and jointly learn transferable features and multilinear relationships of tasks and features using deep networks and adversarial networks. As future work, we will develop methods to learn transferable and discriminative representations in multi-task learning scenarios of classifying classify sleep, sedation, and pathologies (e.g., seizures) simultaneously.

Recent studies [55, 79, 110, 14] investigated the concept of partial domain adaptation, relaxing the assumption that label sets are identical across training and test sets. In their methods, they assumed that the training label set contains the test label set while Busto et al. [68] introduced "unknown" classes in both training and test sets, and assumed common classes between two sets were known in the training phase. However, practical scenarios, where two datasets may not have similar label sets, are more complicated and common for biomedical applications. There are two major challenges in the analysis of biomedical datasets: (1) inherent intra- and inter-subject variability, electrode misplacement, hardware acquisition, leading to a large domain gap; and (2) different patterns in training and test labels, such as sleep and seizure, leading to a large category gap. In summary, the relationship of label sets between the training and test sets is unknown in the presence of a large domain gap. If the training label set is large enough to contain the test label set, partial domain adaptation methods are good choices. However, if the training label set is contained in the test label set or common classes are known, open set domain adaptation methods are good choices. As a future work, we will develop suitable partial domain adaptation methods for biomedical applications, where two datasets may not have the same label sets.

It is well known that deep learning approaches need abundant labeled data. In medical applications, labeling a dataset requires domain expertise, which may lead to the significant inter-expert variability in labeling [108, 13, 13, 102]. Therefore, label noise may manifest itself in various forms. Recent studies in computer vision applications [77, 91, 88, 33, 63] show that deep learning model performance can significantly decrease due to the negative effect of label noise. Studies have shown that the negative impact of label noise can be worse than the feature noise [114, 27]. There are different techniques to combat this issue, such as pre-processing and label cleaning [67, 48, 66], considering a robust loss function to label noise [31, 43, 98], exploiting data similarity to identify incorrect labels [22, 99, 84], or using probabilistic graphical models [106, 60, 95]. Recent studies [64, 59, 70, 19, 70] categorize label noise into three different types:

- Class-independent
- Class-dependent
- Class and feature dependent

Voting among prediction outputs of classifiers is a popular approach to reducing the effect of labeled noise in the training step. Other studies consider label noise as an outlier-detection problem and try to detect mislabeled samples based on K-nearest neighbors (KNN) [103, 104]. Another work focused on determining the effect of outlier samples on the classification error and cleaning labels by removing/reducing weights in those instances. Future work will investigate different techniques such as combining weak label boosting methodologies to handle the label noise problem in enormous physiological datasets, principally applied to seizures, sleep, and identifying brain states where inter-expert variability may be significant.

In summary, having addressed the errors created at acquisition in this thesis, the next step is to address the errors introduced at the labelling stage.

Bibliography

[1] Epilepsy Data and Statistics. `https://www.cdc.gov/epilepsy/data/index.html`. Accessed: 2019-01-25.

[2] 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.

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

[4] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten. Common Spatial Pattern Revisited by Riemannian Geometry. In *2010 IEEE International Workshop on Multimedia Signal Processing*, pages 472–476. IEEE, 2010.

[5] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten. Multiclass brain–computer interface classification by Riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928, 2012.

[6] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten. Multiclass brain–computer interface classification by Riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928, 2012.

[7] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten. Classification of covari-

ance matrices using a Riemannian-based kernel for BCI applications. *Neurocomputing*, 112:172–178, 2013.

[8] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. Marcus, B. V. Vaughn, 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, 2012.

[9] R. B. Berry, R. Budhiraja, D. J. Gottlieb, D. Gozal, C. Iber, V. K. Kapur, C. L. Marcus, R. Mehra, S. Parthasarathy, S. F. Quan, et al. Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events: deliberations of the sleep apnea definitions task force of the american academy of sleep medicine. *Journal of Clinical Sleep Medicine*, 8(5):597–619, 2012.

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

[11] S. N. G. Bolagh, G. Clifford, et al. Subject Selection on a Riemannian Manifold for Unsupervised Cross-Subject Seizure Detection. *arXiv preprint arXiv:1712.00465*, 2017.

[12] S. N. G. Bolagh, M. B. Shamsollahi, C. Jutten, and M. Congedo. 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.

[13] V. Bolón-Canedo, E. Ataer-Cansizoglu, D. Erdogmus, J. Kalpathy-Cramer, O. Fontenla-Romero, A. Alonso-Betanzos, and M. F. Chiang. Dealing with

Inter-Expert Variability in Retinopathy of Prematurity: A Machine Learning Approach. *Computer Methods and Programs in Biomedicine*, 122(1):1–15, 2015.

