OUKA
The University of Osaka
Institutional Knowledge Archive

Title	Building Robust Deep Models of Stress-Related Psycho-Physiological Signals
Author(s)	Hagad, Juan Lorenzo Mutia
Citation	大阪大学, 2021, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/85429
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

The University of Osaka

Building Robust Deep Models of Stress-Related Psycho-Physiological Signals

Submitted to

Graduate School of Information Science and Technology

Osaka University

July 2021

Juan Lorenzo HAGAD

Research Output

Journal Publications

- Juan Lorenzo Hagad, Tsukasa Kimura, Ken-ichi Fukui, and Masayuki Numao. "Learning Subject-Generalized Topographical EEG Embeddings Using Deep Variational Autoencoders and Domain-Adversarial Regularization", Sensors, 21(5):1792, 2021.

International Conference/Symposium (with review)

- Juan Lorenzo Hagad, Ken-ichi Fukui, and Masayuki Numao. "Deep Visual Models for EEG of Mindfulness Meditation in a Workplace Setting." In Precision Health and Medicine, Post Proc. International Workshop on Health Intelligence (W3PHIAI2019) held in conjunction with AAAI2019, pp. 129-137. Springer, Cham, 2019.

- Hayato Uraji, Taweesak Emsawas, Juan Lorenzo Hagad, Ken-ichi Fukui, and Masayuki Numao. "Analyzing the Effect of Video Media on Emotion Using a VR Headset Platform and Physiological Data.", In Theory and Practice of Computation: Proceedings of the Workshop on Computation: Theory and Practice (WCTP 2018), September 17-18, 2018, Manila, The Philippines, p. 151. CRC Press, 2019.

- Jian Yang, Satoru Iwasaki, Anthony Abraham, Juan Lorenzo Hagad, Takuya Obuchi, and Tadashi Nakano. "Performance Optimization of Self-organizing Bioparticles for Multi-Target Detection and Gravitation Problems", In 10th EAI International Conference on Bio-Inspired Information and Communications Technologies, p. 125. European Alliance for Innovation (EAI), 2017.

- Nattapong Thammasan, Juan Lorenzo Hagad, Ken-ichi Fukui, and Masayuki Numao. "Multimodal Stability-Sensitive Emotion Recognition Based on Brainwave and Physiological Signals.", In 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), pp. 44-49. IEEE, 2017.

International Workshop (with review)

- Juan Lorenzo Hagad, Ken-ichi Fukui, and Masayuki Numao. "Multimodal Latent Feature Learning for Psycho-Physiological Stress Modelling and Detection", In Proc. 7th International Workshop on Empathic Computing (IWEC 2016), in conjunction with PRICAI 2016, Phuket, Thailand, Aug. 2016.

Domestic Workshop (with review)

- Juan Lorenzo Hagad, Ken-ichi Fukui, and Masayuki Numao. "Modelling Naturalistic Work Stress Using Spectral HRV Representations and Deep Learning", In: Ohsawa Y. et al. (eds) Advances in Artificial Intelligence. JSAI 2019. Advances in Intelligent Systems and Computing, vol 1128. Springer, Cham., 2019.

Abstract

Recent years have witnessed the rapid proliferation of wearable technologies in consumer markets and the rise of practical machine learning (ML) applications spearheaded by deep learning (DL) models. Many of these advancements have already realized practical application in medical fields, particularly those that use visual data. Another field that stands to benefit from this is human-computer interactions (HCI), where one of the primary goals is to find ways to understand users' psychological states better. One way to achieve this is by building DL models that can detect emotional states through non-invasive physiological measurements such as electrocardiography (ECG) and electroencephalography (EEG). However, the main challenge in training models using these signals is the prevalence of external noise, which can severely limit the generalization performance of most DL models. This dissertation proposes various approaches for building noise-robust DL models using ECG and EEG data. Specifically, it proposes (1) a method to more efficiently discover features based on heart rate variability (HRV) from ECG using DL by employing a spectral data representation based on Lomb-Scargle periodograms, (2) a method for training robust EEG-trained models by utilizing spatial-spectral EEG representations together with visual DL models, and (3) a subject-invariant bi-lateral variational domain adversarial neural network (BiVDANN) that simultaneously tackles the data constraints and subject variability limitations that are common to most EEG-based models. Results indicate that existing DL architectures are susceptible to common types of signal noise and benefit from the increased robustness introduced by the proposed methods. This is reflected by the improved classification performance of the optimized models on their respective target psychological states. Furthermore, detailed analysis of the model activations and embedding spaces of specific models demonstrate how existing architectures could be optimized to counteract common weaknesses of ECG and EEG signals. Overall, these studies demonstrate some of the limitations of existing DL architectures when applied to modeling psychological phenomena using ECG and EEG data and proposes a number of approaches for optimizing model architectures.

Acknowledgement

This dissertation would not have been possible without the guidance and support of a great many people. First and foremost, I would like to express my sincere appreciation to my supervisor, Professor Masayuki Numao, without whom I would not have had the opportunity to study in Japan to pursue higher education. I am also thankful to Professor Tsukasa Kimura who, despite only joining the lab a few years ago, has provided crucial guidance for my academic writing.

I would also like to thank members of the review committee, Professor Ken-ichi Fukui, Professor Jun Tanida, and Professor Yasumasa Fujisaki, who reviewed this dissertation and provided valuable comments which were instrumental to making it the best it could be.

I am also deeply thankful to the wonderful members of Numao laboratory, especially all the students who have become important friends and have left behind a lot of fun memories. I am especially grateful to Dr. Nattapong Thammasan who has been a great colleague and friend, thank you for all the long talks and your unceasing encouragement. I would also like to thank the many graduates of the lab who have been awesome colleagues, to Dr. Wu Hongle for his guidance and wisdom, to Dr. Wasin Kalintha for making the lab a little more interesting, and to Dr. Sirawit Sopchoke for all the delicious Thai food. Know that you made working long hours at the lab a more enjoyable experience. And to Ekasit Phermphoonphiphat, thank you for all the trust you put into our academic discussions, the road is long and hard but it will be worth it in the end. To Taweesak Emsawas, thank you for helping me find the joy in making academic diagrams. I would also like to acknowledge all the laboratory secretaries who have helped me throughout the years, especially Ms. Megumi Tanabe and Ms. Mika Kusakabe who went above and beyond towards assisting me during my first few years in Japan. Transitioning to life in Japan was much easier thanks to all your support.

An extended thanks also goes to my friends, colleagues and mentors back in De La Salle University. I am especially thankful to Dr. Roberto Legaspi who has always encouraged me to complete my PhD studies and greatly influenced my academic writing style. Similarly, I am thankful to Dr. Rafael Cabredo and Dr. Yamie Suarez, who aside from being my mentors for both work and life, were also the ones who encouraged me to even pursue a PhD in Japan. I am also grateful to Dr. Paul Salvador Inventado

for providing the groundwork for my early research at Numao Lab. Know that you all inspired my academic activities in various ways.

Of course, living in Japan would not have been possible without the support of the Humanware Innovation Program at Osaka University, a Program for Leading Graduate Schools (LGS) under MEXT. The program has not only supported me financially, but has also contributed towards forming me into a researcher who embraces learning about other scientific fields. It has exposed me to both the academic and the business sides of science through its mentorship program, domestic and overseas study-tour programs, conference travel grants, and short-term internship program. Speaking of internship, I cannot forget to thank Udana Bandara for his mentorship and guidance during my internship at Rakuten. Although we only worked together for a short time, your wisdom and guidance continues to shape the person I am today. I would also like to extend my deepest thanks to all the support staff of the Humanware Innovation Program at Osaka University, especially Professor Kazufumi Hosoda for his patience and support during the various activities. I would also like to thank everyone involved in the Center of Innovation (COI) Program at Osaka University of which I have had the pleasure of working with. Thank you for supporting my research activities and for helping to open up more opportunities for me beyond graduation. I look forward to doing more collaborative works in the future.

Finally, my deepest gratitude goes to my family and friends here in Japan and in the Philippines for the constant support and encouragement. Especially to Kimberly Cornelio, for always believing in me and encouraging me to push onward. Thank you so much for being patient with me especially through the most difficult times. I am also very grateful to my mother Josephine and my sister Inin for visiting me in Japan often and bringing a little slice of home. To my dearly departed father Daniel, you are the reason I enjoy science and keep striving to learn more about the world. Thank you for the wisdom you left behind, for it continues to guide my path. And of course thanks be to God, the almighty Father, thank you for all the blessings I have received and for keeping us safe through these difficult times.

Contents

List of Figures **iv**

List of Tables **vii**

1 Introduction **1**
1.1 General Introduction 1
1.1.1 Affective Computing and Physiological Sensing 2
1.1.2 Deep Learning on Physiological Signals 3
1.1.3 Domain Adaptation and Domain Generalization 5
1.1.4 Objectives of this Research 6
1.2 Organization 8

2 Detecting Naturalistic Work Stress with Deep Spectral HRV Models **10**
2.1 Introduction 10
2.2 Related Works 11
2.2.1 HRV Stress Analysis 11
2.2.2 Machine Learning for Stress Recognition 11
2.2.3 ECG-based Machine Learning Models 12
2.3 Methodology 15
2.3.1 Signal Processing 16
2.3.2 Model Architectures 20
2.4 Data Acquisition 22
2.5 Experiments and Results 24
2.5.1 Model Training and Optimization 24
2.5.2 Comparison Experiments 25
2.5.3 ROC Analysis 29
2.6 Summary and Future Work 30

3 Detecting Meditative States with EEG and Deep Visual Models **32**
3.1 Introduction 32
3.1.1 Challenges for Building Deep Models for EEG 33
3.2 Related Works 34

3.2.1 Image-based Deep EEG Architectures . . . 34
3.2.2 EEG Models for Detecting Meditation and Attention . . . 36
3.3 Methodology . . . 37
3.3.1 Signal Preprocessing . . . 38
3.3.2 Generating EEG Visualizations . . . 39
3.3.3 Model Architectures . . . 40
3.4 Dataset Acquisition . . . 41
3.4.1 Exploratory Analysis of Linear Features . . . 43
3.5 Experiments and Results . . . 44
3.5.1 Dataset and Models . . . 44
3.5.2 Deep Model Results . . . 46
3.5.3 Analysis of Deep Filter Activations . . . 46
3.5.4 Shallow Model Results . . . 47
3.5.5 Subjectivity Analysis . . . 49
3.6 Discussion . . . 50

4 Building Subject-Generalized Emotion Models using Variational and Adversarial Regularizers 52
4.1 Introduction . . . 52
4.2 Related Works . . . 53
4.2.1 Domain Adaptation . . . 54
4.2.2 Domain Generalization . . . 57
4.2.3 Contributions of the Work . . . 61
4.3 Methodology . . . 62
4.3.1 Input Data . . . 62
4.3.2 Deep Learning Architecture . . . 62
4.3.3 Regularizing for Normally Distributed Features . . . 63
4.3.4 Optimizing for Subject-Independence . . . 67
4.3.5 Training Phases . . . 69
4.4 Datasets and Models . . . 71
4.4.1 Datasets . . . 71
4.4.2 Data Preparation . . . 72
4.4.3 Model Implementation . . . 72
4.5 Experiments and Results . . . 74
4.5.1 Training Hyperparameters . . . 74
4.5.2 Comparison Results . . . 75
4.5.3 Qualitative Analysis of Embeddings . . . 78
4.5.4 Model Ablation Study . . . 83
4.6 Discussion . . . 84

5 5. Summary and Discussion 87

5.1 Summary 87
5.1.1 Primary Contributions 89
5.2 Conclusion 90
5.3 Limitations 91
5.4 Recommendations for Future Work 92

Bibliography **95**

List of Figures

2.1 PQRST-complex of normal ECG wave. The delay between consecutive R peaks indicates the R-R beat interval (or simply, RR). 13
2.2 Overview of HRV-DNN model architecture. Input ECG signals are segmented using sliding window segments, which are then passed to a spectral feature extractor in order to generate the LS-spectrogram data. The spectral data is used as input together with the data labels to train a deep learning model for stress detection. 15
2.3 Colorized sample RR spectrogram extracted from ECG data. Blue regions indicate spectral lower power, while Red regions indicate higher spectral power. 19
2.4 Residual block architecture originally described in He et al. [1]. The skip-connection (Top), enables passing gradients through the identity function, which preserves their magnitude. 20
2.5 ResNet-50 residual stack component. Each stack is comprised of 3 consecutive residual blocks (green). The 50-layer ResNet-50 model uses 4 of these stacks in the entire architecture. 20
2.6 The shallow network architecture based on VGG-Net. Each block is comprised of a 3x3 convolutional layer, a 2x2 max-pooling layer and a dropout layer. To compensate for the reduced input dimension received by deeper layers, the number of filters is doubled for each consecutive block. 21
2.8 Image samples of extracted ECG spectrograms at different window widths. 23
2.9 Comparison of test performance results between models trained on standard HRV features (MLP and DA) and deep models trained on STLST heart rate spectrum data (ResNet and Shallow). 26
2.10 Training and testing performance of ResNet-50 model. 27
2.11 Training and testing performance of the 5-layer CNN model. 28
2.12 Overlapped performance results of ResNet-50 and the shallow 5-layer CNN model. 29
2.13 ROC for the classification results of the ResNet-50 model. 30

3.1 Architecture diagram of the proposed deep visual EEG-Meditation model. 38

3.2 3-band EEG topomap image of a subject during post-meditative rest. RGB color channels represent theta, alpha and beta brainwave power, respectively. In this work, the spherical mapping (Left) is converted into a rectangular mapping (Right) to better accommodate rectangular convolutions. 39
3.3 Standard 10-20 electrode configuration used by the Ultracortex Mark IV EEG headset 39
3.4 Shallow model architecture 41
3.5 Overview of data gathering experiments and data segmentation. 42
3.6 EEG data mapping based on top 2 principal components. 43
3.7 EEG data mapping based on top 2 linear discriminants. 44
3.8 Model training results for ResNet-50. Error ranges represent 1-unit standard deviation over 10-fold cross-validation. 45
3.9 Input layer convolutional filter activations. 46
3.10 Model training results for shallow model. Error ranges represent 1-unit standard deviation over 10-fold cross-validation. 48
3.11 ResNet-50 model performance results compared to Shallow-Wide model. Error ranges represent 1-unit standard deviation over 10-fold cross-validation. 49
3.12 Subject-independent test results for 16 randomly selected subjects. Many of the models barely perform above random chance (50%) on this binary classification problem (rest vs. meditation). 49

4.1 Alpha, beta, and gamma power are extracted from the electrode data, plotted onto separate topographic maps, and are combined into a 3-channel topograph used as input for the proposed model. 62
4.2 The high-level architecture diagram of the proposed model. From left-to-right, (**a**) EEG topograph inputs, (**b**) convolutional variational encoders, (**c**) lateralized embeddings, (**d**) bi-lateral feature concatenation and lateral convolution, (**e**) subject-independent dense embeddings, (**f**) emotion classifier, and (**g**) adversarial subject classifier. 63
4.3 Variational autoencoders are composed of an encoder–decoder pair. The encoder outputs mean and s.d. components, which are used to sample the latent encoding, $\boldsymbol{z}$, stochastically from a target distribution. It combines losses for the reconstruction of the input, $\boldsymbol{x}$, and the KL-loss between the encoding, $\boldsymbol{z}$, and the target distribution (i.e., Gaussian). 64
4.4 The lateralized beta-VAE implementation used in this work. Each hemisphere is composed of an encoder and decoder pair with the encoder generating a stochastically sampled embedding, z, based on outputs, μ and σ. Separate weights are used to encode the features for each of the hemispheres to adhere to neuroscience theories of lateralization. 67

4.5 The domain adversarial network (DANN) implementation used in this work. The feature embedding layer is shared by the emotion classifier and the adversarial subject domain classifier. Each classifier is trained to classify for its respective task; however, the gradient is reversed and scaled with parameter λ, in effect training upstream layers for subject-independence. 68
4.6 Sample training progression for DEAP data. The graphs indicate the training/validation accuracy/losses for the emotion classifier/subject discriminator over training epochs. The black line indicates the point used for early stopping. 75

List of Tables

2.1 Comparison of Related Physiological Data Modeling Approaches. 14
2.2 Stress Annotations Distribution 23
2.3 Window Size Test Results Confusion Matrices 24

3.1 Comparison Results 48

4.1 Summary of Related EEG Domain Adaptation and Generalization Methods. 60
4.2 BiVDANN Implementation Details. 73
4.3 SEED Comparison Results (3-class) 77
4.4 DEAP Comparison Results (2-class) 77
4.5 SEED embedding visualizations for the tested models. Different colors correspond to their respective emotion class (**negative**, **neutral**, **positive**) or unique subject. The axis values represent the arbitrary 2D feature mapping performed by the dimension reduction algorithms. . . . 80
4.6 DEAP embedding visualizations for the tested models. Different colors correspond to their respective emotion class (**negative**, **positive**) or unique subject. The axis values represent the arbitrary 2D feature mapping performed by the dimension reduction algorithms. 82
4.7 SEED Ablation Results (3-class) 84
4.8 DEAP Ablation Results (2-class) 84

Chapter 1

Introduction

1.1 General Introduction

In just the past decade, deep learning (DL) has grown to dominate the market for image and audio analysis solutions due to its unrivaled ability to learn salient features from large, complex datasets. More recently, its reach has begun extending toward medical applications, and particularly those that involve image data, such as medical imaging, radiology and even microscopy [2]. Already, it is set to revolutionize the way we process medical datasets by presenting models that exceed human performance [3, 4]. Recently, we have also seen the rapid growth and proliferation of wearable sensors, and together with it, various attempts at applying machine learning to the signal data for various applications in human-computer interactions (HCI). However, much of the existing work still relies on shallow models and high-level extracted features [5]. One reason for the slow progress is that while it is generally accepted that most DL architectures work well with image-based problems, much less is understood about how modern machine learning architectures react to physiological signals such as electrocardiographs (ECG) and electroencephalographs (EEG) [6, 5]. The major challenge for training models using most physiological signals is the prevalence of unrelated noise in the data. When compounded with the scarcity of large open physiological datasets, this severely limits the generalization performance of traditionally resilient DL models. Meanwhile, the problem of small noisy datasets is not unique to psychological signals, in fact most signal processing fields tend to suffer from the same issues. Therefore, methods found effective for physiological data could potentially have a ripple effect towards other adjacent signal processing fields.

While some studies are attempting to reach practical usability by building end-to-

end architectures trained on large medical datasets [3, 4, 7], others propose to bypass the most difficult low-level feature learning stages by using intermediate data representations [8, 9]. Our work falls under the latter type. Specifically, we investigate the challenges of building deep psycho-physiological models using ECG and EEG data. Through a number of studies, we identify some of the common sources of noise in these data representations and propose noise robust architectures for each. We show how effective DL technologies typically used in image processing can be applied towards psycho-physiological signal modeling, and how models can be trained efficiently from relatively small datasets.

1.1.1 Affective Computing and Physiological Sensing

Emotions play an important role in our daily lives, not only do they regulate our interactions with others, but they also influence our perception of the world around us. The general experience, as well as our own recollection of events, are shaped by the emotions we felt at the time [10, 11]. Computer systems may also benefit from learning and understanding human emotions, enabling them to facilitate smooth human–computer interactions, leading to an improved overall user experience [12, 13]. On the other hand, negative emotions can not only cause psychological strain, but may also bring about adverse physical symptoms. Findings from a large number of clinical studies [14, 15] have linked cardio-vascular disorders to psychological stressors. In particular, long-term exposure to stressful environments, such as those commonly found in offices, has been shown to be a major risk factor for major depressive disorder (MDD), a disorder that affects an estimated 350 million people across the world [16]. In response to this, we have witnessed a surge in the development of ambient intelligence technologies for the purpose of supporting mental health and monitoring stress [12, 17]. These technologies, which attempt to apply computing technologies to recognize and understand human emotions, generally fall under the field of affective computing [13].

We humans convey emotions to each other through a range of channels, including: facial expressions, vocal tone, gestures and, in recent years, even textual emoticons. While the study of these channels constitute fields of their own, physiological signals may hold even greater expressive potential. Unlike many of the other channels, these signals usually run quietly in the background and are largely autonomic [13]. The main drawback, however, is that these signals are less thoroughly understood, and are also much more

difficult to isolate and clean. Fortunately, modern sensors and data-processing techniques are capable of finding obfuscated patterns within these obscure data channels, which could then be linked to emotions. Naturally, each data channel has its pros and cons. For example, channels like facial expression and vocal tone are typically only interpretable in specific contexts. Subjects working at a computer may feel a wide range of emotions, but may not necessarily show them through facial expression or through active vocalization. Meanwhile, gestures are the most likely to differ across cultures. In the case of physiological signals, the problem most often encountered is high complexity and a low signal-to-noise ratio (SNR) [5, 18, 19]. This is to be expected since most of these non-invasive measures rely on surface-level readings resulting from the underlying physiological activity. For context, EEG is usually measured as microvolt(μV) variations in the scalp, while ECG is measured as millivolt(mV) changes at different points of the body usually passing through the torso and heart. This exposes the signal to various external voltage sources such as powerline noise, radio signals, and even nearby humans which can act as electric capacitors. In addition, we must also consider variations due to changes at the point of contact of the electrodes. Such changes include, variations in conductivity, variations in precise locations of electrodes, and electrode shifting due to motion. Finally, there are also various internal factors that can affect the signal, the most prominent of which are electrical signals from the skeletal muscles. Thus, it is often necessary to apply a wide range of signal cleaning filters, perform manual analysis of temporal and spectral features, and remove noisy samples before even basic physiological analysis can take place. It is this complexity that makes them ideal for processing with robust machine learning (ML) architectures. ML methods have proven themselves to be much more efficient than humans at detecting salient patterns from complex aggregate signal data, especially when the appropriate architectures and regularizers are used [5, 2, 1]. As a result, recent advancements in ML have contributed significantly to the state-of-the-art of physiological sensing and decoding [11, 5].