[14] Z. Cao, M. Long, J. Wang, and M. I. Jordan. Partial transfer learning with selective adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2724–2732, 2018.

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

[16] C.-C. Chang and C.-J. Lin. LIBSVM: a library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[17] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[18] D. Chen, S. Wan, J. Xiang, and F. S. Bao. A high-performance seizure detection algorithm based on Discrete Wavelet Transform (DWT) and EEG. *PloS ONE*, 12(3):e0173138, 2017.

[19] J. Cheng, T. Liu, K. Ramamohanarao, and D. Tao. Learning with bounded instance-and label-dependent label noise. *arXiv preprint arXiv:1709.03768*, 2017.

[20] M. Congedo, A. Barachant, and R. Bhatia. Riemannian geometry for eeg-based brain-computer interfaces; a primer and a review. *Brain-Computer Interfaces*, pages 1–20, 2017.

[21] H. Danker-hopfe, P. Anderer, J. Zeitlhofer, M. Boeck, H. Dorn, G. Gruber, E. Heller, E. Loretz, D. Moser, S. Parapatics, et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *Journal of Sleep Research*, 18(1):74–84, 2009.

[22] A. Drory, O. Ratzon, S. Avidan, and R. Giryes. The Resistance to Label Noise in KNN and CNN Depends on its Concentration. *arXiv preprint arXiv:1803.11410*, 2018.

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

[24] A. Emami, N. Kunii, T. Matsuo, T. Shinozaki, K. Kawai, and H. Takahashi. Seizure detection by convolutional neural network-based analysis of scalp electroencephalography plot images. *NeuroImage: Clinical*, 22:101684, 2019.

[25] I. Exarchos, A. A. Rogers, L. M. Aiani, R. E. Gross, G. D. Clifford, N. P. Pedersen, and J. T. Willie. Supervised and unsupervised machine learning for automated scoring of sleep-wake and cataplexy in a mouse model of narcolepsy. *Sleep*, 2019.

[26] F. Fahimi, Z. Zhang, W. B. Goh, T.-S. Lee, K. K. Ang, and C. Guan. Inter-subject transfer learning with an end-to-end deep convolutional neural network for EEG-based BCI. *Journal of Neural Engineering*, 16(2):026007, 2019.

[27] B. Frénay and M. Verleysen. Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2013.

[28] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette,

M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

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

[30] M. M. Ghassemi, B. E. Moody, L.-W. H. Lehman, C. Song, Q. Li, H. Sun, R. G. Mark, M. B. Westover, and G. D. Clifford. You Snooze You Win - The PhysioNet Computing in Cardiology Challenge 2018. In *2018 Computing in Cardiology Conference (CinC)*, volume 45, pages 1–4. IEEE, 2018.

[31] A. Ghosh, H. Kumar, and P. Sastry. Robust loss functions under label noise for deep neural networks. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[32] 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*, 101(23):e215–e220, 2000.

[33] J. Goldberger and E. Ben-Reuven. Training deep neural-networks using a noise adaptation layer. 2016.

[34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[35] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 131–135. IEEE, 2017.

[36] U. Herwig, P. Satrapi, and C. Schönfeldt-Lecuona. Using the International 10-20 EEG system for positioning of transcranial magnetic stimulation. *Brain Topography*, 16(2):95–99, 2003.

[37] M. Howe-Patterson, B. Pourbabaee, and F. Benard. Automated detection of sleep arousals from polysomnography data using a dense convolutional neural network. In *2018 Computing in Cardiology Conference (CinC)*, volume 45, pages 1–4. IEEE, 2018.

[38] H. Huang and M. Ding. Linking functional connectivity and structural connectivity quantitatively: a comparison of methods. *Brain Connectivity*, 6(2):99–108, 2016.

[39] Y. Huang, J. Zhang, Y. Cui, G. Yang, L. He, Q. Liu, and G. Yin. How different EEG references influence sensor level functional connectivity graphs. *Frontiers in Neuroscience*, 11:368, 2017.

[40] Z. Huang and L. Van Gool. A Riemannian Network for SPD Matrix Learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[41] Z. Huang, R. Wang, S. Shan, X. Li, and X. Chen. Log-Euclidean Metric Learning on Symmetric Positive Definite Manifold with Application to Image Set Classification. In *International Conference on Machine Learning*, pages 720–729, 2015.

[42] R. Hussein, H. Palangi, R. Ward, and Z. J. Wang. Epileptic seizure detection: a deep learning approach. *arXiv preprint arXiv:1803.09848*, 2018.

[43] H. Izadinia, B. C. Russell, A. Farhadi, M. D. Hoffman, and A. Hertzmann. Deep classifiers from image tags in the wild. In *Proceedings of the 2015 Workshop on Community-Organized Multimodal Mining: Opportunities for Novel Solutions*, pages 13–18, 2015.

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

[45] D. Jiang, M. Yu, and W. Yuanyuan. Sleep stage classification using covariance features of multi-channel physiological signals on Riemannian manifolds. *Computer Methods and Programs in Biomedicine*, 178:19–30, 2019.

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

[47] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[48] K.-H. Lee, X. He, L. Zhang, and L. Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5447–5456, 2018.

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

[50] Y. Li, K. M. Wong, and H. de Bruin. Electroencephalogram signals classification for sleep-state decision–a Riemannian geometry approach. *IET Signal Processing*, 6(4):288–299, 2012.

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

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

[53] A. Llera, T. Wolfers, P. Mulders, and C. F. Beckmann. Inter-individual differences in human brain structure and morphology link to variation in demographics and behavior. *Elife*, 8:e44443, 2019.

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

[55] Z. Luo, Y. Zou, J. Hoffman, and L. F. Fei-Fei. Label efficient learning of transferable representations acrosss domains and tasks. In *Advances in Neural Information Processing Systems*, pages 165–177, 2017.

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

[57] L. v. d. Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

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

[59] A. K. Menon, B. Van Rooyen, and N. Natarajan. Learning from binary labels with instance-dependent noise. *Machine Learning*, 107(8-10):1561–1595, 2018.

[60] I. Misra, C. Lawrence Zitnick, M. Mitchell, and R. Girshick. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2930–2939, 2016.

[61] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[62] S. Mousavi, F. Afghah, and U. R. Acharya. Sleepeegnet: Automated sleep stage scoring with sequence to sequence deep learning approach. *PloS one*, 14(5), 2019.

[63] R. Müller, S. Kornblith, and G. E. Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, pages 4694–4703, 2019.

[64] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, pages 1196–1204, 2013.

[65] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856, 2002.

[66] C. G. Northcutt, T. Wu, and I. L. Chuang. Learning with confident examples: Rank pruning for robust classification with noisy labels. *arXiv preprint arXiv:1705.01936*, 2017.

[67] P. Ostyakov, E. Logacheva, R. Suvorov, V. Aliev, G. Sterkin, O. Khomenko, and S. I. Nikolenko. Label denoising with large ensembles of heterogeneous neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[68] P. Panareda Busto and J. Gall. Open Set Domain Adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 754–763, 2017.

[69] D. Pascual, A. Aminifar, D. Atienza, P. Ryvlin, and R. Wattenhofer. Synthetic epileptic brain activities using generative adversarial networks. *arXiv preprint arXiv:1907.10518*, 2019.

[70] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1944–1952, 2017.

[71] 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. In *Advances in Neural Information Processing Systems*, pages 4415–4426, 2019.

[72] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, 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*, 27(3):400–410, 2019.

[73] S. Raghu, N. Sriraam, Y. Temel, S. V. Rao, and P. L. Kubben. EEG based multi-class seizure type classification using convolutional neural network and transfer learning. *Neural Networks*, 124:202–212, 2020.

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

[75] P. Rodrigues, M. Congedo, and C. Jutten. A Data Imputation Method for Matrices in the Symmetric Positive Definite Manifold. 2019.

[76] P. Rodrigues, M. Congedo, and C. Jutten. "When does it work?": An exploratory analysis of transfer learning for BCI. 2019.

[77] D. Rolnick, A. Veit, S. Belongie, and N. Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017.

[78] E. Saifutdinova, M. Congedo, D. Dudysova, L. Lhotska, J. Koprivova, and V. Gerla. An Unsupervised Multichannel Artifact Detection method for Sleep EEG Based on Riemannian Geometry. *Sensors*, 19(3):602, 2019.

[79] K. Saito, S. Yamamoto, Y. Ushiku, and T. Harada. Open set domain adaptation by backpropagation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 153–168, 2018.

[80] S. Sanei and J. A. Chambers. *EEG Signal Processing.* John Wiley & Sons, 2013.

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

[82] M. L. Seghier and C. J. Price. Interpreting and utilising inter–subject variability in brain function. *Trends in Cognitive Sciences*, 22(6):517–530, 2018.

[83] A. H. Shoeb. *Application of machine learning to epileptic seizure onset detection and treatment.* PhD thesis, Massachusetts Institute of Technology, 2009.