1.1.2 Deep Learning on Physiological Signals

In recent years, DL has gained rapid adoption in fields such as machine vision and natural language processing where it can make full use of the large labeled datasets. This has produced many of the state-of-the-art models still being used today in a wide range of practical applications [20, 1]. In stark contrast, prominent generalizable models

for physiological signals are much rarer, and none have gained anywhere near the same prominence as some of the machine vision models. A likely reason for this relatively slow growth is the lack of large, open datasets for training physiological models. The scarcity of physiological datasets is primarily due to the difficulty of acquiring good physiological data. First, as mentioned in the previous section, physiological data is inherently noisy and thus requires sufficient expertise and a significant amount of time just to set up properly. On top of this, the type of data gathered is typically perceived to be more personal since the sensors require direct contact with the subject's skin and can possibly reveal underlying pathologies. This can deter participants from even considering volunteering in the experiments and can limit the scope of experiments. And even when the data is finally gathered, there is still the challenge of finding the relevant patterns in the multi-channel data. It is not surprising that considering all the overhead costs, there is currently little financial incentive to build large physiological datasets. An exception to this are the models produced for healthcare applications where data is relatively plentiful, such as processing clinical images, decoding health records, and genomics [2]. Many of the common limitations listed do not incur as much additional overhead in hospitals and medical institutions since they are mostly already accounted for, so data is abundant. In the case of ECG decoding, one of the few notable works was a study by Hannun et al. [4], where they used convolutional neural networks (CNN) to analyze raw ECG sequences and identify 14 different types of arrhythmia. Using a large dataset of over 60,000 clinical ECG recordings, they were able to train a model that, on average, was more accurate at identifying arrhythmia than actual cardiologists recruited for the experiment. This work demonstrated how DL could help physiological models reach practical usability. However, it is important to note that heart arrhythmia is characterized by significantly irregular heart rhythm which is visibly apparent from the raw EEG signal. Therefore it lends itself better to end-to-end modeling, unlike stress analysis (of normal subjects) where ECG morphology is expected to be consistently normal. Instead, the natural but subtle variations in beat intervals can be used to estimate autonomic nervous system activity in response to stress [21]. In fact, this is supported by findings in a recent survey [22] that found that it was much easier to build ML models from beat interval data than raw ECG signals. Therefore, in this work, we apply noise-robust spectral representations of beat intervals to train deep models for multi-level stress classification.

Meanwhile, among recent EEG modeling studies, a work by Schirrmeister et al. [9] stands out for having produced an effective end-to-end model for modeling motor visualizations from EEG. Their study showed how DL models could be built from relatively small datasets by training more efficiently. While the results were not as striking as in [3, 4], they provided valuable insight into how physiological signals can affect learning in DL models. Nevertheless, end-to-end models may not be ideal for the type of signal analysis that do not rely on structural elements of the signal [23]. Unlike motor visualizations, where the exact timing of brain events can be known beforehand (i.e, event-related potentials), emotions are less predictable and so are typically analyzed through spectral methods. Therefore, a key challenge for EEG-emotion studies moving forward is finding high-level spectral features that are generalizable and mostly free from unrelated morphological noise. Domain adaptation methods could offer some solutions for filtering out unrelated features during ML training.

1.1.3 Domain Adaptation and Domain Generalization

Domain adaptation and *domain generalization* approaches attempt to solve the generalization problem by allowing models to optimize for features that are invariant across different domains [24]. Here a *domain* represents a probability distribution from which one can obtain samples used to train a model (e.g., training and testing). Domain adaptation aims to produce optimized models for the target domain by applying knowledge learned from a labeled source domain, while also leveraging unlabeled samples from the (possibly unlabeled) target domain. Its potential for extracting relevant knowledge from multiple sources has lead it to become a hot topic in the field of computer vision [24, 25]. This is driven by realizations that along with the great learning potential and flexibility of deep learning models comes the increased propensity to overfit to a single sample distribution (i.e., the training set) [26]. These characteristics may also be useful for physiological data analysis where multiple domain definitions exist. For instance data from different devices may define different domains, so too could data from similar devices but from a different group of subjects. Even within a single dataset, samples from different users can show domain-like variations, so adaptation techniques can help sanitize the features learned form these irrelevant fluctuations. Therefore, in this work, we attempt to adapt some of these concepts to build generalized EEG-emotion models.

1.1.4 Objectives of this Research

As discussed in the previous sections, application of ML to physiological data analysis is still in the early stages, so much work is yet to be done. One of the fundamental problems we must first solve is how to efficiently prevent the ML models from overfitting to signal noise. At present, most efforts are focused on either building end-to-end models using large amounts of data, or training ML models on the traditional extracted features used in past works. In this work, we propose a middle ground solution lying between these two extremes: to apply low-level feature extraction to minimize the effects of common noise sources, and to apply machine learning to learn high-level features from training data. Since the proposed methods operate primarily on generic signal representations, the problems and optimizations discussed may also apply to other signal-processing disciplines with similar noise characteristics.

This dissertation proposes various approaches to building optimized DL architectures for detecting stress-related psychological phenomena through the use of physiological signals such as ECG and EEG. In general, the objective of the work is to design approaches for building ML architectures which are capable of resisting common sources of noise in their respective modeling tasks. Through the course of the study we conduct 3 separate sets of experiments on different datasets. We start with modeling a relatively simple single-channel signal (i.e., ECG) and work our way towards more complex multi-channel EEG signals. The details of the experiments are outlined as follows:

First, we investigate the viability of using image-based DL models for measuring stress from naturalistic ECG. The objective of this study is to better understand the learning requirements of spectral ECG DL models, and to identify ways of efficiently building deep models for multi-level stress detection. We hypothesize that using a robust spectral-temporal representation of heart beat interval patterns to train DL models will enable it to learn features that would exceed the robustness and expressiveness of traditional heart rate variability (HRV) indices. Specifically, we propose the use of short-time Lomb-Scargle (LS) spectrograms to train CNNs on a naturalistic work stress dataset. To test the reactivity of different DL architectures, we use pre-trained deep models designed for image recognition as well as smaller optimized architectures. The results from this experiment, aside from verifying the efficacy of the method on ECG data, will also help to guide the design of our next studies on more complex EEG data.

In the second study, we propose a method for building optimized models of medita-

tion using novice EEG data. The main objective of this work is to develop an architecture that allows more efficient use of EEG datasets by using an intermediate topographical representation thereby enabling training on partially noisy data samples. Specifically, we hypothesize that the ability of visual DL models to adapt to large degrees of occlusion will manifest in the proposed model as robustness to missing data channels. To test the hypothesis, first we build a naturalistic dataset of novice meditation practice and train a gold-standard DL architecture to detect meditative EEG. Next, we use the findings from this initial model to guide the design of a more optimized architecture. By enabling models to train from a dataset that includes noisy EEG samples with missing electrode channels, we can promote more efficient use of noisy EEG data samples. Future works can benefit from more robust models and the reduced costs of acquiring large EEG datasets. These experiments will also lay the groundwork for the final study in this dissertation.

Finally, for the third study, we expand the previous method by proposing a subject-invariant bi-lateral variational domain adversarial neural network (BiVDANN) designed for modeling emotion from EEG. The objective of the study is to develop a DL architecture that reduces the effects of subject variability and allows training generalized models from small datasets. By combining variational autoencoder (VAE) components and subject-adversarial training, we expect the model to learn subject-generalized embeddings for the EEG data. Through the work, we develop an efficient framework for building generalized EEG-based models from relatively small datasets, while preserving the ability to train on larger datasets by adjusting regularization parameters.

To summarize, through this work, we explore ways to efficiently build DL models from noisy signal data such as ECG and ECG. Specifically, we address the following noise components which could negatively affect ML models in their respective contexts:

- Typical morphological ECG signal noise, including occasional short-term artifacts resulting from motion. Addressed in Chapter 2 through spectral-temporal feature extraction using LS spectrograms and deep CNNs for multi-level stress classification.

- Typical morphological EEG signal noise, particularly partial data loss due to unstable data channels. Addressed in Chapter 3 through spatially-interpolated spectral topology maps and shallow-wide CNNs for modeling meditation.

- EEG signal variability due to subjective differences in neurological responses to emotional stimuli. Addressed in Chapter 4 through a novel framework for building subject-generalized EEG models, featuring bi-hemispheric feature learning combined with joint optimization of variational and subject-independent embeddings tested on emotion classification datasets.

To this end, we propose and test various noise-robust intermediate data representations and DL architectures to model stress and emotion. Through our experiments with a ECG-Stress model we test the robustness of the approach on single-channel temporal data. In the second study, we build an EEG-Meditation model and show how a similar approach to the first can be adapted to datasets with missing or noisy data channels. And finally, in the final model we propose a DL architecture that extends the previous methods towards subject-generalization. Through the experiments, we show that the proposed methods easily outperform ML models trained on traditional extracted features, while also allowing DL architectures to train on relatively small physiological datasets while avoiding overfitting. Overall, the work presents various modeling methodologies that could be useful for studying psycho-physiological phenomenon with noisy data, while the experiments on BiVDANN may offer some solutions to other DL works where generalization is a primary concern.

1.2 Organization

The rest of this dissertation is organized as follows: In Chapter 2, we introduce a method for efficiently discovering HRV-related features from ECG using DL by employing a spectral data representation based on LS periodograms. Through comparisons with HRV-based models, we show that this low-level spectral-temporal feature representation allows DL models to avoid overfitting on morphological variations while still enabling them to discover features that can outperform traditional gold-standard HRV indices. Based on the promising results obtained from these initial experiments, in the next study we attempt to apply a similar methodology on the more complex but also more expressive EEG data. In Chapter 3, we present a method for training robust EEG-meditation models by utilizing spatial-spectral EEG representations together with optimized visual DL models. In this section, we show how to address persistent localized noise in EEG channels by implementing a noise threshold method combined with a spatially-interpolated topology map representation. In addition, we demonstrate how

deep learning architectures can be optimized for the proposed spectral representations. By using a combination of performance comparisons and DL layer activation visualizations, we show how features learned by our models differ from traditional image-based models and how architectures can be optimized based on these differences. Despite the promising subject-dependent results, we show the shortcomings of the model in regards to subject-generalized performance and propose to address these limitations with the third study. In Chapter 4, we propose a novel EEG architecture featuring a subject-invariant bi-lateral variational domain adversarial neural network (BiVDANN) for learning subject-generalized EEG embeddings. Through embedding visualizations, we show that our approach more effectively learns a generalized subject-independent embedding space. And, through comparisons with similar generalized embedding approaches, we show that this directly translates to better subject-independent classification performance. Furthermore, we include ablation experiments to affirm the merits of the ML components used in the proposed method. Finally, in Chapter 5, we summarize the results and contributions of the three studies and present some possible directions for future work.

Chapter 2

Detecting Naturalistic Work Stress with Deep Spectral HRV Models

2.1 Introduction

Prior to the popularization of deep learning, most studies attempting to use ML to model psychological states from physiological signals like ECG utilized ML methods such as support vector machines (SVM) and simple neural networks (NN). This often required the use of well-studied medical indices like heart rate variability (HRV) which were often selected after prior works established direct correlations with the target state being modelled [27]. However, while these were found effective for characterizing serious medical pathologies, studies found limited success in establishing consistent predictive links to comparatively ambiguous physiological changes like those from stress or emotion [28, 29]. Detecting stress, in particular, could have widespread applications in ambient healthcare and self care. It is already well established that long term exposure to stressful environments, such as those commonly found in offices, could be a major risk factor for the development of various cardio-vascular disorders [14, 15]. Practical stress detection technologies [17] could be the solution to reducing work-related pathologies. Or, at the very least, they may help encourage people to take care of their psychological well-being, in the same way that step trackers are currently helping people manage their physical health.

Because of recent advancements in DL techniques, there has been an influx of successful applications of ML in a wide range of healthcare applications, such as processing

clinical images, mining health records, predictive genomics, and even works using mobile and wearable sensor data [2]. Similar technologies are even finding increasing use in the field of affective computing [30].

However, a primary limitation tends to be the large training datasets which are required to produce practical levels of reliability such as in Hossain and Muhammad [31]. In this study [32], our goal is to test whether DL approaches can be effectively applied to a smaller naturalistic dataset, and particularly for the detection of comparatively subtle physiological changes caused by stress.

2.2 Related Works

2.2.1 HRV Stress Analysis

Historically, heart rate variability (HRV) spectrograms were used by medical practitioners as the primary tool for measuring cardiac neural regulation in a non-invasive manner. This was a result of studies that linked mental and physical states to cardiac rhythms which could be easily and unintrusively measured using electrocardiographs (ECG) [33]. In the past few decades, studies have established the connection between HRV and the sympatho-vagal stress response as being visible in both the time and frequency domains. Through experiments and empirical anlaysis, these works have identified indices which show high correlations with physical and mental stress, particulary with the high and low frequency ranges [27]. However, despite our growing understanding of physiological signals, even now there is no universally accepted standard for stress evaluation based on HRV features. In recent surveys [34, 35], it has been found that neurobiological evidence links some components of HRV to activity in cortical regions related to stress appraisal as well as other brain mechanisms. But overall, it seems that one of the main challenges is identifying which components of HRV may be relevant to evaluating stress and to what extent.

2.2.2 Machine Learning for Stress Recognition

One of the earliest attempts to use physiological signals to detect stress states was Shi et al. [36]. In that work, they used a combination of physiological signals, including a number of ECG and HRV indices, to train SVMs for detecting stress states. The final models were able to distinguish stress and rest states at up to 62% accuracy. In addition, it was also discovered that it was possible to increase the performance of the models

by an additional 6% by adding personalized data during training. While this would be useful for building personalized models, it has limited use for generalization.

Since then, most works on stress analysis have shifted towards generalizable features in the form of visual or auditory data [37]. In that study, researchers built a multimodal model for recognizing emotions (valence and arousal) using an ensemble of deep neural networks for audio-visual analysis. The visual channel was modelled using a residual CNN while the speech channel used a CNN with a long-term short-term memory (LSTM) network. After training separately, they topped the CNN networks from both channels with a LSTM layer to learn the multimodal features to build an end-to-end model. They found that pre-training the networks separately helps speed up training while multimodal end-to-end training takes advantage of correlations between channels. These types of applications have the advantage of using powerful DL architectures for general vision and audio, and simply adapting them for facial expression and speech prosody [31]. Currently, there are a few works attempting to build similar baseline end-to-end models for ECG [38, 3, 4].

2.2.3 ECG-based Machine Learning Models

In one of the most notable works, Hannun et al. [4] used a 34-layer residual convolutional neural network (CNN) to analyze raw ECG sequences in order to identify 14 different types of arrhythmia. Using a very large dataset of over 60,000 recordings they were able to train a model that performed better on average than actual cardiologists recruited for the experiment. Their work demonstrates how deep learning is able to identify structural patterns in heart rate data. After training on their comprehensive dataset, the model was able to identify structural patterns in the signal linked to the pathology of arrhythmia. In the case of the current study, unlike in medical conditions like arrhythmia, there are currently no widely accepted structural ECG patterns linked to stress states. Instead, experts typically rely on the analysis of higher level frequency domain features in order to minimize the effects of physical variations between subjects.

Keren et al. [7] also attempted to build an end-to-end ECG model for emotion recognition using a combination of CNNs, RNNs and FC layers. However, their findings showed that models that used raw ECG signals as input performed significantly worse than those trained on a set of HRV features. This was consistent for both valence and arousal detection. In a related work [18], they used multiple signals including galvanic

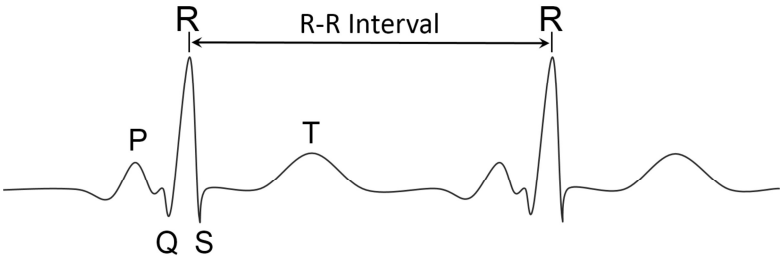


Figure 2.1: PQRST-complex of normal ECG wave. The delay between consecutive R peaks indicates the R-R beat interval (or simply, RR).

skin response(GSR), electromyograms (EMG), ECG, and respiration to model naturalistic stress during driving. The study found that after applying feature selection to optimize features, heart rate and respiration were one of the most consistently selected features. And while the study did not directly investigate the relation between the two features, prominent clinical studies such as Schipke et al. [39] have shown direct correlations between respiration and certain HRV components. A recent survey [22] also noted that across studies, classifying raw ECG signals was more difficult compared to classifying based on heartbeat intervals. This is thought to be caused by innate differences in signal morphology between each person. Thus, it was concluded that end-to-end models were more suited for classifying the signal itself (e.g., to classify noise level), and that derived measures such as beat intervals were better for classifying generalizable phenomena based on heart rate. In addition, they survey also noted weaknesses in standard wavelet decomposition techniques when applied to unevenly sampled signals such as heart beats. A summary of these related works is shown in Table 2.1.

Findings from existing studies indicate that end-to-end ECG models may not yet be suitable for diagnosing changes that do not rely on morphological aspects of the ECG signal. This could also explanation why models trained on raw ECG signals are more difficult to train than those that use an intermediate representation like HRV or RR intervals, as found in a survey [22]. It brings us back to the issue of weak SNR. The raw shape of the signal can change significantly due to external factors like variations in specific electrode positioning, sensor gain, and even the physiology of subject (e.g., height). Individuals can be trained to intuitively recognize the different components of a normal ECG waveform shown in Figure 2.1. And because we understand the concept of heartbeats, it is easy to extrapolate to the concept that intervals between peaks represent heart rate. In addition to this, since regular heart rhythms are relatively well-

Table 2.1: Comparison of Related Physiological Data Modeling Approaches.

Work	**Input Features**	**Model Type**	**Target Label**	**Label Type**	**Noise Robust**	**Feature Discovery**	**Small Data**
Hannun et al. [4]	ECG	CNN	Arrhythmia	Categorical	△	◯	×
Keren et al. [7]	ECG + HR + GSR	CNN + RNN	Emotion	Categorical	△	◯	×
Ghaderi et al. [18]	ECG + GSR + EMG + Respiration Index Features	KNN, SVM, Feature Selection	Stress	Multilevel	△	×	◯
Melillo et al. [27]	HRV Index Features	Linear Discriminant Analysis (LDA)	Stress	Binary	◯	×	◯
Shi et al. [36]	ECG + HRV Index Features	Personalized SVM	Stress	Binary	◯	×	◯
Tzirakis et al. [37]	Audio + Video	CNN + LSTM	Emotion	Categorical	△	◯	△
Proposed Method	ECG Spectrogram	CNN	Stress	Multilevel	◯	△	◯

◯ = strongly supports property; △ = partially supports property; × = weakly supports or does not support property

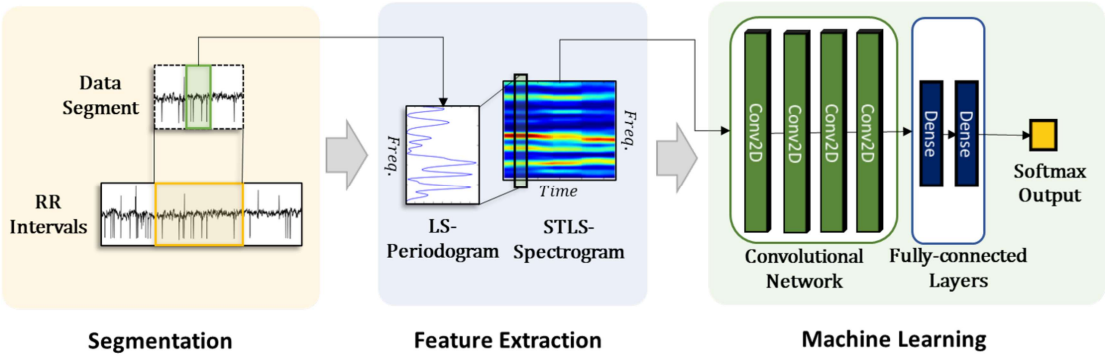


Figure 2.2: Overview of HRV-DNN model architecture. Input ECG signals are segmented using sliding window segments, which are then passed to a spectral feature extractor in order to generate the LS-spectrogram data. The spectral data is used as input together with the data labels to train a deep learning model for stress detection.

understood, tagging and correcting missing or false beats can be efficiently handled by simple algorithms (assuming the subject is normal) [21]. Meanwhile, detecting normal beats and correcting them constitutes an ML problem in itself. It may be too much to expect an end-to-end model to learn all these correlations, in addition to learning to recognize spectral-temporal phenomena like HRV, and isolating patterns that relate to stress. This is further compounded by the small datasets available for analyzing naturalistic mental stress. On the other hand, HRV features are consistently found to be effective regardless of the ML technique used. The need for human interpretability has reduced HRV to a small number of indices representing spectral and temporal features of heart rate. This could be limiting the expressiveness of the data itself since a lot of information is lost when the spectral-temporal values are aggregated into these indices. Thus, in the current work we propose to train a model on an intermediate spectrogram representation in order to limit its exposure to irrelevant morphological variations of ECG while retaining its ability to learn HRV-like features. In particular, we propose the use of Lomb-Scargle (LS) spectrograms for building an intermediate data representation for training visual DL models to recognize multi-level stress.

2.3 Methodology

This work aims to use spectral data representations derived from ECG together with deep learning models to predict mental stress. The goal is to allow the model to focus on discovering spectral-temporal features of heart beat patterns, similar to the HRV analysis done in majority of stress studies. This should reduce the need for noise filtering

which would also lower the data requirements of the model. In this case, we propose the use of short-time LS spectrograms to train CNNs on a naturalistic work stress dataset. Specifically, we use a sliding window Lomb-Scargle approach to build a spectrogram of heart rate, in lieu of the traditional short-term Fourier transform (STFT) as a way to avoid the interpolation artifacts common in standard wavelet approaches. Next, we use pre-built deep models designed for image recognition and compared them with smaller simplified architectures, and compare their performance when applied to our data representation for multi-level stress recognition. An overview of the proposed method is illustrated in Figure 2.2. The following sections describe the details of the experiments and tools used in this research.