[84] J. Shu, Q. Xie, L. Yi, Q. Zhao, S. Zhou, Z. Xu, and D. Meng. Meta-Weight-Net: Learning an Explicit Mapping for Sample Weighting. In *Advances in Neural Information Processing Systems*, pages 1919–1930, 2019.

[85] S. Sivaranjini and C. Sujatha. Deep learning based diagnosis of Parkinson's disease using convolutional neural network. *Multimedia Tools and Applications*, pages 1–13, 2019.

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

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

[88] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*, 2014.

[89] M. E. Tagluk, N. Sezgin, and M. Akin. Estimation of sleep stages by an artificial neural network employing EEG, EMG and EOG. *Journal of Medical Systems*, 34(4):717–725, 2010.

[90] X. Tang and X. Zhang. Conditional Adversarial Domain Adaptation Neural Network for Motor Imagery EEG Decoding. *Entropy*, 22(1):96, 2020.

[91] K. K. Thekumparampil, A. Khetan, Z. Lin, and S. Oh. Robustness of conditional gans to noisy labels. In *Advances in Neural Information Processing Systems*, pages 10271–10282, 2018.

[92] P. Thodoroff, J. Pineau, and A. Lim. Learning robust features using deep learning for automatic seizure detection. In *Machine Learning for Healthcare Conference*, pages 178–190, 2016.

[93] N. D. Truong, L. Zhou, and O. Kavehei. Semi-supervised seizure prediction with generative adversarial networks. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2369–2372. IEEE, 2019.

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

[95] A. Vahdat. Toward robustness against label noise in training deep discriminative neural networks. In *Advances in Neural Information Processing Systems*, pages 5596–5605, 2017.

[96] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[97] U. Von Luxburg. A Tutorial on Spectral Clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[98] X. Wang, Y. Hua, E. Kodirov, and N. M. Robertson. IMAE for Noise-Robust Learning: Mean Absolute Error Does Not Treat Examples Equally and Gradient Magnitude's Variance Matters. *arXiv preprint arXiv:1903.12141*, 2019.

[99] X. Wang, E. Kodirov, Y. Hua, and N. M. Robertson. Derivative Manipulation for General Example Weighting. *arXiv preprint arXiv:1905.11233*, 2019.

[100] X. Wang, L. Li, W. Ye, M. Long, and J. Wang. Transferable attention for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5345–5352, 2019.

[101] Z. Wei, J. Zou, J. Zhang, and J. Xu. Automatic epileptic EEG detection using convolutional neural network with improvements in time-domain. *Biomedical Signal Processing and Control*, 53:101551, 2019.

[102] S. L. Wendt, P. Welinder, H. B. Sorensen, P. E. Peppard, P. Jennum, P. Perona,

E. Mignot, and S. C. Warby. Inter-Expert and Intra-Expert Reliability in Sleep Spindle Scoring. *Clinical Neurophysiology*, 126(8):1548–1556, 2015.

[103] D. R. Wilson and T. R. Martinez. Instance pruning techniques. In *ICML*, volume 97, pages 400–411, 1997.

[104] D. R. Wilson and T. R. Martinez. Reduction techniques for instance-based learning algorithms. *Machine Learning*, 38(3):257–286, 2000.

[105] J. Wolpaw and E. W. Wolpaw. *Brain-computer interfaces: principles and practice*. OUP USA, 2012.

[106] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2691–2699, 2015.

[107] S. You, B. H. Cho, S. Yook, J. Y. Kim, Y.-M. Shon, D.-W. Seo, and I. Y. Kim. Unsupervised automatic seizure detection for focal-onset seizures recorded with behind-the-ear EEG using an anomaly-detecting generative adversarial network. *Computer Methods and Programs in Biomedicine*, page 105472, 2020.

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

[109] P. Zanini, M. Congedo, C. Jutten, S. Said, and Y. Berthoumieu. Transfer learning: a Riemannian geometry framework with applications to Brain-Computer Interfaces. *IEEE Transactions on Biomedical Engineering*, 2017.

[110] J. Zhang, Z. Ding, W. Li, and P. Ogunbona. Importance weighted adversarial nets for partial domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8156–8164, 2018.

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

[112] T. Zhang, W. Chen, and M. Li. Generalized Stockwell transform and SVD-based epileptic seizure detection in EEG using random forest. *Biocybernetics and Biomedical Engineering*, 38(3):519–534, 2018.

[113] M. Zhou, C. Tian, R. Cao, B. Wang, Y. Niu, T. Hu, H. Guo, and J. Xiang. Epileptic Seizure Detection based on EEG signals and CNN. *Frontiers in Neuroinformatics*, 12:95, 2018.

[114] X. Zhu and X. Wu. Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review*, 22(3):177–210, 2004.