2.3.1 Signal Processing

Previous works on the spectral anlaysis of HRV applied the fast fourier transform (FFT) [40] to produce periodograms from which the power of different frequency bands can be measured and compared. However, one limitation of the FFT algorithm is that it assumes that all samples used for analysis are evenly spaced in time. This poses a fundamental issue for heart rate since the RR intervals used to represent detected heart beats are not evenly sampled. To better illustrate the problem, refer to the Segmentation phase described in Figure 2.2. In the RR interval graph, the vertical axis represents the beat interval (in ms), while the horizontal axis describes the temporal order of the beats. The RR interval data is unlike the usual signals processed with FFT, where the data can be safely interpolated at the horizontal axis since they represent regular measurement intervals. Instead, they are more similar to a record of events which occur at mildly varying regular intervals. In Clifford and Tarassenko [21], they show that data resampling methods used to align RR interval data for FFT analysis can lead to artifacting and a significant increase in error rates in the final PSD estimation. A better alternative has been put forward in the form of the Lomb-Scargle periodogram, a method based on FFT that does not require evenly sampled data.

Lomb-Scargle Periodogram

Periodograms are some of the most basic tools used for spectral analysis, the most common of which is the discrete Fourier transform (DFT). The DFT can be defined for any sampled dataset,

$X(t_i), i = 1, 2, ..., N_0$ as,

$$FT_X(\omega) = \sum_{j=1}^{N_0} X(t_j) \exp(-i\omega t_j). \tag{2.3.1}$$

The periodogram can then be defined as,

$$\begin{aligned} P_X(\omega) &= \frac{1}{N_0} |FT_x(\omega)|^2 \\ &= \frac{1}{N_0} \left| \sum_{j=1}^{N_0} X(t_j) \exp(-i\omega t_j) \right|^2 \\ &= \frac{1}{N_0} \left[\left(\sum_j X_j \cos \omega t_j \right)^2 + \left(\sum_j X_j \sin \omega t_j \right)^2 \right] \end{aligned} \quad , \tag{2.3.2}$$

which, while sufficient for calculating an accurate estimation of the FFT of most signals [41], was found to be sensitive to noisy signals and other artifacts [42]. In practice, it is necessary to apply additional spectral smoothing equations, such as *spectral window functions*, to reduce the variances of the spectrum. Most of these methods are designed to be applied on evenly sampled data. Though some may also be applied to the unevenly sampled case, such as in RR intervals, it has been found that it can lead to higher error and increased sensitivity to signal noise [43]. This can also be mitigated by the addition of more data for calculating the spectra as a way to taper off the signal-to-noise ratio, but this may not be an option for data-scarce or time-intensive domains. A reasonable alternative is the Lomb-Scargle (LS) periodogram [44], a notable method designed for use in calculating frequency spectrograms from unevenly sampled data streams such as heart beats.

The equation is based on the FFT and is wholly similar but addresses some of the noise sensitivity and *spectral leakage* issues of the standard FFT equation. FFT generally assumes uniform sampling of the data in order to determine an orthogonal set of frequencies for which the values of the Fourier coefficients are independent. However, for non-uniformly sampled data such as RR intervals, such an orthogonal set generally does not exist [45], leading to spectral leakage[46]. This can be mitigated to some extent by applying data interpolation, however this has been shown to lead to artifacts such as an underestimation of high-frequency amplitudes and an overestimation of low-frequency amplitudes [46]. The LS transform (LST) addresses these limitations by employing a time delay (τ) for each frequency (f) on order to determine that a pair of sinusoids is mutually orthogonal. The LST defines a generalized form of the classical periodogram

as,

$$P_X(\omega) = \frac{1}{2}\left\{ \frac{\left[\sum_j X_j \cos\omega(t_j - \tau)\right]^2}{\sum_j \cos^2\omega(t_j - \tau)} + \frac{\left[\sum_j X_j \sin\omega(t_j - \tau)\right]^2}{\sum_j \sin^2\omega(t_j - \tau)} \right\} \tag{2.3.3}$$

where τ is defined for each f as,

$$\tan(2\omega\tau) = \left(\sum_j \sin 2\omega t_j\right) \Big/ \left(\sum_j \cos 2\omega t_j\right). \tag{2.3.4}$$

This modified equation outputs similar values to the original FFT in most cases, however it has a simpler statistical behavior and is equivalent to the reduction of the sum of squares in least-squares fitting of sin waves to data [44]. A key difference can be observed in their denominators, where the DFT assumes complete phase sampling at each frequency f with divisor N_0, while LST applies $\sum_j \cos^2\omega(t_j - \tau)$ and $\sum_j \sin^2\omega(t_j - \tau)$ which accounts for time-shifts. As a result, it is also time-translation invariant and reduces to a similar representation as DFT in cases where data is evenly spaced. This makes it ideal for HRV analysis and is the method used in this research. The next issue is how to tackle the non-stationarity of heart rate. For this, we construct spectrograms from the periodograms.

Spectrogram Analysis

Spectrograms are visual representations of frequency spectra over time. They are tools that are used extensively in applications that deal with audio and radio signals. Recently, they have also seen frequent application in the analysis of physiological signals such as brainwaves and heart rate. When applied to RR intervals, they can reflect the activity of the autonomic nervous system. Specifically, they have been shown to correlate well with the interactions between the sympathetic and parasympathetic nervous systems, which are then related to the stress response [28, 34]. The most common method used for building spectrograms is the short-time Fourier transform (STFT). Put simply it is built by applying the DFT over a pre-defined window of data and sliding the windowed function over the data series using a fixed amount of overlap. Formally, for any series $x(n)$ we can defined the STFT as,

$$X_w(mS, \omega) = \sum_{n=-\infty}^{\infty} x(n) w(n - mS) \exp(-j\omega n) \tag{2.3.5}$$

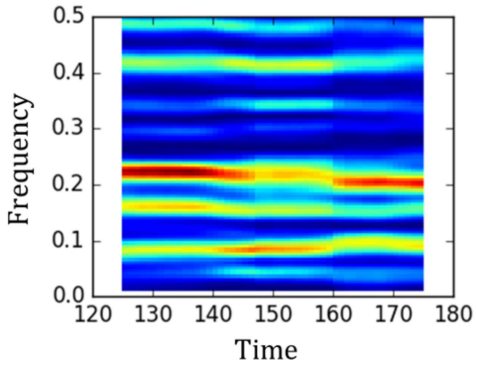


Figure 2.3: Colorized sample RR spectrogram extracted from ECG data. Blue regions indicate spectral lower power, while Red regions indicate higher spectral power.

for analysis window w, frame index m and step size S [47]. The internal equation $\sum_{n=-\infty}^{\infty} x(n)w(n - mS)\exp(-j\omega n)$ represents the Fourier transform of $x_w(mS, \omega)$ and is stored in the matrix $\hat{X}$ whose columns are Fourier periodograms $P_X(\omega)$. To adapt the spectrogram to the unevenly sampled heart rate data, we replace this equation with the Lomb-Scargle equation,

$$X_w(mS, \omega) = \frac{1}{2}\left\{ \frac{\left[\sum_j x(n)\cos\omega(t_j - \tau)\right]^2}{\sum_j \cos^2\omega(t_j - \tau)} + \frac{\left[\sum_j x(n)\sin\omega(t_j - \tau)\right]^2}{\sum_j \sin^2\omega(t_j - \tau)} \right\}. \tag{2.3.6}$$

In this way, we mitigate the sensitivity of the standard STFT while maintaining its ability to capture the nonstationary patterns in the temporal data. Figure 2.3 shows a colorized sample of an RR spectrogram. In effect, we end up with a very high dimensional representation that could potentially have more expressive and appropriate features for detecting stress than simple HRV indices. For example, the spectrogram in Figure 2.3 features a prominent peak at around 0.22Hz. Its frequency range and consistency throughout the window indicates that it may be correlated with changes to heartbeat due to breathing. If this is the case, the spectrogram data not only allows machine learners to potentially discover HRV patterns, but also respiration patterns. As noted in Ghaderi et al. [18], these are some of the top features discovered to be strongly indicative of naturalistic stress states. The next step is to use machine learning to discover these features and compare the performance difference with models based on traditional HRV features. Due to the 2D spatial nature of data representation, convolutional neural networks should provide an efficient learning architecture.

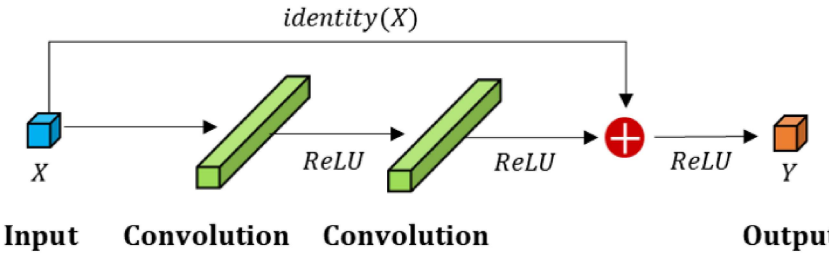


Figure 2.4: Residual block architecture originally described in He et al. [1]. The skip-connection (Top), enables passing gradients through the identity function, which preserves their magnitude.

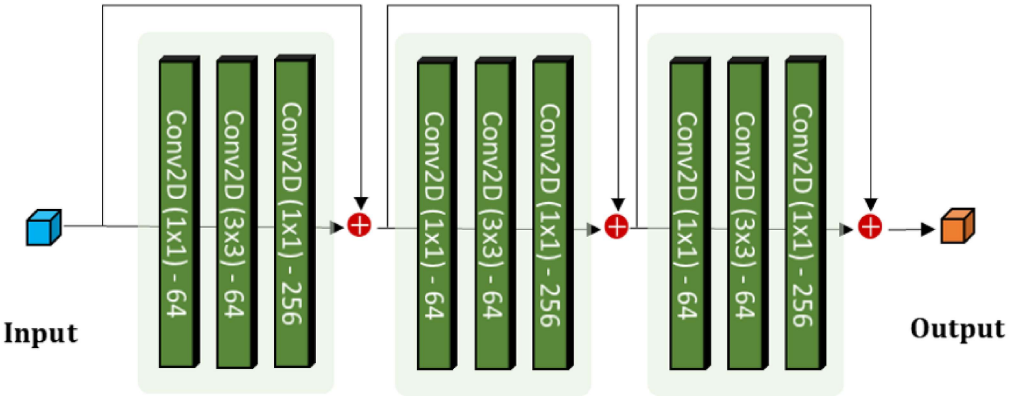


Figure 2.5: ResNet-50 residual stack component. Each stack is comprised of 3 consecutive residual blocks (green). The 50-layer ResNet-50 model uses 4 of these stacks in the entire architecture.

2.3.2 Model Architectures

Deep Model Architecture

A residual network (ResNet) models were selected as the benchmark for the expected performance of visual models adapted to our ECG-stress recognition architecture. Initially formalized by He et al. [1], ResNets are a subset of ConvNet architectures that make use of convolutional layer blocks designed to learn to output a *residual* representing the changes from the previous layer's output. A diagram of this component architecture is shown in Figure 2.4, hence the "residual" naming convention used to refer to the block architecture and the overall network. A key component of each residual block is the inclusion of a *skip connection* that allows passing of the unchanged input from the previous block. This schema enables the construction of very deep architectures while minimizing the effects of the *vanishing gradient problem* [48].

In this work, we use ResNet-50 a 50-layer implementation based on the original

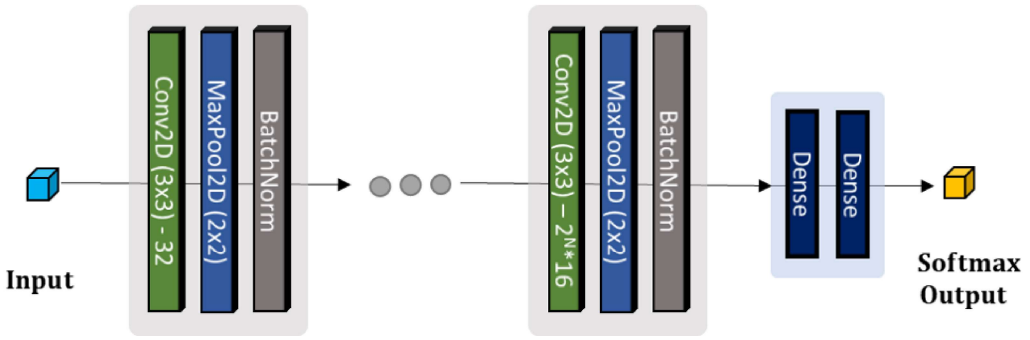


Figure 2.6: The shallow network architecture based on VGG-Net. Each block is comprised of a 3x3 convolutional layer, a 2x2 max-pooling layer and a dropout layer. To compensate for the reduced input dimension received by deeper layers, the number of filters is doubled for each consecutive block.

ResNet architecture. Here, residual blocks are defined using sets of [Conv2D(1×1)] + [Conv2D(3×3)] + [Conv2D(1×1)] layers, as seen in Figure 2.5. Next, consecutive residual blocks are grouped into *stacks* where each consecutive stack features convolutional layers with double the filter counts of the previous. In total, the traditional ResNet-50 architecture uses 4 residual stacks with [3, 4, 6, 3] residual blocks each, topped with an average pooling layer and finally a dense embedding layer. In our implementation, we replace the final non-convolution components with $2 \times$ [Dense (128)] layers followed by a soft-max layer for classification across all the models tested.

Shallow Model Architecture

One of the disadvantages of deep ML architectures is their tendency towards overfitting, particularly when the number of trainable parameters far exceeds the amount of data encoded by the training set. Essentially, this makes the resulting models capable of "copying" the training data entirely through their model parameters, including any variations due to noise. This can be mitigated to some extent through batch-normalization or dropout layers. However, considering the relatively small dataset used in this study, we also test if a smaller model comprised of a few stacked convolutional layers could be optimized to approximate the performance of even deeper models. In previous studies [9], this has been shown to be the case. Specifically, here we adopt a similar architecture to VGG-Net [20], however in a shallow model architecture shown in Figure 2.6. VGG-Net is one of the first successful DL models and was one of the first to show that stacking 3×3 convolutional layers with increasing filter counts can be sufficient for modeling even complex image representations. In this work, we adopt the same basic structure



Figure 2.7: The 2-lead wearable ECG sensor used in the experiments and developed by Imec[1].

and test a range of model depths, from 3 × [Conv2D(3 × 3)] to 15 × [Conv2D(3 × 3)], and compare the most optimal configuration against ResNet-50 and the other baseline models. Our models feature [Conv2D(3 × 3), 32] + [MaxPool2D(2x2)] + [BatchNorm] layer blocks, with consecutive blocks doubling the filter count of the previous. Similar to the other models tested, the shallow models are topped with 2 × [Dense (128)] layers followed by a soft-max layer for classification.

2.4 Data Acquisition

For the following experiments we use the same work stress dataset we introduced in an earlier work [49]. It consists of ECG recordings of subjects performing naturalistic desk work activities in front of a PC. The protocol was reviewed and approved by the Research Ethics Committee at Osaka University prior to actual implementation. Data was collected from 4 healthy male participants aged 20-32. All subjects were graduate school students from Osaka University.

The physiological recording device used was a wearable ECG sensor developed by Imec[1]. The device is capabale of recording ECG at a sampling frequency of 256Hz as well as accelerometer motion data at 128Hz. The 3-lead wearable ECG sensor is worn with replaceable leads placed diagonally over the heart as shown in Figure 2.7. The wearable configuration enables unconstrained and continuous measurement of the participants' heart rate while they perform various tasks. RR intervals were then extracted from the raw ECG signal using a QRS detection toolkit [50], which were then used to build the

[1] `https://www.imec-int.com/nl/expertise/wearables`

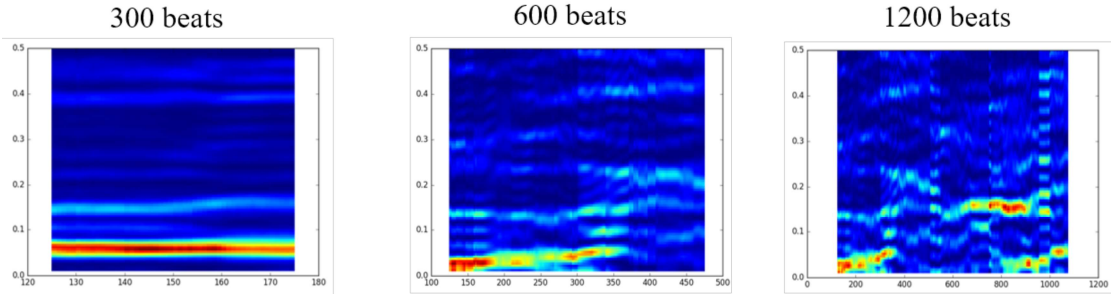


Figure 2.8: Image samples of extracted ECG spectrograms at different window widths.

LS spectrograms.

Table 2.2: Stress Annotations Distribution

	Tasks	**Very Low**	**Low**	**High**	**Very High**
Total	196	46	78	53	18
Percentage		23.47%	39.80%	27.04%	9.18%

During the observation sessions, participants were urged to perform only work activities and to avoid large movements like standing up or walking. However, specific activity restrictions were not enforced so subjects could perform natural activities. After the recording session, subjects were requested to perform activity and stress-level annotations, as well as answer a set of psychology questionnaires. Based on the provided annotations, majority of the stressful activities consisted of debugging programs and preparing academic materials (i.e., papers and presentations). Subjects identified a total of 196 work task segments, from which 184 usable segments were extracted. Extracted task segments were required to be within 5 to 30 minutes long in order to allow for reliable short-term HRV analysis [51]. Segments shorter than 5 minutes were removed in order to maintain the stability of the HRV features based on recommendations from previous literature [52]. Annotated stress levels were labeled based on a 4-point Likert scale ranging from 1 (very low stress) to 4 (very high stress). A summary of the label distributions is shown in Table 2.2. In total, this provided a total of 20 hours of ECG data for training and testing models.

Table 2.3: Window Size Test Results Confusion Matrices

(a) 300-beats

*	Very Low	Low	High	Very High
Very Low	**1.00**	0.00	0.00	0.00
Low	0.06	**0.62**	0.23	0.09
High	0.06	0.25	**0.62**	0.08
Very High	0.00	0.00	0.00	**1.00**

(b) 600-beats

*	Very Low	Low	High	Very High
Very Low	**0.97**	0.00	0.00	0.03
Low	0.00	**0.71**	0.12	0.17
High	0.04	0.24	**0.59**	0.13
Very High	0.00	0.00	0.00	**1.00**

(c) 1,200-beats

*	Very Low	Low	High	Very High
Very Low	**0.44**	0.11	0.44	0.00
Low	0.00	**0.65**	0.32	0.03
High	0.00	0.18	**0.79**	0.02
Very High	0.00	0.03	0.11	**0.86**

2.5 Experiments and Results

2.5.1 Model Training and Optimization

For the following experiments, we applied a sliding window method to divide the sessions into segments. This method could potentially be extended towards real-time processing in the future. To identify the appropriate window lengths, a number of trials were conducted, and it was observed that decreasing window lengths would increase the stability of the representation at the cost of expressiveness. Meanwhile, longer segments were more expressive but tended to overlap across samples with different labels (See Figure 2.8).

To determine the best window length, simple CNN models were trained on various window lengths from 60 to 1,200 beats. Notable differences in performance are summarized in Table 2.3 which shows the confusion matrices of models trained on 300, 600, and 1,200-beat windows. The finer empirical tests showed that 505-beat windows (∼6 mins) presented the best balance between stability and expressiveness while staying above the 5-minute minimum length recommended for accurate short-term HRV analysis. Thus, all following experiments would use this window length configuration.

For labeling the overlapping windows segments, a modal labelling scheme was adopted. Additionally, segments with overlapping labels were marked as either “stable” or “transitions” based on a 70% majority threshold. That is, segments with a majority label that constitute less than the threshold should be considered to be ambiguously labeled and should not be included in training its corresponding class label. Instead, an additional “transitions” class could be used for the ambiguous segments. After segmentation and rejection of short or noisy segments, the final experiment dataset was comprised of

~1,500 windowed samples which were either labeled based on the 4-level stress annotation or as transitions.

2.5.2 Comparison Experiments

One objective of the comparison experiments is to test whether the proposed method performs better than traditional approaches that utilize manually extracted HRV features. To build these baseline models, multi-layer perceptrons (MLP) and denoising autoencoder (DA) networks were selected to train on the following HRV features: Average of NN intervals (AVNN), Standard deviation of NN intervals (SDNN), root-mean-squared differences between adjacent NN intervals (rMSSD), percentage of differences between adjacent NN intervals greater than 50ms (pNN50), spectral power measures of NN intervals of varying frequencies (ULF, VLF, LF, HF) and the ratio of low to high frequency power (LF/HF).

The MLP architecture mirrored the supervised component of the deep models, with 2 hidden layers containing [128,128] hidden units and rectified linear unit (ReLU) activation. ReLU is a piecewise linear function that outputs the input directly if it is positive, and otherwise outputs zero. It is the preferred activation function for hidden units in most modern architectures since it encourages sparsity and reduces the effects of the vanishing gradient. Meanwhile, the DA features a 2-layer stacked architecture with [10,6] units for enforcing a slightly undercomplete embedding for feature selection. This was topped by a dense layer with 128 units for supervised feature learning. Overcomplete DAs (i.e., >12 embedding units) were also tested, however these had severely degraded performance compared to the slightly compressed model. The MLP provides a good baseline for traditional methods that discover shallow non-linear features from the input data, similar to SVMs. On the other hand, the DA model should provide a good representation of how shallow feature selection and discovery models would perform on the same input data.

Next, to provide a benchmark for state-of-the-art pre-built architectures adopted to the proposed approach, a deep network based on ResNet-50 was trained. Meanwhile, to represent the potential of compact optimized models, a shallow 5-layer CNN was trained. The compact model may also be used to verify the hierarchical feature requirements of the dataset. In summary, the following models were trained for comparison:

- HRV-MLP—similar to traditional SVM and NN models used in most preceding

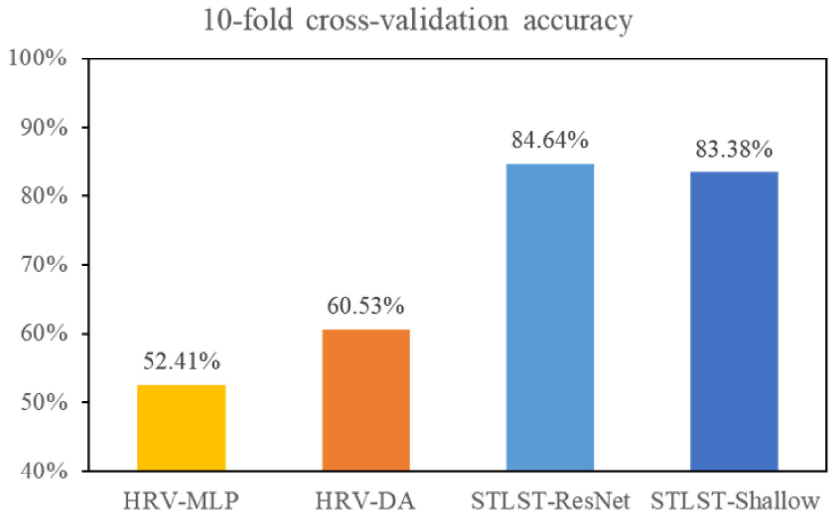


Figure 2.9: Comparison of test performance results between models trained on standard HRV features (MLP and DA) and deep models trained on STLST heart rate spectrum data (ResNet and Shallow).

studies

- HRV-DA [53]—relatively simple DL architecture that prioritizes feature selection and discovery of compact embeddings by randomly corrupting data during training
- STLST-ResNet [1]—one of the gold-standard DL architectures designed for learning deep image features, comprised of 50-layers including residual blocks and skip connections
- STLST-Shallow—relatively shallow 2D CNN featuring 5 $\times$ [Conv2D(3 $\times$ 3) + MaxPool(2 $\times$ 2) + Dropout] layer blocks

All models were topped with a 2 $\times$ [Dense (128)] layers and a final soft-max classification layer to conduct the final prediction task. All models were trained until convergence or until early stopping is triggered based on the validation loss. Performance was validated and measured using stratified 10-fold cross-validation accuracy.

Summary of Results and Comparisons

Based on the summary of results for the baseline comparisons shown in Figure 2.9, the STLST-trained models performed significantly better than the baseline models trained

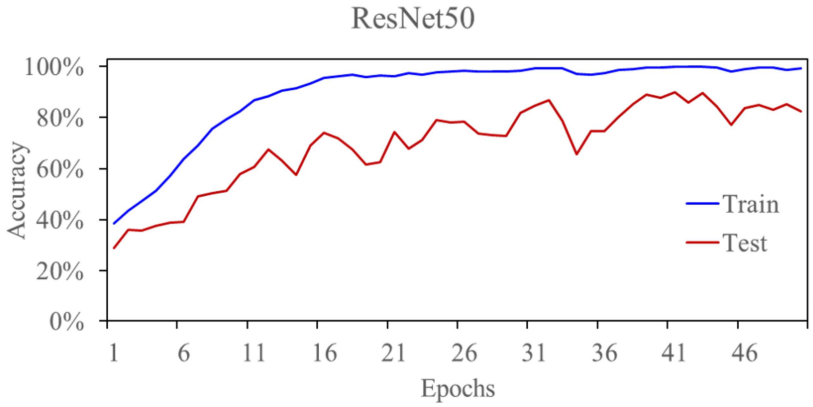


Figure 2.10: Training and testing performance of ResNet-50 model.

on HRV. In the case of the baselines trained on traditional extracted HRV features, the HRV-MLP model achieved 52% test accuracy. Meanwhile the HRV-DA model saw increased accuracy to about 60%. While these results may seem weak at first glance, it should be noted that this was accuracy on a multi-class problem. Given that there were effectively 5 classes, 4 stress levels and "transitions", a random classifier would be expected to perform at around ∼20% accuracy on the same problem. Therefore, these could be taken as notably better than some previous works [36] which achieved similar percentages, but for binary classification. The performance improvement of DA over MLP is interesting in that it shows that, despite the relatively small number of input features (14), performance can still be improved simply through the more efficient latent feature selection of modern ML architectures. Meanwhile, in the case of the DL models, both models performed much better than the HRV-trained baselines. Even the shallow model was almost able to match the performance of ResNet-50 despite having much fewer parameters. The following sections discuss in more detail the performance of the STLST models.

Deep Model Results

For the spectral model results (See Figure 2.10), ResNet-50 was able to achieve over 80% testing accuracy, which was well above both of the traditional HRV baselines. These significant improvements over the baseline models highlight the potential of the increased expressiveness provided by learning from raw spectral data using the feature discovery ability of DL algorithms. While the traditional indices used in HRV analysis are useful for helping humans understand physical phenomena, they can also obfuscate some fea-

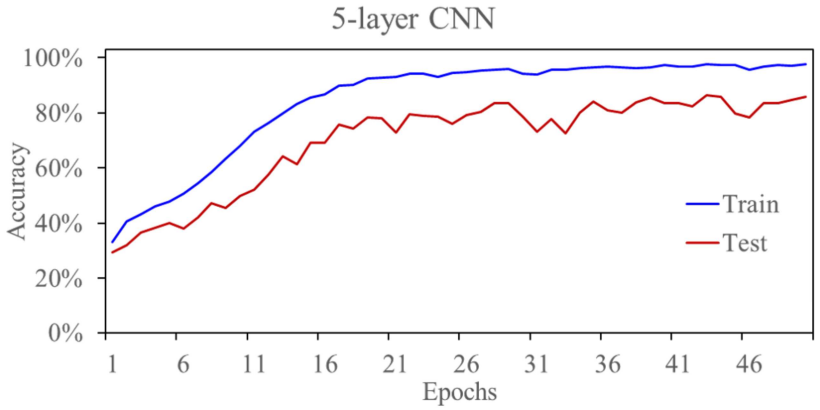


Figure 2.11: Training and testing performance of the 5-layer CNN model.

tures variations that could also be useful. By allowing the ML model to learn from a spatial-spectral representation, we are able to train an effective model using a relatively small dataset. Additionally, the proposed data representation which obfuscated ECG morphology allowed the DL model to largely avoid overfitting to signal noise features. But aside from effectiveness, it is also necessary to compare models basd on efficiency. Building deeper models naturally comes at the cost of increased complexity. In the case of our standard 50-layer ResNet model, it involved training over 20 million parameters. While it showed good performance on our small dataset, the large number of parameters can increase the risk of the model to overfitting to irrelevant noise features as the signal-to-noise ratio of the training dataset shifts.

Shallow Model Results

To counteract the problem of increasing complexity, we attempted to train a more compact CNN model based on the VGG-Net architecture. After testing various architecture depths, the most efficient model consisted of 5 convolutional layers. Here, a layer is actually a block of [Conv2D(3×3)] + [MaxPool2D(2x2)] + [BatchNorm] layers. Shallower models less than 3 layers were not able to provide enough variability to learn all the variations from the training set, while models deeper than 8 layers showed increasing overfitting. The deeper models probably also suffered from vanishing gradients due to the relatively simple construction of our models which did not use residual blocks. Models below 8 layers were shallow enough that residual blocks were unnecessary. An additional benefit of these shallow models was that training was much faster than ResNet.

Interestingly, the optimal 5-layer shallow model was able to achieve comparable

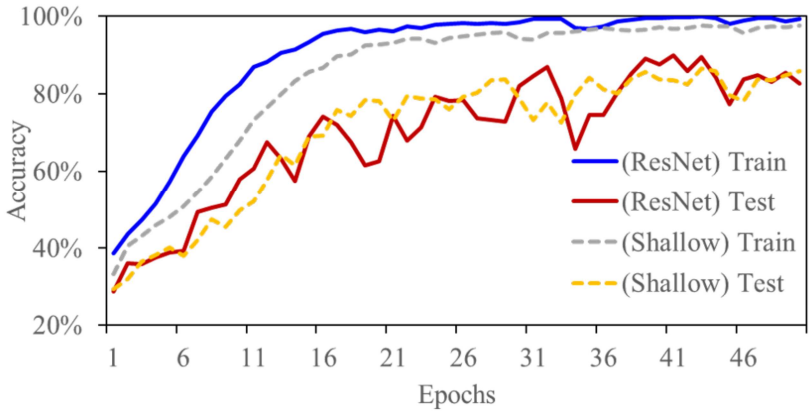


Figure 2.12: Overlapped performance results of ResNet-50 and the shallow 5-layer CNN model.

performance to the much deeper ResNet-50 model (See Figure 2.11). Figure 2.12 shows a direct comparison of the training process of both the deep and shallow models. Despite using less than 1 million parameters, as opposed to the over 25 million of ResNet-50, the shallow model is able to match the performance of the much larger ResNet model while also avoiding overfitting. Actual training time could also reduced by up to 5x which may be useful in situations where occasional retraining is necessary.

2.5.3 ROC Analysis

Next, we wish to analyze how well the model is able to discriminate between the different stress classes. For this we look at the receiver operating characteristic (ROC) curves for each of the classes, using a one versus all methodology. As seen in Figure 2.13, using the deep ResNet model has a high rate of discrimination for the edge labels (i.e., very low and very high). Meanwhile, it is slightly more challenging to correctly distinguish the median labels (low and high). This could indicate that there is more ambiguity with regards to the physiological signals shown for moderate stress levels. Given that the stress levels were expected to be ordinal, it would be easier to separate classes on either side of the spectrum rather than in the middle. This may be a sign that the features discovered by the model follow the same intended ordinality as the stress labels.

Alternatively, this could also be due to labelling inconsistencies common to psychological labels. Simply, it is easier to write accurate labels for extreme emotions rather than moderate ones. Future work should focus on exploring this further in order to determine a more efficient annotation methodology, or possibly even correct or omit

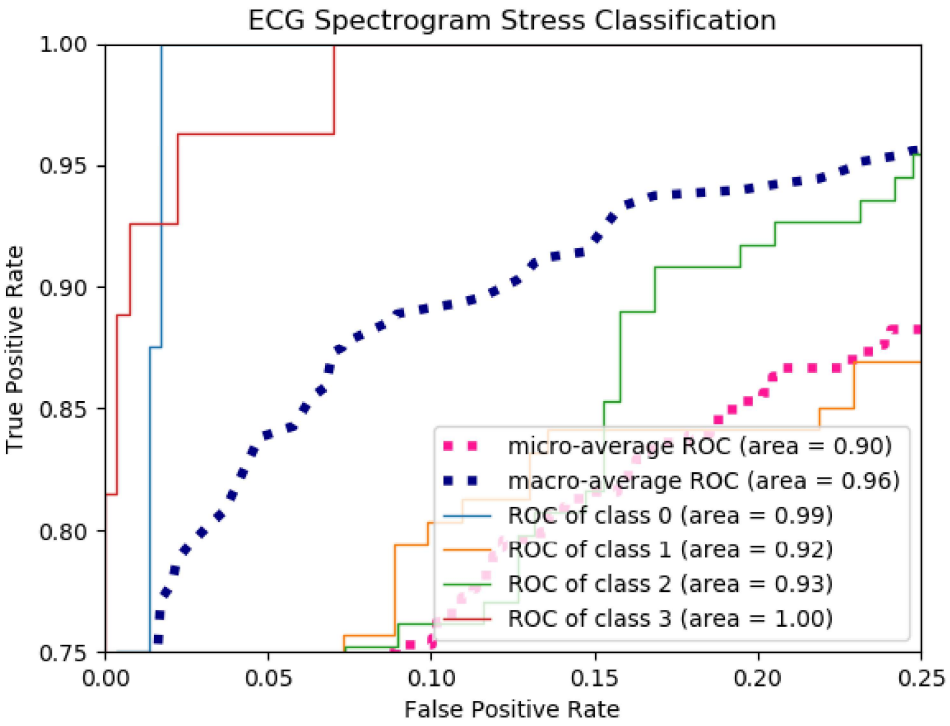


Figure 2.13: ROC for the classification results of the ResNet-50 model.

inconsistent labels.

2.6 Summary and Future Work

In this work we trained a deep model of work stress using a spectral representation of heart rate extracted using short-time Lomb-Scargle transform. To test the approach, stress data was gathered from subjects performing naturalistic work activities, and reactivity tests were performed on both shallow and deep ML architectures. The experiments showed that the deep models trained on the spectral representation could easily outperform those trained on the gold-standard HRV features for detecting naturalistic stress levels. Additionally, we show that even relatively shallow models featuring at least 5 convolutional layers can match the test performance of the more complex ResNet-50. This could indicate only a few shallow hierarchical features may be important for effectively modelling naturalistic stress from heart rate, or that deeper architectures are negatively affected by the non-stationarity of the signal. It should also be noted that one-dimensional signals such as RR intervals are inherently limited in their expressiveness and the amount of variations they can represent. Moreover, as a secondary effect

of autonomic neural activity, the ECG signal is fundamentally limited in how much information it can encode about the psychological state of a subject. While it is effective as a non-invasive method of observing stress, it may not be as effective for other muted responses such as emotions. Therefore, it may be worthwhile to investigate the potential of this method when applied to more complex multi-channel signals which are closer to the seat of neural activity, the brain. EEG signals offer a non-invasive measure of neural activity and could be the future of HCI, however it also suffers greatly from the effects of signal noise. If the proposed methods can adapted to and be proven effective on EEG signal noise, this may allow us to work towards more direct measures of psychological states, and possibly even emotion.

Therefore, in the next studies, we attempt to apply and refine the proposed methods for EEG signals. Due to the increased sensitivity of EEG signals, we first attempt to model meditation as this provides a good benchmark for relatively stable EEG readings. This also allows us to contribute to a rarely studied field in modern affective computing and propose methods to expand it further.

Chapter 3

Detecting Meditative States with EEG and Deep Visual Models

3.1 Introduction

In the previous work, we proposed a method for training ECG-Stress models using a robust training scheme for learning HRV-like features. We noted its efficacy at allowing DL models to learn from the relevant features of the data, while avoiding the common pitfalls of end-to-end models when learning on noisy signals. To build on top of the work, we now propose an adaptation of the model applied to EEG data. In particular, we use EEG to model meditation, as an analogue to stress. This change in focus is necessary since compared to ECG, EEG is a much more sensitive multi-channel signal that can be easily affected by various stimuli. In addition, it introduces an additional spatial dimension that is dependent on the number and positioning of electrodes used in the EEG device. Therefore, it can be difficult to measure the EEG of subjects doing work, since it involves the compounding of multiple stimuli that could cause overlapping brainwave responses. Meditation provides a good starting point for investigation since it generally requires subjects to focus and avoid external distractions. It can also be performed with the eyes closed, which rules out the effects of visual stimuli. This helps to simplify the data gathering methodology and enables the dataset to be more focused on the effects of meditation. At the same time, it allows us to identify the most basic limitations and noise sources in the data and provides a good baseline for the applicability of the method to other psychological phenomena. Essentially, any issues encountered in the meditation data, with regards to noise and signal quality, will most likely be present under other experimental conditions. Therefore, the solutions

we present through this work will likely be relevant to most other EEG works moving forward.

In this study, we focus on a type of meditation which falls under mindfulness meditation. Among the many interventions and treatments currently being used to treat stress-related mental and physical illnesses, one that has gained wide adoption in recent years is mindfulness-based stress reduction (MBSR) [54]. Essentially, these are sets of mindfulness meditation practices designed to reduce a subjects' reactivity to stressors. In less than a decade, MBSR has grown into a billion-dollar industry backed by a growing number of major U.S. companies [55]. Its growth is due in part to the belief in its effectiveness at reducing mental and emotional strain, as well as its ability to improve workplace productivity. However, despite the widespread faith in the results of the practice, there is yet to be a consensus on how to objectively measure its effects.

One way to build an unbiased measure of meditation effects is to measure the pyschological and physiological impact of the practice. The former is generally the target of psychological studies and is typically solved through the use of questionnaires and statistical diagnosis tools [56]. Meanwhile, the latter offers an opportunity to apply modern technologies to track and better understand the long term effects of meditation on stress [57]. Wearable EEG devices are one such technology that could hold the key to better understanding the effects of meditation. But before that, it may be necessary to discover effective methods for building scalable models from physiological data. Until now, studies on building models of the neurological effects of meditation remain limited due to a wide range of challenges including signal noise, label subjectivity, inter-subject differences and the difficulty of obtaining clean and complete real world EEG datasets [58]. Many of these challenges are inherent to physiological sensor data, which are known to be sensitive to various external sources of noise. This is especially the case for EEG, where even partial degradation of the signal in a single channel often results in the exclusion of entire sessions of data. This not only limits the size of usable datasets for training ML models, but also hampers the practical usability of consumer EEG headsets.

3.1.1 Challenges for Building Deep Models for EEG

EEG is prone to many types of noise; this necessitates the use of some combination of signal cleaning and enhancement techniques in the processing pipeline. The most common of these are signal filters designed to filter out nuisance signals from certain frequency

bands not directly related to brain signals. Additionally, it is occasionally necessary to decompose and analyze the components of the cleaned signal to further improve the signal-to-noise ratio. Of these, some of the most commonly used are independent component analysis (ICA) [59], artifact subspace reconstruction (ASR) [60], common average reference, surface Laplacian, and principal component analysis (PCA) [61]. These techniques can be used to identify noise sources and remove them from the signal. However, most of these have the disadvantage of requiring manual intervention by the researcher to distinguish and remove the noise sources from the decomposed signal.

In recent years, similar forms of manual filtering and feature engineering have largely been made obsolete in other signal-processing fields, such as computer vision, due to the prominence of deep learning (DL). In response, recent studies in EEG emotion analysis have also explored machine learning as a strategy to automatically discover the most relevant features from the EEG signal through different DL techniques and architectures [62]. In this work [63], we propose a method for building practical EEG-based ML models of mindfulness meditation by applying visual EEG signal representations and visual DL models. We show how the model is able to overcome some of the most common issues surrounding EEG datasets, namely noise sensitivity and scalability. In addition, we analyze the types of features discovered by the model and show how model architectures can be optimized for higher efficiency.

3.2 Related Works

3.2.1 Image-based Deep EEG Architectures

Bashivan et al. [8] were one of the first to successfully implement deep visual models for EEG analysis. In their study, they proposed the use of EEG spectral topography maps to train convolutional and recurrent neural networks (CNNs and RNNs) to model the effects of mental workload. Their study found that using the spatial representation allowed their models to cope with inter- and intra-subject differences as well signal noise. This was reflected by performance improvements, particularly in subject-independent classification. This study was the primary inspiration for the proposed work, and provides the groundwork for the general methodology. Our work differs in the phenomenon being studied and the specific architectural optimizations applied to reach a suitable model. In a similar work, Thodoroff et al. [64] used spectral representations of EEG for detecting seizures. Their approach closely mirrored that of [8] in that they used

a combination of CNNs and RNNs to train the models. However, [64] used long-term short-term memory (LSTM) components in place of RNNs. While this would normally be detrimental to the performance of a model trained on a small dataset, they applied various techniques for sample-efficient training, including pre-training at multiple stages, transfer learning to specific patients, as well as using an ensemble of 3 models with different weight initializations. They found that their method not only beat a benchmark method for cross-patient classification, but the model was also robust to missing channels. However, while the model showed good sensitivity to seizures, small parameter changes could have a large impact on positive rates. It's possible that the high data requirements of recurrent networks could be detrimental for training on small datasets. Thus, in the current work we first focus on building optimized representations based on CNN models.

Among works that used purely CNN models, Vilamala et al. [65] stands out as one of the most notable yet straightforward approaches. In [65], researchers used multitaper spectral analysis to create visually interpretable images of sleep patterns from EEG signals and used these to train a visual DL model. Specifically, they used 2D spectrograms from single-channel EEG to train VGGNet, an off-the-shelf model for image classification. Their experiments show that VGG16, a 16-layer version of VGGNet, can effectively be trained to accurately detect 5 different sleep stages from EEG data at up to 86% accuracy. Tests comparing the use of VGGNet as a pure feature extractor and as a fully-trained model also show that fully re-training the model can improve performance slightly. In our work, we conduct a similar comparison, however instead we compare a fully-trained visual model and an optimized shallow model. A more advanced approach was proposed in Jiao et al. [66] where they used 3D convolutions and applied to modeling cognitive workloads. While their approach is also promising, for now we focus on training spatial-spectral models to avoid the additional data requirements of adding a temporal component.

Schirrmeister et al. [9] propose a different approach, instead proposing an end-to-end model for EEG motor visualizations. Their approach involved analyzing DL model activations and plotting the correlation between those model activations and a set of known EEG signal measurements. Their work was among the first to prove that DL models can learn the spatial topology of EEG spectral features. Furthermore, results showed that DL models are capable of automatically discovering traditional EEG features used

by humans. It was a big step towards creating interpretable deep EEG models. However, the primary limitation of the work was the time and expertise necessary to test and filter all the complex correlations between model weights and known EEG signal features. While we do not go as deep into feature analysis in the current work, we also analyze model feature activations in order to design a more optimized model.

3.2.2 EEG Models for Detecting Meditation and Attention

One of the earliest modern investigations on the applicability of ML to meditation was by Dixit [67]. The work studied how ML models could be used to study the effects of meditation. Using a single-electrode EEG sensor (Neurosky), they measured the meditation practice of 31 experienced meditators from Tibetan and Indian monastic backgrounds. To build the dataset, the subjects each performed a 5-minute eyes-closed resting period followed by 15 minutes of free form meditation. The samples were then segmented into 1-second windows and processed to extract power spectrum over eight common brainwave frequency bands. Results from training an SVM model showed that subject-dependent classification performance could reach an average of 75.7% accuracy. This was one of the first works to show the potential of EEG-based ML models for detecting meditation. Our work follows a similar methodology, however we investigate the meditation practice of novice and intermediate practitioners, therefore the effects are expected to be much more subtle. Additionally, we gather data in a natural office setting with natural noise sources. However, to compensate for the additional noise, we use an EEG device with more electrodes (16 channels) and combine this with our proposed ML approach.

Ahani et al. [68] also used SVMs, but this time to detect the long-term effects of mindfulness meditation practice. Using data gathered from 34 novice meditators who conducted a 6-week meditation workshop, as well as a set of control subjects, they trained models to distinguish between EEG signals of meditators and non-meditators. They found that spectral features of respiration rate were just as effective as EEG rates for the classification task. Additionally, their results supported previous findings that theta, alpha, and beta frequencies (4-32Hz) are generally enough to assess meditation. In another work, Djamal et al. [69] attempted to detect states of inattention using EEG and ML by using data from blue light stimulation to simulate attentive and inattentive states. Using data recorded with a 20-channel EEG device and by extracting power spectrum

coefficients of 5 brainwave bands between 5-30Hz, they trained an SVM model which reached an average of ∼76% accuracy. Their study found that features in the 5-30Hz range were optimal for this classification task, and noted that Theta bands (5-8Hz) were some of the most distinctive features. While attention may be functionally different from meditation, there may be an overlap in the brainwave bands they affect. Our approach is similar in that we target the same general frequency ranges (i.e., theta, alpha, and beta). Also, similar to the manual analysis done in [68], we use an ML approach that includes spatial information about EEG signals to better classify meditative states.

Another more recent work by Lin and Li [70] proposed a method for distinguishing meditation practice from other states. They compared results from experienced meditators and novices, and found that entropy was effective at characterizing the difference in regularity between meditative and non-meditative states. They trained SVM and GMM models on different window sizes and found that classification performance improved with larger windows and tapered off at around 10 seconds. This coincides with our own findings where window sizes above 4 seconds improved model performance. However, a notable weakness in their analysis is that they did not measure test accuracy, so there is no way to confirm the generalizability of their models. In contrast, our benchmarking approach more closely matches the method in Dixit [67], where cross-validation is used to measure the generalization performance of the models. In addition, we use a similar target frequency range to [68, 69], while including the automated equivalent of the spatial analysis done by [69].

3.3 Methodology

This work proposes an automatic approach that facilitates robustness to external noise sources by utilizing spatial-spectral EEG representations together with visual DL models to model mindfulness meditation. To the best of our knowledge, this is the first attempt to apply modern DL models for modeling naturalistic stress from novice meditators. To limit the effects of isolated noise that affects only a subset of electrodes, we adopt a similar intermediate spectral-spatial representation to Bashivan et al. [8]. However, unlike past works we also apply a framework that involves actively removing noisy electrodes while still retaining the remainder of the data to help train the ML model. Rather than allowing the noisy channel to negatively affect its neighbors, we allow the remaining neighbors to account for the data of the missing channel. The results of the

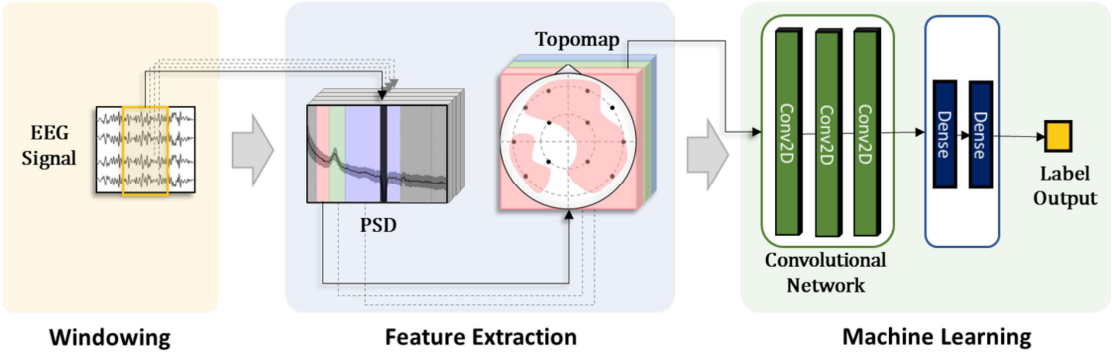


Figure 3.1: Architecture diagram of the proposed deep visual EEG-Meditation model.

experiments can help to assess the viability of building training datasets from partially noisy EEG data. If successful, it can lead to a significant reduction of the costs of acquiring EEG datasets by enabling use of partially noisy samples. Additionally, the results should provide a good benchmark for assessing the strengths and limitations of CNN-based models for analyzing EEG data. Many of the existing works require the use of temporal DL structures like LSTMs, however this often drastically increases the data requirements of the model and makes it more vulnerable to overfitting. Figure 3.1 shows the overall architecture diagram of the proposed method. It involves windowing of EEG streams to capture fluctuations of brainwaves over time, extraction of spectral topomaps to encode the spectral-spatial components of the data, and finally modeling of meditation using DL architectures.

3.3.1 Signal Preprocessing

The raw EEG data recorded by the Ultracortex Mark IV consisted of 16 channels sampled at 250Hz. Basic noise filtering was performed using a combination of high and low pass Butterworth filters to remove signal noise below 2Hz and above 75Hz, as well as a notch filter to remove 50Hz powerline noise. Next, artifact subspace reconstruction (ASR) [71] was applied primarily to reduce occasional motion-related artifacts. ASR is a non-stationary, real-time method that is good at removing occasional large-amplitude artifacts such as those resulting from movement or loose electrode contact. It is often used in conjunction with independent component analysis (ICA) to perform signal source localization. Finally, a threshold method was used to completely remove channels that exhibited excessive variance, possibly due to poor contact with the scalp. Bi-cubic interpolation would replace these missing values with signals from other nearby elec-

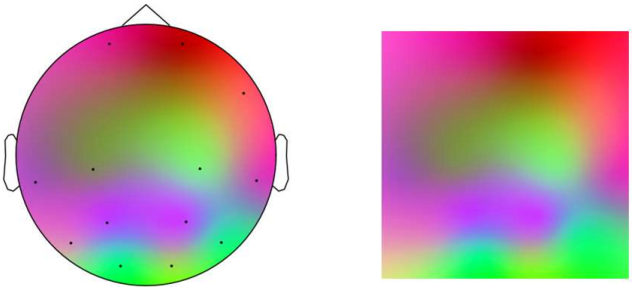


Figure 3.2: 3-band EEG topomap image of a subject during post-meditative rest. RGB color channels represent theta, alpha and beta brainwave power, respectively. In this work, the spherical mapping (Left) is converted into a rectangular mapping (Right) to better accommodate rectangular convolutions.

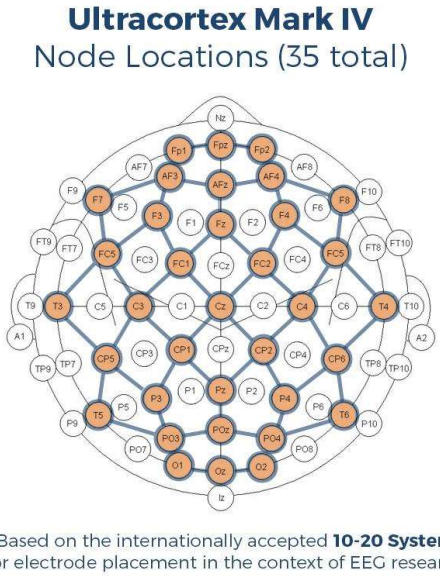


Figure 3.3: Standard 10-20 electrode configuration used by the Ultracortex Mark IV EEG headset

trodes. This was done to reduce the effect of channels containing mostly just noise on the topographs extracted in the next phase. Based on the findings from previous studies [8, 64], visual DL models should be able to cope well despite the missing data channels. This approach maximizes utilization of data samples, including those with common data degradation issues which are commonplace for EEG signals.

3.3.2 Generating EEG Visualizations

We generated EEG visualizations using a 4-second window with 2-second overlap to extract spectral features from the data by multitaper spectral analysis. This produces a power spectrum for each EEG channel, for a total of 16 EEG spectra. Next, for

each channel, we average the values of frequency bins corresponding to brainwave bands related to meditation, namely: theta (4-7Hz), alpha (8-12Hz), and beta (13-30Hz). The spectral power values of all channels are then mapped to a 2D projection based on the locations of each electrode on the scalp (See Figure 3.3). These values are then made to fill the 2D plane using bicubic interpolation to produce a spectral topograph for each brainwave band. By combining the topographs from each brainwave band into a single matrix, the resulting 3-channel data-structure can be visualized as an image (See Figure 3.2). In this experiment, the visualizations were generated using the MNE-Python package [72]. Rectangular topomaps were generated using extrapolation of outside points beyond the head skirt and subsequent cropping. Circular cropping was used to visualize and verify the topomaps, however a rectangular crop was used for generating the dataset in order to better accommodate the rectangular convolutional kernels. Additionally, we down-sample the resulting images to 64×64 px and normalize their values based on the z-score values across the dataset to facilitate efficient ML training.

3.3.3 Model Architectures

Deep ResNet Model

ResNet [1] was chosen as the baseline machine learning model due to its reputation as a good overall network for a wide range of visual classification tasks. Thus, it serves as a good representation for most off-the-shelf yet state-of-the-art visual DL models. It should also serve as a good demonstration of some of the drawbacks of simply using DL models on EEG datasets [73]. ResNet enables building very deep architectures by using residual blocks (See Figure 2.4) which circumvents the vanishing gradient problem [48] by featuring skip connections. While this allows it to learn very complex patterns, it also leaves it open to learning unrelated patterns in the training data. So, despite their flexibility, very deep networks are not always ideal due to their increased complexity, longer training times, and susceptibility to overfitting. So, as an alternative, we also train a more compact optimized network benchmarked on ResNet-50.

Shallow VGG-Net Model

Similar to our previous work [63], we adopt an architecture based on VGG-Net [20] another classical modern image network. VGG-Net provides a simple but scalable ar-

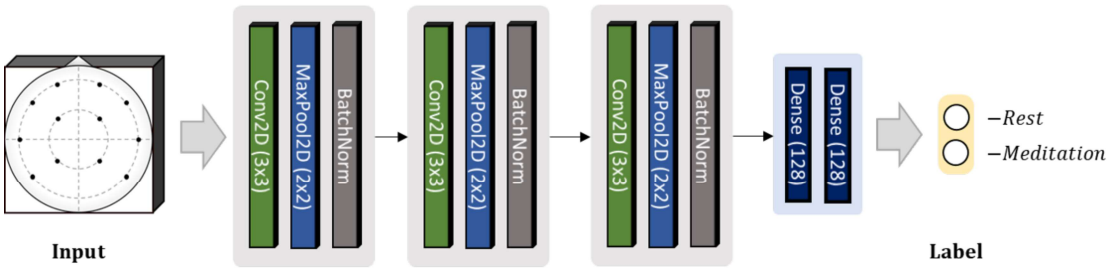


Figure 3.4: Shallow model architecture

chitecture that uses stacked convolutional layers with a fixed 3 × 3 receptive field, 1 pixel stride, and ReLU activation (See Figure 3.4). Adding a 2 × 2 maxpool layer on top of each convolutional layer expands the receptive field of successive layers. The original implementation also made use of a dropout layer to prevent co-adaptation of network weights. However, in this implementation we found batch normalization layers to be more effective and also help the model converge faster. By doubling the number of filters after each additional convolutional layer, the network is able to encode more complex patterns with fewer overall parameters.

The shallow model shown in Figure 3.4 is comprised of 3 convolutional layers topped with 2 dense layers for combining all the features to make the final prediction. This model was tuned to closely match the learning capacity of ResNet-50 the dataset. Testing showed that exponential linear unit (ELU) activation performed better than ReLU, so it was used for all non-linear activations. Unlike ReLU, ELU features a smoother natural exponential activation curve at inputs below zero. This results in more consistent training gradients than ReLU, at the cost of slightly weaker sparsity constraints. Finally, we inserted 2 dense layers each with 128 units before the final 2-output softmax classification layer. Dropout was tested but was ultimately not included since it adversely affected the performance results, in its place batch normalization worked better on this dataset. Further regularization did not seem necessary since very little overfitting was observed in the final model, likely due to its compactness. We also tested average pooling layers, however results were not significantly better than max pooling and even degraded performance in some scenarios.

3.4 Dataset Acquisition

The dataset featured in this study was gathered using the OpenBCI framework, specifically we use the Ultracortex Mark IV EEG headset in a 16-channel configuration[74].

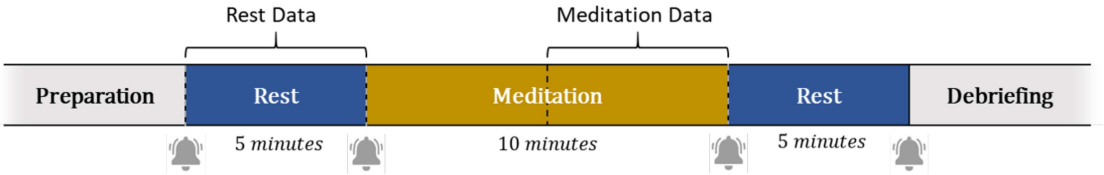


Figure 3.5: Overview of data gathering experiments and data segmentation.

The channel layout follows the standard 10-20 electrode configuration shown in Figure 3.3. Specifically, we used the following EEG channels: Fp1, Fp2, C3, C4, T5, T6, O1, O2, F7, F8, F3, F4, T3, T4, P3, P4, and linked-mastoids as reference.

To conduct the meditation sessions, we devised an easily-reproducible experimental protocol based on the UCLA breath meditation procedure described in Winston [75]. Each mindfulness meditation session included the following sequential phases: 1) 5-minute baseline eyes-closed relaxation phase, 2) 10-minute guided breath meditation phase, 3) 5-minute relaxation phase. The beginning and end of each phase were timed and signaled by ringing a bell. Refer to Figure 3.5 for an overview of the experimental flow. For the entire duration of the session, subjects wore the 16-channel dry electrode EEG headset which recorded their brainwave data. Only data from the first rest period are considered for the rest dataset. Meanwhile, since the onset of meditative states can be hard to predict, only the last 5 minutes of the meditation segments were used for the meditation class. This also had the positive side-effect of helping to balance the class distribution of the dataset. The relatively simple procedure is designed to make it easy to replicate in hopes that it could be extended into a larger dataset in the future. The protocol was reviewed and approved by the Research Ethics Board of the Japanese company in order to comply with safety and data privacy standards prior to actual implementation.

The entire procedure was conducted with the guidance of an experienced meditation instructor who guided the participants to perform the meditation practice. Aside from the physiological data, we acquired psychological wellness profiles from each subject using a number of standardized psychology questionnaires to measure mindfulness, stress, and overall well-being. This data can be used to characterize the subjects' mental health and could be used as ground truth for later work. From the survey, we obtained psychological profile samples (n = 164) which were analyzed using bivariate correlation (Pearson's R). Here, we found strong correlations (r = -0.66) between trait anxiety

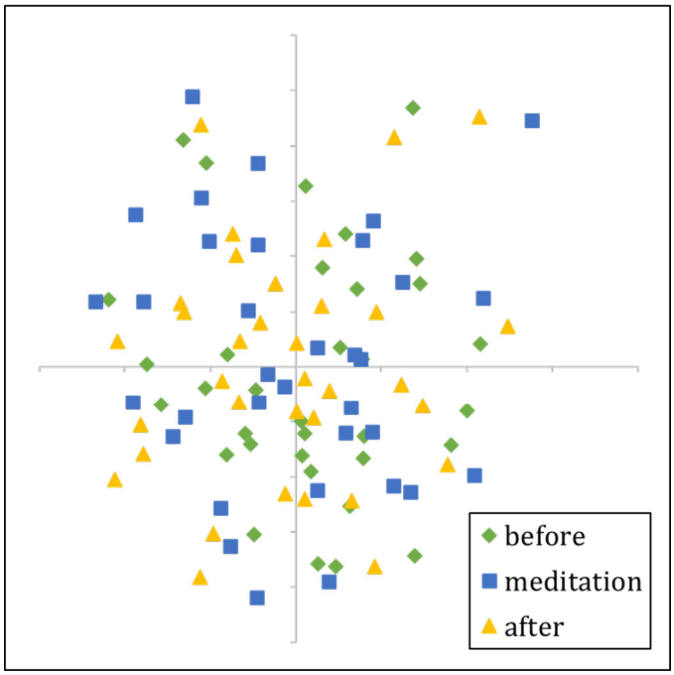


Figure 3.6: EEG data mapping based on top 2 principal components.

(i.e., long-term stress) and five-factor mindfulness, as well as significant negative correlations ($r < -0.4$) between trait anxiety and the other wellness indices (life satisfaction, psychological safety, and work meaning).

Among those who answered the survey, 36 subjects (23 males and 13 females) volunteered to take part in the EEG meditation experiment. Moreover, from the 36 subjects, data from 2 subjects (2 males) had to be excluded due to excessive signal loss in more than 6 channels. All participants were aware that we were studying the effects of mindfulness.

3.4.1 Exploratory Analysis of Linear Features

Initial visual and statistical analysis of the raw data did not reveal any strong and consistent distinction between resting and meditative states across subjects. On the other hand, the differences between samples from different users was more obvious. This was confirmed through principal component analysis (PCA) applied to the visual data, as seen in Figure 3.6. The primary axes of variance appear to completely overlap all 3 classes representing the different phases of the session. To test the effectiveness of supervised learning for finding an effective linear mapping, we also tested linear discriminant analysis (LDA), which resulted in the 2D feature map shown in Figure 3.7. It shows a clear distinction between the meditation segments (blue) and the rest

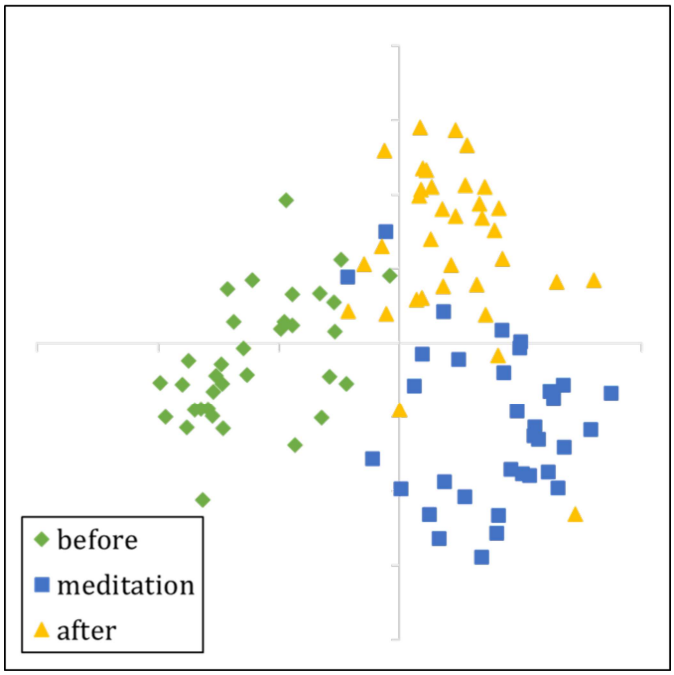


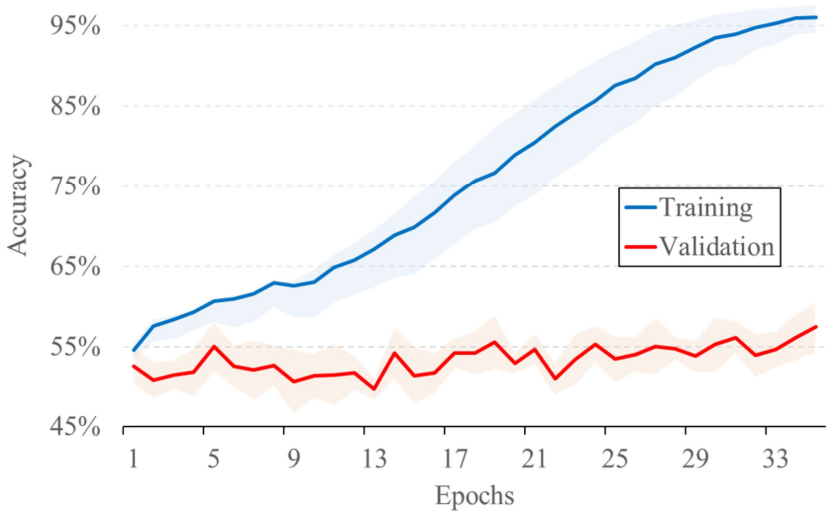
Figure 3.7: EEG data mapping based on top 2 linear discriminants.

segments before meditation (green). Furthermore, both classes use for training exhibit an overlap with the data distribution for rest segments after meditation (yellow). This supports our initial decision to exclude those segments from the ML dataset as they may exhibit overlapping effects from both rest and meditation. It should, however, be noted that these visualizations were made using the entire dataset. The ML models will need to generalize an effective mapping from only the test data, which is challenging because of the noisy features affecting the primary data distribution as demonstrated by Figure 3.6.

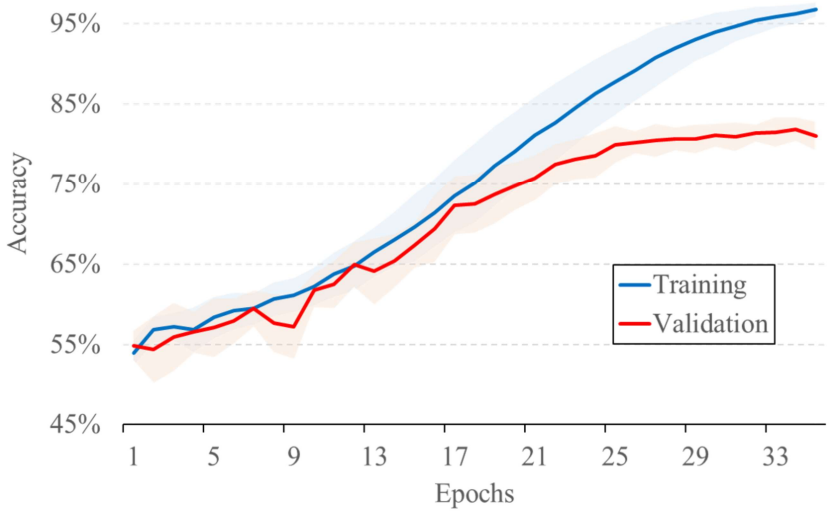
3.5 Experiments and Results

3.5.1 Dataset and Models

The final dataset for ML testing consisted of ∼9000 instances extracted from 4-second windows from the 5 minutes non-meditation baseline and the latter 5 minutes of guided meditation. For the model tests, we first tested ResNet-50 with locked and unlocked weights, and later followed up with a shallower model with 3 convolutional layers. Performance was validated using stratified 10-fold cross-validation.

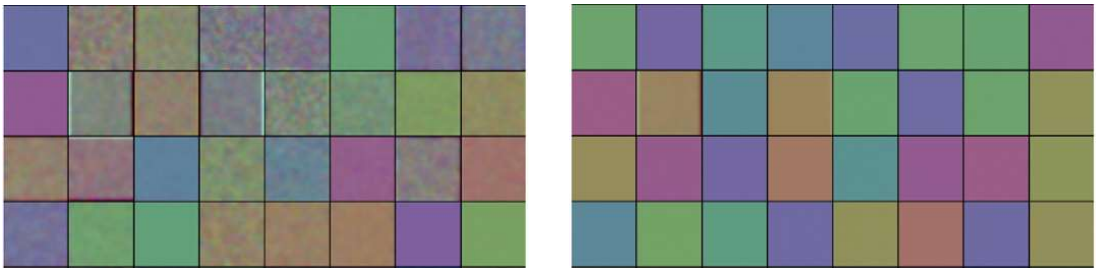


(a) ResNet-50 with pre-trained weights



(b) ResNet-50 with learned weights

Figure 3.8: Model training results for ResNet-50. Error ranges represent 1-unit standard deviation over 10-fold cross-validation.



(a) ImageNet pre-trained filters (b) Filters trained on EEG visualization

Figure 3.9: Input layer convolutional filter activations.

3.5.2 Deep Model Results

We first tested the dataset with ResNet-50 using locked pre-trained weights (from ImageNet) and attached a pair of dense layers on top for supervised training. The averaged classification results across all 10 folds shown in Figure 3.8a indicate that the model was overfitting. It shows a good progression for the training accuracy, but validation results stayed very low and close to random classification (∼50%). This could indicate that regular image CNN features are not compatible with the EEG visualizations, possibly due to their lack of prominent high frequency spatial features. To test this hypothesis we trained a second model with the same architecture but with all weights learned from scratch. The results for the second test, shown in Figure 3.8a, demonstrate that by training all model weights using the current dataset we are able to obtain significantly higher test accuracies easily reaching 75% and almost 90% for the optimized model. Based on these findings, we attempt to further optimize the model by analyzing the layer activations of the model.

3.5.3 Analysis of Deep Filter Activations

By analyzing filter activations of the convolution layers closest to the image input (See Figure 3.9), we observe that compared to the image-trained model, the one trained on the EEG visualizations focuses more on inter-channel gradients than on spatial textures. We noted how the maximal layer activations from the pretrained model favour richer spatial textures, while the EEG-trained model's activations prefer flat colors. This matches traditional methods of EEG analysis where there it is often more effective to compare the characteristics of different brainwave bands. In this case, those bands are represented by the different image channels. As such, it may not be necessary to use very deep models to learn complex spatial patterns. Rather, it may be more practical

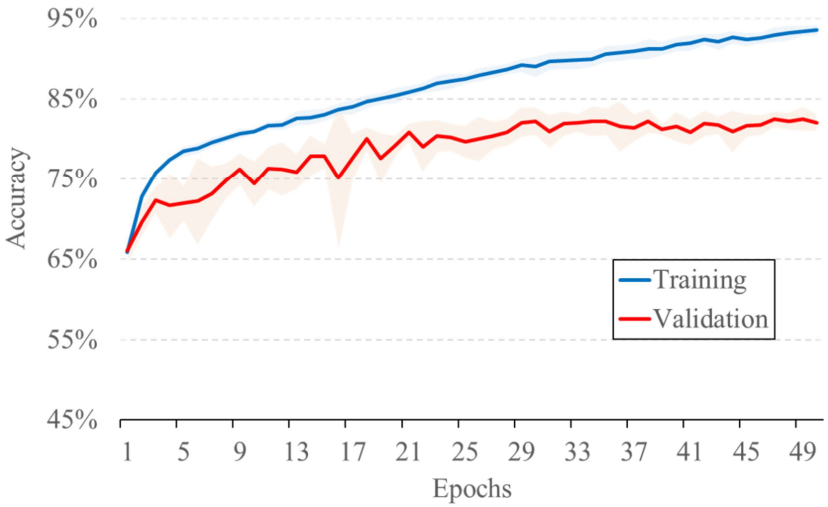
to add more filters to each layer in order to allow the model to learn more variations of inter-channel patterns. So, for the final test we built and trained a shallow networks and compared a narrow and a wide implementation of the network.

3.5.4 Shallow Model Results

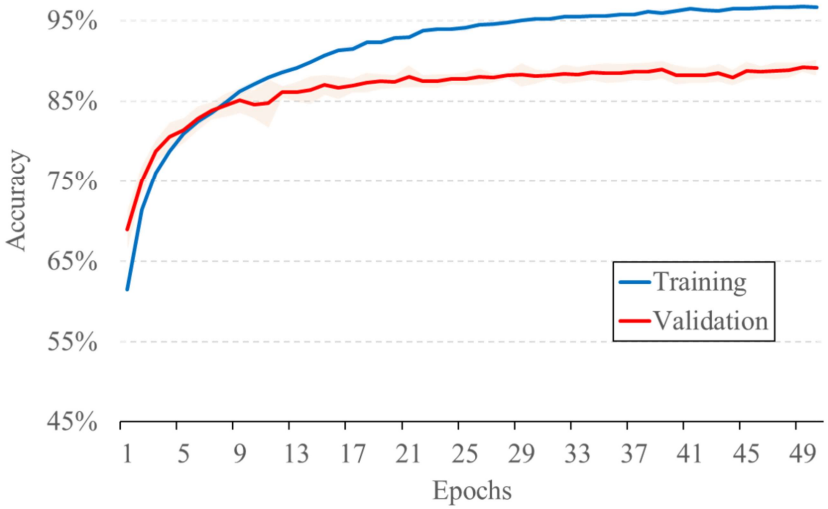
Based on the results from the activation analysis of the deep models, this learning problem may not require a very deep model. Instead, it may be more efficient to provide a sufficient number of filters to model the complex relationships between brainwave channels. In other words, our dataset may favor models that scale through width rather than the more typical depth. To test this hypothesis, we compare shallow models comprised of 3 convolutional blocks and 2 dense layers as shown in Figure 3.4. First, we tested a narrow shallow model with only [8, 16, 32] filters for each of the convolutional layers. To compensate for the reduction in parameters, we doubled the number of neural units in the dense layers to 256. The training results shown in Figure 3.10a indicate that while the model was more resistant to overfitting than either of the ResNet models, it was learning much more slowly and showed very high bias. So, while its final test performance was better than the ResNet models, the difference was not very large, reaching an average of roughly 83% accuracy across all 10 folds.

Thus, we also test a wide implementation with [64, 128, 256] filters in the convolutional layers and 128 units in each of the dense layers, similar to the ResNet models. The results in Figure 3.10b show that the wide model performed much better than the narrow model in both training and validation phases. Additionally, it showed lower variation in test performance across the training epochs and so was easier to train with early stopping. It also achieved significantly better test performance and convergence characteristics than both of the deep ResNet models, albeit over longer training epochs. In just a little over 30 epochs it reached >85% test accuracy, exceeding even the peak performance of the ResNet models, and eventually converged at almost >89% accuracy. Meanwhile, the best ResNet-50 model saw testing performance begin to degrade after 31 epochs. Figure 3.11 shows a comparison of the test performance of the best ResNet and shallow models, with the x-axis normalized based on epochs-to-convergence.

Based on these results, we can conclude that our optimized shallow models were more successful at avoiding overfitting than the deeper ResNet models. These results also confirm our initial observations that our approach favors horizontal scaling, favoring wide



(a) Narrow model with [8, 16, 32] filters



(b) Wide model with [64, 128, 256] filters

Figure 3.10: Model training results for shallow model. Error ranges represent 1-unit standard deviation over 10-fold cross-validation.

Table 3.1: Comparison Results

	ResNet (Pretrained)	**ResNet (Learned)**	**Shallow (Narrow)**	**Shallow (Wide)**
Mean	55.70%	81.37%	82.21%	88.91%
SD	1.23%	0.34%	0.29%	0.17%

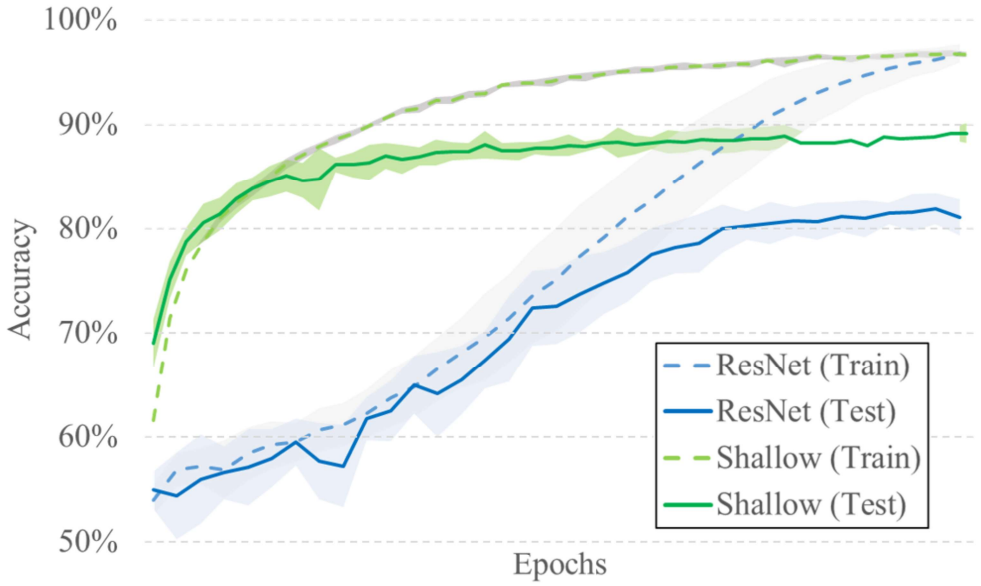


Figure 3.11: ResNet-50 model performance results compared to Shallow-Wide model. Error ranges represent 1-unit standard deviation over 10-fold cross-validation.

models with more filters rather than deep models with more layers. A final summary of test results across all the models is presented in Table 3.1.

3.5.5 Subjectivity Analysis

Aside from 10-fold cross-validation, we also tested leave-one-subject-out cross-validation, however the results were not very promising. Figure 3.12 shows the results from 16 random leave-one-subject-out tests. For each test, a model was trained with all data except from one subject. The model would then be tested on the data from the remaining

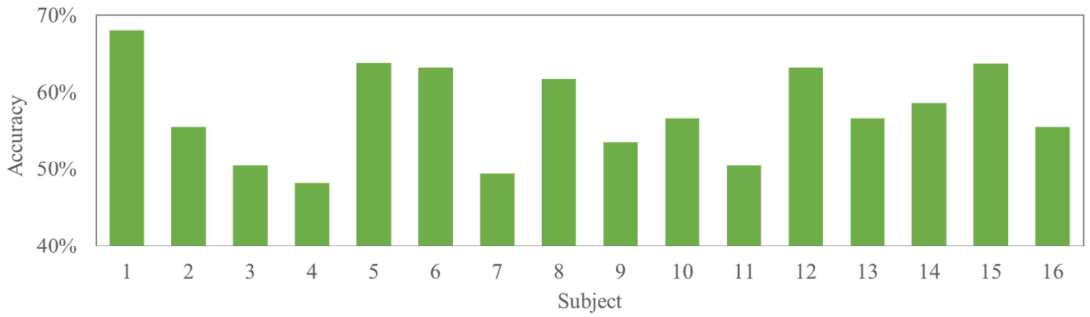


Figure 3.12: Subject-independent test results for 16 randomly selected subjects. Many of the models barely perform above random chance (50%) on this binary classification problem (rest vs. meditation).

subject. The weak performance results reveals that although our proposed model was effective at generalizing when given training data from all subjects, it was not very robust to the subjective differences between users. Indeed, at no point in the methodology was subject-independence specifically enforced. Additionally, it should also be noted that performance may have also been affected by labeling noise inherent to most human-centered datasets. In this case, it may be possible that not all samples gathered from the latter 5-minutes of the meditation phase may be representative of true meditation and this may have affected the learning of the model. In practice, it is difficult to guarantee and identify the onset of meditative states, even for experienced practitioners, though some guidelines have been proposed. Nevertheless, these could be good topics for future works.

3.6 Discussion

In this study, we built a small naturalistic dataset for mindfulness meditation and trained various deep model architectures to detect meditative states. Using a visual model of meditation EEG and an optimal shallow-wide CNN, we address the noise and subjectivity limitations of our naturalistic EEG dataset. By observing the activation behaviours of a deep off-the-shelf ResNet model, we designed an optimized shallow-wide architecture that allowed our model to scale through width rather than depth. The final model achieved an average accuracy of 88.91% in 10-fold cross-validation tests for distinguishing between rest and meditative segments. Considering that out of the 34 usable sessions, 14 actually contained one or more (but less than six) unusable data channels, this is a very promising result that corroborates findings by [64] regarding the noise-robustness of image-based EEG models. We also showed that it is possible to utilize data samples with lost segments by leveraging the resilience of CNNs to spatial noise. By enabling models to learn even from samples containing noise, future works can benefit from the reduced costs of acquiring large EEG datasets for training DL models. Moreover, the sliding STFT window method as well as all preprocessing steps used can be applied for real-time tracking, albeit with an initial delay corresponding to the length of the sliding window. This can eventually lead to generalized health trackers that can track a user's mental meditative states in the same way that most existing devices track steps. The proposed model is currently limited primarily by its subject-independent performance. In order to work towards a generalized deep model for EEG that is comparable to exist-

ing image models, our next efforts focus on finding ways to enforce subject generalization in deep EEG models.

Chapter 4

Building Subject-Generalized Emotion Models using Variational and Adversarial Regularizers

4.1 Introduction

As discussed in the previous study, the field of human-computer interactions (HCI) can benefit greatly from advancements in wearable EEG technologies by offering the ability to directly decode the mental and emotional needs of users through their brainwaves. In the previous study, we proposed an approach that successfully enabled the training of models using EEG data with noisy channels. However, results also revealed that while the model performed well on cross-validation tests, it still showed subject-independent performance, indicating that it was likely to over-fit to users in the training set. This is a challenging problem that cannot be solved simply through spatial or spectral processing methods. In fact, the true extent of this noise is difficult to measure since it is the result of intrinsic personal differences and makes up some of the most pervasive challenges across EEG-Emotion studies [23].

Machine learning, particularly deep learning, offers a solution to many of the signal noise problems by offering to bypass some of the manual EEG feature extraction steps through simply learning which features are relevant to the task by training to ignore noise sources in the data distributions. Therefore, one of the primary challenges for EEG-based deep learning methods is finding architectures that effectively guide the model to recognize and extract the most suitable EEG signal features while resisting the various sources of noise and variability inherent to the dataset. Domain adaptation

is a branch of modern machine learning methods that could hold the solution to this problem. Domain adaption models aim to bring optimized models to a given target domain by applying knowledge learned from another, possibly better understood source domain. They address the problem of contextual variance by learning to ignore features that are unrelated to the learning task and instead focusing on features that are useful across different domains. The growing interest in this field is driven by realizations that along with the great learning potential and flexibility of deep learning models comes the increased propensity to overfit to a single sample distribution (i.e., the training set) [26]. In the case of EEG-emotion studies, regardless of the form of the input features, of key importance is finding high-level features that are invariant across subjects. This is a basic requirement for building practical EEG-based systems for general use, and is where most existing systems tend to fall short.

In this work [76], we take effective concepts from domain adaptation and apply them towards domain generalization by combining adversarial training techniques with variational regularization. This is the first work to apply both methods simultaneously to an EEG dataset and serves as a good measure for how the combined methods can apply to multi-channel spectral-spatial datasets. The proposed architecture tackles the data constraints and subject variability simultaneously by employing variational autoencoder (VAE) components at the input-layer to provide regularized Gaussian-distributed feature embeddings, while using subject-adversarial training enforces subject-independence on the deeper features. The Gaussian embeddings prevent the model from collapsing into tight clusters due to the small sample size, while the subject-independent regularizer ensures that the primary variance of the embedding is free from subjective effects. Through comparative experiments, we show that our proposed method leads to embeddings that generalize better than comparable state-of-the-art methods. Meanwhile, through embedding space visualizations, we show how our model successfully learns a variational, subject-independent embeddings of EEG features related to emotions.

4.2 Related Works

While most domain adaptation techniques were originally intended for other signal processing applications such as image processing and sensor drift [77], a growing number of these techniques are gradually being applied successfully to EEG analysis. Roy et al. [62] provides a comprehensive survey of recent advancements in DL for EEG analysis,

while Lan et al. [78] provides a good benchmark of domain adaptation methods applied to EEG-emotion data. In this section, we discuss a number of recent methods for extracting domain-invariant features from EEG and narrow down which techniques can potentially lend themselves to domain generalization.

4.2.1 Domain Adaptation

Domain adaptation constitutes a subset of transfer-learning techniques, which attempt to directly adjust learned model parameters to better accommodate a target dataset. The underlying principle is that data from two domains collected under different conditions can be assumed to have different feature distributions. The general goal of adaptation is to find a feature transformation where the domains are indistinguishable but still retain some of the original characteristics of the data [77]. In the case of EEG, it is often desirable to remove factors stemming from subjective characteristics while maintaining the effects of the phenomenon being studied (e.g., emotion). The majority of techniques currently being explored in this field are either kernel methods or DL methods.

Kernel-Based Methods

Kernel methods are a class of algorithms that use kernel functions to project data onto a high-dimensional feature space where they are more likely to be linearly separable. These projections are typically calculated from inner product spaces, also known as Hilbert spaces, $\mathcal{H}$, which also encode data similarity metrics. For more efficient computation, $\mathcal{H}$ is paired with a corresponding *reproducing kernel*, $k(x, x')$, which when satisfying certain properties of linearity and continuity, defines a reproducing kernel Hilbert space (RKHS) [79]. This is the core of the *kernel trick*, which enables substitution of the kernel function, $k(x, x')$, with a more efficient feature mapping, $< \Phi(x), \Phi(x') >$, where Φ is the feature map. This enables the kernel projection to be more efficiently applied to machine learning classification methods such as support vector machines.

Among domain adaptation kernel methods, one powerful kernel-based approach is proposed by transfer component analysis (TCA) [80]. TCA implements domain adaptation in RKHS by optimizing kernel parameters to minimize the maximum mean discrepancy (MMD) metric between domains. MMD essentially measures the difference between the means of the probability distributions of source and target feature map-

pings $\Phi(X_S)$ and $\Phi(X_T)$. As the distance between means approaches zero, the RKHS becomes more universal [78]. To preserve the original properties of the data, variance is maintained by maximizing the trace of the covariance matrix of the projected samples. However, an important limitation of TCA is that adaptation can only be applied to binary domains since the MMD metric is a pairwise calculation only suited for domain transfer.

A more recent work that builds upon TCA is maximum independence domain adaptation (MIDA) [81]. Similar to TCA, MIDA is an unsupervised domain adaptation technique that uses RKHS projection and combines it with a domain independence criterion. In this case, domain independence is calculated using the Hilbert-Schmidt Independence Criterion (HSIC) [82], a non-parametric test statistic that is also effective as a kernel independence measure. MIDA works by first encoding the domain labels into "*domain features*" which are concatenated to the original features. Next, a nonlinear mapping is applied through a kernel matrix, $K_x = \Phi(X)^T\Phi(X)$, and the samples are projected to a subspace with dimension h using a learned projection matrix M. The final projection Z is equivalent to $W^T K_x$. Finally, domain adaptation is achieved by regularizing the kernel matrices K_x, K_d and projection matrix W by minimizing the HSIC. Additionally, a centering matrix adjusts the projected space to approximately zero mean. The main advantage of MIDA lies in the encoding of domain features into the projection, thus allowing the HSIC to work across multiple domains. A semi-supervised extension of MIDA (SMIDA) also allows for concurrent optimization of the mapping using target labels. Recent studies [78, 83] have shown that MIDA outperforms other comparable methods like TCA and MMD in EEG-emotion datasets.

Deep Learning Methods

In recent years, DL methods have steadily taken over the range of state-of-the-art of domain adaptation techniques [77]. In the context of EEG analysis, one of the first promising deep domain adaptation methods to be proposed was the Wasserstein generative adversarial network (WGAN) [84]. This method uses a pair of *generators*, neural networks trained to output a feature vector. A pretraining step initializes separate generators for the source and target domains. Both are set to output similarly sized vectors, therefore creating a shared feature space. Next, an adversarial training step adjusts the parameters of the generators such that the outputs more closely match each

other based on a Wasserstein distance. Thereafter, the combined outputs of the adapted source and target feature spaces are used as input features for the output classifier network. Experiments on two of the most popular EEG-Emotion datasets (i.e., SEED and DEAP), showed that their Wasserstein generative adversarial network domain adaptation (WGANDA) framework showed significantly better results in subject-independent tests than those of the standard KPCA, TCA, and transductive parameter transfer (TPT)[85] approaches. However, the need for training different networks for each new domain makes it difficult to scale the technique to multiple domains. A separate generator would be necessary for each subject in the training set, increasing both memory and processing demands.

In another study [86] that used a similar topography representation to our proposed work, domain adversarial networks (DANN) [8] were used to regularize intermediate feature embeddings for subject invariance. The proposed architecture featured separate feature embeddings for each hemisphere, as well as a combined embedding for the final whole brain. Each embedding was trained with a separate DANN, each of which was trained to recognize subject labels. By fine-tuning the embeddings adversarially to each target subject during testing, domain distribution differences could be minimized. As a result, the BiDANN model performed better than state-of-the-art approaches, including plain DANN and dynamic graph convolution networks [87], on the SEED dataset, and particularly in the subject-independent tests.

A more recent work [88] applied a similar hemispheric approach that also optimized for latent representation similarity between the source and target domains. This approach treated the different sides of the brain as independent sources of data, resulting in a bi-hemisphere architecture that better reflects hemispheric brain theories that attribute different roles in the formation of emotion to different sides of the brain [89]. In their proposed work, the architecture features source and target networks which are treated as separate models during training; however, they feature shared model weights. Shallow layer output features were adapted using marginal distribution adaptation through an adversarial domain predictor DANN. Meanwhile, the deep layers up until the final output features were adapted to the conditional distribution of the target label by combining the outputs of the source and target networks with association reinforcement. Results on SEED and DEAP show that their method is more effective than not only the baseline KPCA, TCA, TPT, but also recent neural network approaches

like DANN, MMD and associative domain adaptation (ADA)[90]. However, once again parameter fine-tuning simply adapts the distribution of the intermediate features to the target domain.

In many of these approaches, the model both identifies salient features from the source data and optimizes them to accommodate the target subject, in a binary domain transfer scheme. One key limitation of this scheme is that it requires retraining the model for each new subject. Systems designed for use by multiple users may require a model that shows good generalization across multiple subject domains. The base DANN method [8] is a notable exception since it is not necessarily limited to binary adversarial domains. In the following section, we discuss domain adaptation techniques that can be applied toward domain generalization.

4.2.2 Domain Generalization

Often with consumer products, it is necessary to have a system that is immediately usable by as many users as possible. This requires a domain agnostic or domain invariant model with reasonably good generalization performance. *Domain generalization* approaches achieve this by leveraging data from multiple sources to learn a universal representation that generalizes well across unseen target domains [91].

Domain Agnostic Methods

In addition to methods that explicitly use domain labels, general-purpose techniques that are agnostic to the source/target domains can also be used for generalization. For instance, in traditional EEG analysis, better generalization is often achieved by using ICA [59] to decompose the signal into its source components and subsequently selecting only the most relevant components. Another typical use-case of ICA is for the removal of irrelevant noise effects from electromyography (EMG) or electrooculography (EOG) (i.e., muscle or eye movements). However, the need for manual intervention limits its applicability in automatic machine learning methods.

Fortunately, linear decomposition methods like PCA, which include inherent feature selection, have shown some promise for EEG domain generalization. Particularly, kernel PCA (KPCA) [92], has been shown to be somewhat effective for cross-domain classification [83]. While it is not specifically designed for domain adaptation, it is often used in comparisons to these types of methods [78].

KPCA is an extension to PCA that applies the kernel trick $K(x, y) = \Phi x^T \Phi(y)$ to project data non-linearly onto a kernel space where it may be more linearly separable. This space is a RKHS projection, similar to the one employed in MIDA. PCA is applied within this kernel space by computing the covariance matrix and determining the "*principal components*" of the projected data by ranking eigenvectors by their corresponding eigenvalues. By selecting a subset of eigenvectors which contribute most of the variance of the dataset and using them as basis for the final projection, a more generalized projection can be achieved. An important limitation of PCA is that it is sensitive to outliers since they may have a strong influence on the variance of the dataset.

Another study [93] proposed the use of variational mode decomposition (VMD) [94], a technique that combines signal decomposition and component selection, and combines it with DL. VMD is a time-frequency decomposition approach that algorithmically detects and decomposes a signal into its principal "*modes*". This involves an iterative process where each mode is updated directly in Fourier domain using a narrow-band Wiener filter corresponding to the current estimate of the mode's center-frequency. Meanwhile, the center frequency is re-estimated based on the center-of-gravity of the mode's power spectrum. As a result, each intrinsic mode function (IMF) exhibits certain sparsity properties with regards to their bandwidths. That is, they have a limited bandwidth and are assumed to be compact around a center pulsation which is determined together with the decomposition. IMFs can serve the same purpose as principal components in PCA [61]; however, the former preserves time-frequency information by isolating the modes of the signal and can be used to generate a meaningful PSD. In Pandey and Seeja [93], the top three IMFs were selected and their PSD features were combined into feature vectors which were subsequently used to train a deep neural network (DNN). Through experiments on the DEAP dataset, they found that the VMD-based model performed better than comparable models trained on raw PSD or differential entropy features.

Deep Latent Variable Methods

Another possible approach to domain generalization is to use neural networks to discover latent features in the data. This is comparable to the domain agnostic models, however DNNs also tend to scale better with larger datasets.

In a recent study [95], they proposed the use of variational autoencoders (VAE) [96]

to extract features for classifying EEG motor imagery. VAEs are generative machine learning structures that learn an n-dimensional embedding of data constrained to a distribution. An encoder network transforms data to an embedding space while a decoder network reconstructs the original input from the embedding. An additional constraint is applied to the distribution of the embeddings. By combining these objectives, the encoder is able to generate a well-distributed set of features. Their work shows that Gaussian VAEs can be used to effectively extract salient features from 2D EEG spectrographs, resulting in a model that outperforms the state-of-the-art subject-independent results for the BCI Competition IV 2b dataset. They also showed empirically that the ideal EEG features naturally follow a normal distribution based on a Kolmogorov-Smirnov test. We adopt a similar approach with VAEs but with an additional subject-invariant regularizer.

Alternatively, we can take a more unsupervised approach by allowing the machine learner to discover domain-invariant features. In Chai et al. [97], they proposed the use of autoencoders to enforce a consistency constraint, combined with subspace alignment techniques, to minimize domain differences. Their proposed subspace alignment autoencoder (SAAE) uses MMD for domain-invariant regularization and showed competitive performance on the SEED dataset, outperforming baselines like TCA. A limitation of the work is that the subspace alignment requires data from all domains to be loaded onto memory simultaneously. This limits its scalability to multiple domains, so only binary domain transfer was explored.

Multi-Domain Adaptation

To build generalized models, it may be necessary to consider marginal distributions across multiple domains. Fortunately, certain approaches used for domain adaptation exhibit characteristics that could be applied for domain generalization [83]. Notably, domain agnostic techniques and those that support multi-domain regularization can be used to enforce domain generalization. Among these, KPCA, VMD, and MIDA provide good comparative performance while offering varied approaches to generalization. Methods discussed in this section are summarized in Table 4.1.

Table 4.1: Summary of Related EEG Domain Adaptation and Generalization Methods.

Method	**Algorithm Type**	**Domain Invariance**	**Adaptation Type**	**Distribution**
TCA [80]	Kernel-based	MMD	Binary domain	Centered distribution
MIDA [81]	Kernel-based	HSIC	Multi-domain	Centered distribution
WGAN [84]	Deep Learning	Wasserstein distance	Binary domain	None
KPCA [92]	Kernel-based	Principal components	None	Centered distribution
VMD [93]	Signal Decomposition	IMF decomposition	None	Approx. Gaussian
VAE [95]	Deep Learning	None	None	Gaussian
SAAE [97]	Deep Learning	MMD	Binary domain	None
DANN [98]	Deep Learning	Adversarial training	Multi-domain	None
BiDANN [86]	Deep Learning	Adversarial training	Binary domain	None
BiVDANN [Ours]	Deep Learning	Adversarial training	Multi-domain	Gaussian

4.2.3 Contributions of the Work

In summary, while recent advancements in DL have brought about marked improvements in EEG emotion recognition models, complex methods with many parameters still suffer due to the relatively small labeled datasets and the high subject variability of EEG. Topographic representations provide a practical middle ground that allows machine learning algorithms to discover high-level spatial-temporal context-specific features, while also reducing data requirements by utilizing manually extracted, but well-researched, effective low-level features (i.e., PSD). Meanwhile, domain adaptation has proven to be quite effective for handling subject variability; however, most methods require fine tuning of the model for each target subject. Contrarily, label-agnostic normalization methods do not cope well with small datasets with imbalanced feature distributions. For applications where subject-specific data are unavailable or model retraining is not computationally feasible, a flexible general-purpose model is needed.

The proposed work offers a flexible domain-generalized deep EEG emotion model that builds upon previous works by learning a bi-hemispheric variational embedding optimized for subject invariance. To work around the small-data constraints of EEG-emotion datasets, instead of building an end-to-end model, we use topographic inputs similar to Bashivan et al. [8] while also adopting the bi-hemispheric feature split of Li et al. [86]. This also enables better flexibility for combining trained models with other datasets in the future. To optimize for generalized feature embeddings that are not overfit to the training distributions, VAEs with Gaussian regularization similar to Dai et al. [95] are applied to the hemispheric topography features. These variational embeddings are further refined by combining them with domain adversarial training [98] across subjects to learn subject-independent feature embeddings. This novel combination of regularization approaches increases the robustness of the proposed model to the effects of subject variability compared to purely distribution-driven methods like VMD and KPCA. Additionally, unlike domain adaptation methods like MIDA, the addition of VAE regularization should increase the resistance of our model to the overfitting effects of small datasets with non-Gaussian feature distributions. We validate the performance of the models by focusing on subject-independent classification performance when combined with a DL classifier, similar to the approach used in Pandey and Seeja [93]. Furthermore, we perform qualitative analysis of the embedding spaces using manifold learning algorithms to demonstrate the differences in the feature distributions.

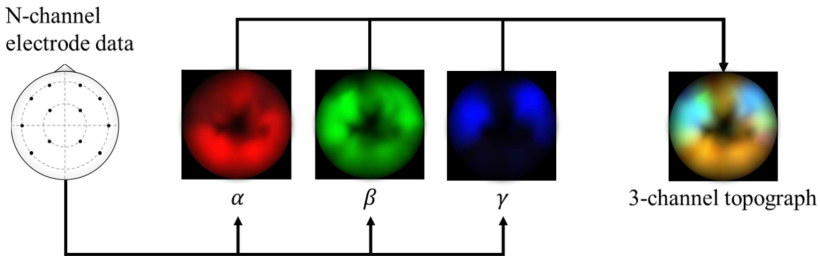


Figure 4.1: Alpha, beta, and gamma power are extracted from the electrode data, plotted onto separate topographic maps, and are combined into a 3-channel topograph used as input for the proposed model.

4.3 Methodology

4.3.1 Input Data

The framework in this study uses spectral topography data to maximize dataset inter-compatibility and to improve robustness to localized electrode noise. The pre-processing and topography generation were performed using MNE-Python [99] with the PSD calculated using the multitaper method [100] and bilinear spatial interpolation. For the target channel bands, we refer to findings from previous works [101, 86, 102, 103] that have repeatedly shown that relatively high frequency bands typically lead to significantly improved emotion recognition performance. Therefore, we generate three 64 $\times$ 64 px topographs based on the following brainwave bands: alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–50 Hz). Subsequently, these topographs are combined into a single 3-channel image with pixel values scaled between 0 and 255, as shown in Figure 4.1. This encoding has two main advantages: (1) it acts as a form of normalization to scale the values within a standard range, and (2) it allows the data to be visualized as an RGB image. This may be useful later for visualizing the DL model's activations. These data are passed as input to a custom deep CNN.

4.3.2 Deep Learning Architecture

An overview of the architecture of the proposed model is shown in Figure 4.2. Similar to Li et al. [86], we adopt a bi-lateral architecture to emulate the lateralized brain theories of emotion [89], which are frequently used in neuroscience and neuroimaging studies. In the context of machine learning, asymmetry features have regularly been found to improve results considerably when compared to just using electrodes separately [104, 86]. This is enforced at the input level by cropping the topographs into left and right hemispheres

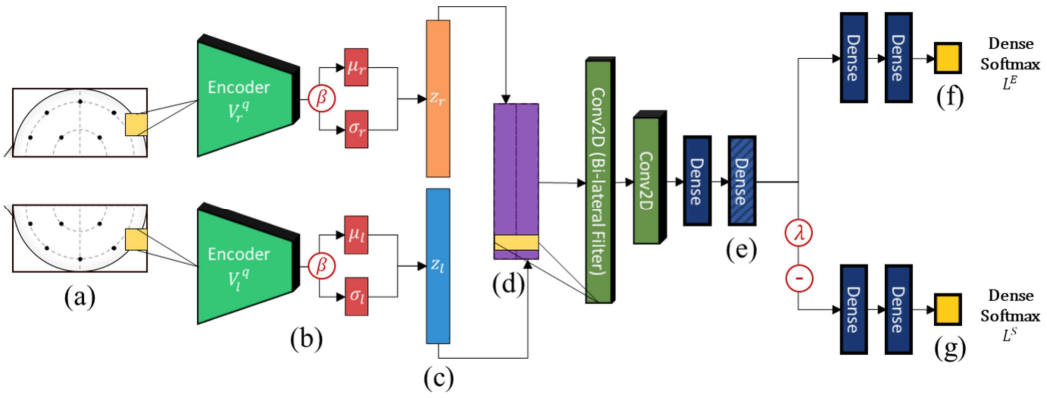


Figure 4.2: The high-level architecture diagram of the proposed model. From left-to-right, (**a**) EEG topograph inputs, (**b**) convolutional variational encoders, (**c**) lateralized embeddings, (**d**) bi-lateral feature concatenation and lateral convolution, (**e**) subject-independent dense embeddings, (**f**) emotion classifier, and (**g**) adversarial subject classifier.

each with their independent VAE weights. This ensures that feature embeddings from either hemisphere are properly regularized to a Gaussian distribution. The outputs of the lateral VAEs with β-scaled regularization (beta-VAE [105]) are then combined and filtered using a bi-lateral convolution to extract lateralized features. Thereafter, we impose generalization across domains through user adversarial regularization on the bi-lateral feature embedding by applying domain adversarial training. In this case, different subjects are treated as different domains. An adversarial classifier attempts to distinguish between different users, while applying a negative gradient to any upstream layers to train them to resist these attempts. Effectively, the model learns a feature embedding that is indiscriminate between users while jointly optimizing the features to distinguish the target emotion labels. Additionally, the Gaussian penalties of the VAEs prevent the model from converging to an unexpected out-of-bounds distribution that may also satisfy the adversarial and target classifiers.

4.3.3 Regularizing for Normally Distributed Features

The proposed model uses a pair of VAEs to generate the initial embeddings for the input EEG topographs. Aside from serving as a feature embedding regularizer, they have also been shown to be effective for generating new data and even a viable alternative to generative adversarial networks (GAN) [106]. Some studies [107, 108] have also shown how VAE-based training can improve classification performance. It is hypothesized that VAE regularization leads to better-disentangled representations, which could explain

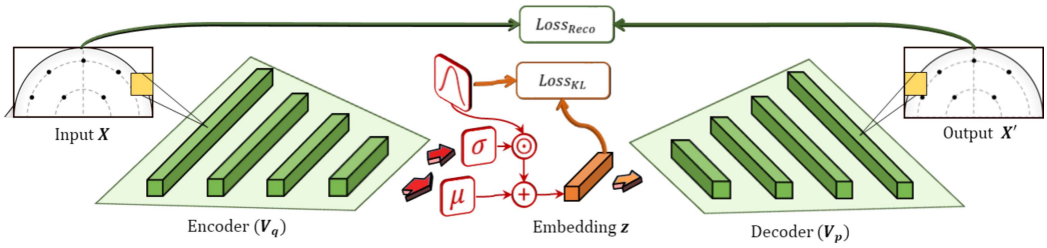


Figure 4.3: Variational autoencoders are composed of an encoder–decoder pair. The encoder outputs mean and s.d. components, which are used to sample the latent encoding, $\boldsymbol{z}$, stochastically from a target distribution. It combines losses for the reconstruction of the input, $\boldsymbol{x}$, and the KL-loss between the encoding, $\boldsymbol{z}$, and the target distribution (i.e., Gaussian).

the improved classification performance. In the context of EEG signals, Aznan et al. [109] demonstrated how VAE can be effective for simulating EEG samples which when used for pre-training models can improve the cross-subject accuracy. This is further supported by another recent work [95] that showed how CNN-VAEs can be effective for classifying motion imagery from EEG spectrograms.

In the proposed work, separate beta-VAEs are used for each hemisphere to allow the model to encode bilateral asymmetry features. We opt for the standard mean squared error (MSE) for the reconstruction loss in the VAEs despite common criticisms to MSE reconstructions when used in image classifiers. This is because the blurry reconstruction common to MSE-based VAEs is a non-issue for the EEG topograph inputs, which are inherently spatially interpolated.

Variational Autoencoders

The VAE is a machine learning structure based on the autoencoder and includes both encoder and decoder components, as shown in Figure 4.3. The encoder learns to generate an embedding that is constrained in both number of dimensions and type of distribution. For instance, the embedding space can be configured to be an n-dimensional vector of normally-distributed values. The decoder then takes the embedding and attempts to generate an accurate reconstruction of the input. By combining these objectives with gradient optimization methods, the model learns an embedding that approximates latent factors inherent to the data, similar to matrix factorization. Additionally, the embeddings are constrained to a Gaussian distribution that allows the model to prioritize factors in the data that are well-distributed.

Originally formalized in Kingma and Welling [96], a VAE is a machine learning

structure that is trained based on learned approximate inference and works by constraining the generated representations to some continuous distribution (e.g., Gaussian). VAEs provide a computationally efficient method to optimize deep latent-variable models jointly with inference models by solving the intractability problem of parameterizing the posterior distribution, $p_\theta(\mathbf{x}|\mathbf{z})$, for the latent variables, $\mathbf{z}$, given a marginal likelihood, $p_\theta(\mathbf{x})$ [110]. Similar to other variational methods, VAEs optimize for the *evidence lower bound* (ELBO) but also include a KL-divergence term, redefining the ELBO loss as,

$$\mathcal{L}_{\theta,\phi}(\mathbf{x}) = \log p_\theta(\boldsymbol{x}) - D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}))||p_\theta(\boldsymbol{z}|\boldsymbol{x}). \tag{4.3.1}$$

The KL term indicates the KL divergence of the approximate posterior, $q_\phi(\boldsymbol{z}|\boldsymbol{x})$, from the true posterior, $p_\theta(\boldsymbol{z}|\boldsymbol{x})$. Due to the non-negative property of the KL term, the ELBO loss acts as a lower bound on the log-likelihood of the data. Here, the VAE employs a parametric *inference model*, $q_\phi(\boldsymbol{z}|\boldsymbol{x})$, also known as the *encoder*, which is parameterized by θ to approximate the intractable posterior, $p_\theta(\boldsymbol{z}|\boldsymbol{x})$. By approximating the posterior with a generative model, $p_\theta(\boldsymbol{x}, \boldsymbol{z})$, the ELBO loss can be applied to enable the joint optimization of parameters θ and ϕ, using stochastic gradient descent, as defined in the AEVB algorithm [96].

Thus, if we assume that $p_\theta(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \mathbf{0}, \boldsymbol{I})$ and that the true posterior for our latent encoding approximates to a Gaussian distribution, then the approximate posterior becomes a multivariate Gaussian defined as,

$$\log q_\phi(\boldsymbol{z}|\boldsymbol{x}_i) = \log \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2 \boldsymbol{I}), \tag{4.3.2}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$ are the mean and s.d. outputs of the encoder for input $\boldsymbol{x}_i$, parameterized by ϕ. To calculate the gradients through stochastic sampling, the *reparameterization trick* is applied by sampling from the posterior, $\boldsymbol{z}_{il} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x}_i)$, with $\boldsymbol{z}_{il} = \boldsymbol{\mu}_i + \boldsymbol{\sigma}_i \bigodot \boldsymbol{\epsilon}_l$, where $\boldsymbol{\epsilon}_l \sim \mathcal{N}(0, \boldsymbol{I})$ and $\bigodot$ denotes an element-wise product. Thus, the resulting loss equation for the Gaussian implementation of VAE can be defined as,

$$\mathcal{L}_{VAE}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}_i) \simeq \frac{1}{2} \sum_{j=1}^{J} \left(1 + \log((\sigma_{ij})^2) - (\mu_{ij})^2 - \sigma_{ij}^2 + \frac{1}{L} \sum_{l=1}^{L} \log p_\theta(\boldsymbol{x}_i|\boldsymbol{z}_{il}) \right), \tag{4.3.3}$$

where J is the feature dimension of the latent embedding, $\boldsymbol{z}$, and L denotes the number of random samples taken (typically, $L = 1$).

Our proposed model implements beta-VAE [105], an extension to the base VAE that

introduces a controlled parameter, β, that allows tuning of the learning constraints of the model. Specifically, it augments the KL-divergence term in Equation (4.3.1) into $\beta D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z}))$. This enables stronger disentanglement between latent factors by increasing the regularization loss and also allows the adjustment of the regularization strength during training. Recent investigations have found that scheduling β to increase linearly during training results in improved model convergence [111]; therefore, we adopt a similar training procedure.

Optimizing these objectives enforces improved sampling of latent features despite biases in the training distribution. Non–Gaussian distributions are often attributed to insufficient data or problems in the data sampling methodology. In effect, they do not generalize well. By constraining the reconstruction objective to finding Gaussian embeddings, the model discovers features that are better suited for generalization. That is, the model can ignore outliers or badly sampled factors and biases inherent to the training data to some extent since it models the distribution of clean data [112]. A model trained on these losses learns a stochastic mapping from the $\boldsymbol{x}$-space defined by the available training dataset, to a latent $\boldsymbol{z}$-space embedding regularized to a distribution that more closely matches what we would expect from a more comprehensive dataset (i.e., Gaussian). When used as a pretraining method, such regularization has been found to improve the classification performance by learning factors of variation that improve data log likelihood [111]. This is particularly useful in our case since the high-dimensional nature of our dataset combined with the relatively small sample size increases the possibility of overfitting to the training distribution. Without proper regularization, there is a high probability that standard deep learning models will develop a strong bias to the training set resulting in poor performance on unseen data.

Next, the hemispheric VAE outputs are vertically concatenated and processed with a bi-lateral convolutional filter to extract lateral brain activity features as shown in Figure 4.4. This is inspired by the lateral brain theories of emotion [89] and is an approach that has proven effective in previous works [8, 86]. However, unlike in previous works, the VAE regularization should force the bi-lateral feature embeddings to be well-distributed, emulating distributions expected from large comprehensive datasets. In addition, to further curtail the effects from subject variability, we apply subject-independent regularization using domain adversarial training.

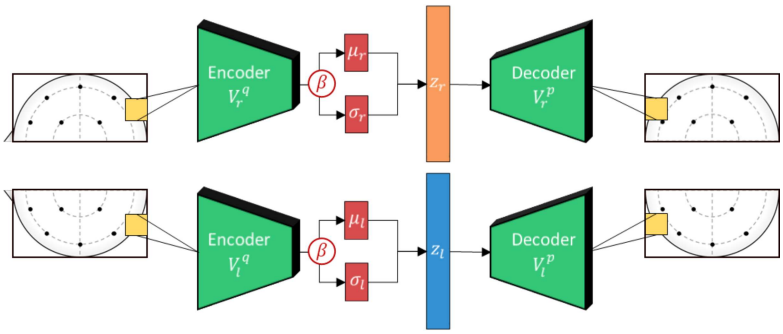


Figure 4.4: The lateralized beta-VAE implementation used in this work. Each hemisphere is composed of an encoder and decoder pair with the encoder generating a stochastically sampled embedding, z, based on outputs, μ and σ. Separate weights are used to encode the features for each of the hemispheres to adhere to neuroscience theories of lateralization.

4.3.4 Optimizing for Subject-Independence

To achieve better generalization across different subjects, we regularize the bi-lateral filter features with DANN training. DANN is a domain alignment approach usually presented in domain adaptation studies as an alternative to GANs [25]. The underlying theory is that effective domain transfer can be achieved by finding features that are indiscriminate between the training (source) and test (target) domains [98]. While it is primarily used for adapting between binary source/target domains, in this work we apply it together with the features extracted from the VAE embeddings to enforce subject-invariance. A similar approach for speaker recognition is found in [113] where it showed promising results for generalizing across different audio datasets. In this case, we treat each subject as an independent domain that is assumed to be sampled from a different distribution and apply the regularization starting from a feature embedding layer shared by the emotion classifier and the adversarial classifier, as shown in Figure 4.5.

Domain Adversarial Neural Networks

One of the earliest works formalizing DANN adaptation [98] tested the technique on optical character recognition (OCR) models by adapting pretrained models to different OCR datasets. As a general-purpose architecture, it is defined primarily by its joint optimization of a *prediction loss* and a *domain loss*. In a simplified case featuring a single hidden layer DANN, the prediction and domain losses can be defined by the

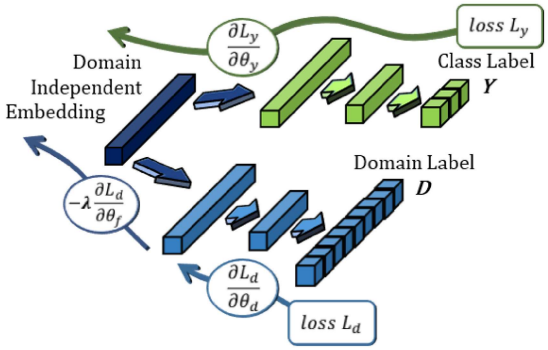


Figure 4.5: The domain adversarial network (DANN) implementation used in this work. The feature embedding layer is shared by the emotion classifier and the adversarial subject domain classifier. Each classifier is trained to classify for its respective task; however, the gradient is reversed and scaled with parameter λ, in effect training upstream layers for subject-independence.

following:

$$\begin{aligned} \mathcal{L}_y^i(\theta_f, \theta_y) &= \mathcal{L}_y(G_y(G_f(\boldsymbol{x}_i, \theta_f); \theta_y), y_i), \\ \mathcal{L}_y^i(\theta_f, \theta_d) &= \mathcal{L}_d(G_d(G_f(\boldsymbol{x}_i, \theta_f); \theta_d), d_i) \end{aligned} \tag{4.3.4}$$

Here, G_f is defined as the neural network feature extractor, G_y is the part of the neural network responsible for label classification, and G_d is the neural network domain classifier. The parameters, $\theta_f, \theta_y, \theta_d$, refer to the trainable parameters of their corresponding neural networks. Training the DANN consists of finding the saddle points for parameters $\hat{\theta}_f, \hat{\theta}_y, \hat{\theta}_d$, such that,

$$\begin{aligned} \hat{\theta}_f, \hat{\theta}_y &= \arg\min_{\theta_f, \theta_y} E(\theta_f, \theta_y, \hat{\theta}_d), \\ \hat{\theta}_d &= \arg\max_{\theta_d} E(\hat{\theta}_f, \hat{\theta}_y, \theta_d) \end{aligned} \tag{4.3.5}$$

As is apparent from the parametric saddle point in Equation (4.3.5), the objective for $\hat{\theta}_d$ runs adversarial to that of the other classifiers. Instead of minimizing the error, it is necessary to maximize it; therefore, the combined objective for the feature generator can be defined simply as,

$$\mathcal{L}_{DANN}(\theta_f, \theta_y, \theta_d) = \mathcal{L}_y^i(\theta_f, \theta_y) - \lambda \left(\mathcal{L}_d^i(\theta_f, \theta_d) \right), \tag{4.3.6}$$

where λ is a scaling factor that controls the strength of the reversed adversarial gradient. Typically, $\lambda = 1$; however, this hyper-parameter can be tuned if either the target classifier cannot reach convergence or the feature embedding is not sufficiently domain-invariant. Empirical testing has also shown that introducing an unsupervised

pre-training step as described in Algorithm 1 leads to better convergence.

Algorithm 1: Bi-lateral VAE Pre-training

Data:
X : EEG topograph training set
Models:
$\mathcal{V}_l$,$\mathcal{V}_r$: Model weights for left and right βVAEs
Parameters:
β : Variational regularization hyper-parameter

Initialize ($\mathcal{V}_l$,$\mathcal{V}_r$) with small random variables
Initialize $\beta = \beta_{target}$
while *target epochs not reached and early-stopping patience not exceeded* **do**
    Generate random training mini-batch X^b from X)
    Partition X^b into left and right bi-hemispheres (X_l^b,X_r^b)
    Train $\mathcal{V}_l$,$\mathcal{V}_r$ with (X_l^b,X_r^b) based on Equation (4.3.3)
    Update losses history with $\mathcal{L}_{VAE}$
    Check if the change in loss is below the threshold and update *patience* variable as necessary
Save model weights $\mathcal{V}_l$,$\mathcal{V}_r$

4.3.5 Training Phases

Unsupervised VAE Pretraining

Before training the full model, it may be beneficial to pretrain the VAEs in an unsupervised manner with the training set, as shown in Algorithm 1. This step reduces the risk of exploding error gradients from the large losses during the initial phases of jointly training the VAEs and the classifiers. We also found that this practice increases the convergence rate of the subsequent supervised and semi-supervised training phases by providing a reusable starting point for supervised training.

Supervised EEG-BiVDANN Training

The proposed model is trained to optimize for generating feature embeddings that are indiscriminate across subjects in the training set. This is hypothesized to also approximate toward indiscriminate features for unseen test sets from the same general sampling population. To optimize for both the variational loss and the adversarial loss, our model uses interleaved adversarial training, VAE reconstruction, and classifier optimization, as described in Algorithm 2. A jointly optimized model was also tested, however it was found that the interleaved training allows for finer control and tuning of the individ-

ual regularizers. Properly tuning the regularization strengths is essential for preventing the independent losses from overwhelming each other, particularly with the target label classifier.

Algorithm 2: Bi-lateral VAE Pretraining

Data:
X : EEG topograph training data
Y : Emotion training labels
Z : Subject context labels
Models:
$\mathcal{V}_l$,$\mathcal{V}_r$: Model weights for left and right βVAEs
$\mathcal{H}$: Model weights for bi-lateral feature extractor
$\mathcal{C}$: Model weights for emotion classifier
$\mathcal{D}$: Model weights for subject discriminator
Parameters:
β : Variational regularization hyper-parameter
λ : Adversarial gradient hyper-parameter

Load pretrained weights for ($\mathcal{V}_l$,$\mathcal{V}_r$)
Initialize $\beta = 0$
Initialize $\lambda = 0$
while *target epochs not reached and early-stopping patience not exceeded* **do**
    Generate random training mini-batch M^X,M^Y,M^Z from (X,Y,Z)
    Partition M^X into left and right bi-hemispheres (M_l^X,M_r^X)
    Fine-tune ($\mathcal{V}_l$,$\mathcal{V}_r$) with (M_l^X,M_r^X) based on Equation (4.3.3)
    Generate hemispheric outputs (H_l,H_r) with encoders ($\mathcal{V}_l^q$,$\mathcal{V}_r^q$)
    Generate bilateral embedding H' by convolving (H_l,H_r) with $\mathcal{H}$
    Train $\mathcal{D}$ with H' and context labels M^Z base on Equation (4.3.6)
    Propagate negative λ-scaled gradients to train $\mathcal{H}$ and ($\mathcal{V}_l$,$\mathcal{V}_r$)
    Train $\mathcal{C}$ with H' and training labels M^Y based on Equation (4.3.7)
    Propagate raw gradients to train $\mathcal{H}$ and ($\mathcal{V}_l$,$\mathcal{V}_r$)
    Update losses history with $\mathcal{L}_y$
    Check if the change in loss is below the threshold, and update *patience* variable as necessary
    if $\beta < \beta_{target}$ **then**
        Slightly increase β
    if $\lambda < \lambda_{target}$ **then**
        Slightly increase λ

EEG-BiVDANN Model Objective

To build a subject-generalized model of EEG emotion, our proposed BiVDANN adopts the joint optimization of three losses: (1) VAE loss (Equation (4.3.3)), (2) subject-independent DANN loss (Equation (4.3.6)), and (3) emotion classifier loss $\mathcal{L}_y(\phi_v, \theta_f, \theta_y)$.

Formally, the combined loss can be defined as,

$$\mathcal{L}_{BiVDANN}(\phi_v, \theta_v, \theta_f, \theta_y, \theta_d) = \\ \mathcal{L}^l_{\beta VAE}(\theta^1_v, \phi^l_v) + \mathcal{L}^r_{\beta VAE}(\theta^r_v, \phi^r_v) + \mathcal{L}_y(\phi_v, \theta_f, \theta_y) - \lambda\left(\mathcal{L}_d(\phi_v, \theta_f, \theta_d)\right) \quad (4.3.7)$$

The adversarial minimization and maximization objectives are thus defined as follows:

$$\min_{\phi_v, \theta_v, \theta_f, \theta_y} \max_{\theta_d} \mathcal{L}_{BiVDANN}(\phi_v, \theta_v, \theta_f, \theta_y, \theta_d) \quad (4.3.8)$$

Internally, Equation (4.3.8) encapsulates the following min-max optimization steps:

$$\begin{aligned} \hat{\phi}_v, \hat{\theta}_v &= \arg\min_{\phi_v, \theta_v} \mathcal{L}_{BiVDANN}(\phi_v, \theta_v, \hat{\theta}_f, \hat{\theta_y}, \hat{\theta}_d), \\ \hat{\phi}_v, \hat{\theta}_v, \hat{\theta}_f, \hat{\theta}_y &= \arg\min_{\phi_v, \theta_v, \theta_f, \theta_y} \mathcal{L}_{BiVDANN}(\phi_v, \theta_v, \theta_f, \theta_y, \hat{\theta}_d), \\ \hat{\theta}_d &= \arg\max_{\theta_d} \mathcal{L}_{BiVDANN}(\hat{\phi}_v, \hat{\theta}_v, \hat{\theta}_f, \hat{\theta_y}, \theta_d), \end{aligned} \quad (4.3.9)$$

where the saddle point is defined by symbols $\hat{\phi}_v, \hat{\theta}_v, \hat{\theta}_f, \hat{\theta}_y$, and $\hat{\theta}_d$. When these symbols are found on the right-hand side of Equation (4.3.9), it indicates that their respective parameters are not optimized by the current operation. Tuning of the models relies primarily on selecting hyper-parameter values that balance the main optimization objectives and may vary based on the dataset.

4.4 Datasets and Models

4.4.1 Datasets

In this study, we use two of the largest and most widely studied EEG-emotion datasets: DEAP [114] and SEED [104]. Both datasets feature multimodal data; however, we constrain this study to only the EEG portion of the data. Similarly, both also include various types of annotations; however, we focus on the ones most relevant to emotion.

For DEAP, the dataset comprises data from 32 subjects, each with a single continuous data gathering session divided into 40 trails that last about 1 min each. Emotion was elicited using music videos and labels were collected with a questionnaire including a self-assessment manekin (SAM) for arousal and valence (9-point scale). The EEG device used for measurement was a 32-channel Biosemi headset capable of recording at a 512 Hz data rate; however, the final cleaned data were downsampled to 128 Hz. Since we focus only on the emotion classification aspect, we train the models to classify a binary

valence where valence annotations > 5 are *positive*, and those < 5 are *negative*. The selection of the valence threshold in other works tends to be rather arbitrary; however, here we opt to use 5 as it is the median value of the SAM questionnaire.

For the SEED dataset, it comprises data from 15 subjects, with each having participated in 3 separate sessions divided into 15 trials, which were approximately 4 min long. Emotions were elicited using emotional film clips pre-labeled as positive, neutral, or negative. For the EEG device, they used a 62-channel ESI Neuroscan system capable of recording data at 1000 Hz but with the final stored data downsampled to 200 Hz.

4.4.2 Data Preparation

To generate the representative topographs from the raw EEG data, we apply a sliding window method. For this, we consider that labels in both datasets are applied across entire trials. Thus, to increase the probability that most segments included the target labeled emotions we employ a 15 s sliding window with a 50% overlap. This is roughly based on the findings of a previous study [115] that found $\sim$12 s to be optimal for valence classification. To extract the time-frequency features and generate the topographic representations we use MNE-Python [72], an open-source package for MEG and EEG analysis. Specifically, we apply discrete prolate spheroidal sequences (DPSS) tapers to extract the PSD spectrum [116]. This method avoids the spectral leakage that is common in methods that use the standard Welch periodogram. A 64×64 px topograph is generated for each brainwave band of interest (alpha, beta, and gamma), resulting in three-channel matrix whose values are then scaled between 0 and 255 to generate the input topograph image.

4.4.3 Model Implementation

The bulk of the machine learning components in this work are implemented in Python, specifically using the Keras machine learning API [117]. All models were optimized using the Adam optimizer [118] since it could best handle the highly variable gradients produced by the joint optimization. For the purpose of discussion, we will divide the proposed BiVDANN into its main components: bi-lateral feature extractor, emotion classifier, and adversarial subject classifier. A description of each is shown in Table 4.2. Model layer depth was optimized to maximize the visual field of the convolution operators. Meanwhile, dense layer and embedding sizes were determined through grid

Table 4.2: BiVDANN Implementation Details.

Component	**Input (Dims)**	**Components (Dims)**	**Output (Dims)**
Unilateral VAE	*Topograph*: (32,64)	*Encoder*: 3 × [Conv2D(3 × 3) + BatchNorm] *VAE Embedding*: [Dense(512) + Mean: Dense(64) + SD: Dense(64) + z: Dense(64)] *Decoder*: 3 × [Conv2DTranspose(3 × 3) + BatchNorm]	*Unilateral Embedding*: (64)
Bilateral Feature Extractor	*Unilateral Embeddings*: 2 × (64)	[Concat + 2 × [Conv2D(2 × 2) + BatchNorm] + 2 × [Dense(128) + BatchNorm]]	*Bilateral Embedding*: (128)
Emotion Classifier	*Bilateral Embedding*: (128)	2 × [Dense(256) + BatchNorm] + 1 × [Dense(64) + BatchNorm] + Dense(Softmax)	*Emotion prediction*: (# of classes)
Adversarial Subject Classifier	*Bilateral Embedding*: (128)	GradientReversal + 2 × [Dense(256) + BatchNorm] + 1 × [Dense(64) + BatchNorm] + Dense(Softmax)	*Subject prediction*: (# of subjects)

search optimization based on model loss from 5-fold cross validation. Other hyperparameters such as learning rates and regularization weights were determined through Bayesian optimization also based on model loss from 5-fold cross validation.

4.5 Experiments and Results

4.5.1 Training Hyperparameters

The proposed model jointly optimizes the multiple objectives specified in Equation (4.3.7); however, properly controlling conflicting gradients requires fine-tuning of various regularization weights, particularly β and λ. Empirical tests show that linearly increasing both β and λ during the early stages of training allows the model to converge to a distribution with the desired characteristics. Specifically, we grow β from zero to full magnitude during the first 100 epochs, while λ had a slightly faster increase reaching max in the first 50 epochs. Testing showed that this configuration prevents the model from starving the target emotion learning objective. Slowly growing beta also prevents the gradient from exploding due to changes in KL-loss while the model is also exploring the main class objective during the initial learning phase. Similarly, slowly increasing lambda allows the model to first learn a strong user discriminator that will later serve as a powerful regularizer when the negative adversarial gradient reaches a critical point where it matches or overpowers the class gradient. For the training epochs, a batch size of 256 samples was used to allow the adversarial batches to contain a good distribution of subject samples, and empirical testing showed that 300 epochs allowed all models to converge to an optimum state with the help of early stopping. We used 15% of the training data ($\sim$1200 samples) for validation and set the early stopping objective based on the emotion class validation accuracy.

Figure 4.6 shows the learning progression of the model on DEAP. Specifically, it shows the testing losses and accuracies (*Loss/Acc*) on the training and validation sets (*Train/Val*) for the emotion classifier and the subject discriminator (*Class/Disc*) over the epochs. Here, the model is learning with 2 emotion classes (*Positive/Negative*) and 32 subject classes. The objective of the model is to first learn a good subject discriminator, and then to optimize the emotion classifier with features that minimize the performance of the subject discriminator. The use of adversarial objectives pushes the emotion classifier toward subject-invariant features.

The early training phase is characterized by a rapid decrease in all losses (red lines),

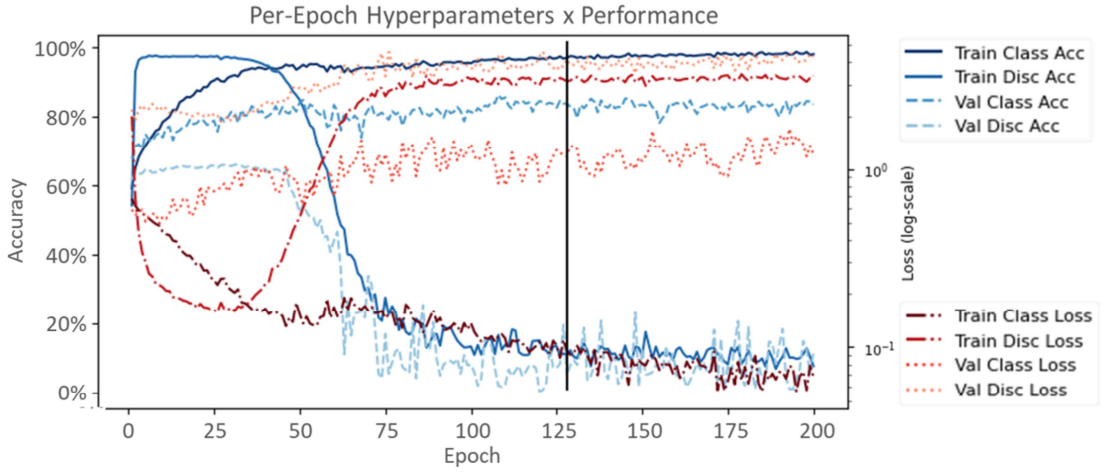


Figure 4.6: Sample training progression for DEAP data. The graphs indicate the training/validation accuracy/losses for the emotion classifier/subject discriminator over training epochs. The black line indicates the point used for early stopping.

particularly the discriminator loss. This demonstrates the strong influence of subject variability on the dataset since the subject discriminator learns much faster than the emotion classifier despite having more subject classes (32 in the case of DEAP). If allowed to continue, there would be an eventual increase in the classification losses due to subjective overfitting. However, as λ reaches a critical point at around epoch 40, the adversarial gradients begin to tune the embedding for subject invariability. Eventually, this converges to a representation that is indiscriminate across subjects but is highly discriminative for emotion classes.

4.5.2 Comparison Results

For the comparison experiments all the featured methods used a similar *Emotion Classifier* component indicated in Table 4.2. This allows the comparison to be more even; thus, we can focus on comparing the efficacy of the different feature embeddings. Specifically, the following techniques are used for comparison:

- CNN—simple 2D CNN featuring 3 × [Conv2D(3 × 3) + BatchNorm] layers with ∼200,000 trainable parameters
- ResNet-50 [1]—popular DL network for image classification with ∼23 million trainable parameters
- KPCA [119]—nonlinear dimensionality reduction technique

- VMD [94]—signal decomposition technique used for EEG analysis in Pandey and Seeja [93]
- MIDA [81]—semi-supervised domain adaptation technique for mapping data to domain-invariant subspace

In this experiment, we wish to compare the generalization ability of the models. Therefore, we apply leave-one-subject-out cross-validation for all the quantitative experiments. That is, the entire dataset is divided such that all samples from a single subject are excluded from training a model. Subsequently, all samples from that subject are used to measure the test performance of the trained model. This is performed for each subject, and we use the mean accuracy and standard deviation to evaluate the models.

We wish to compare our proposed approach to the standard baseline methods, as well as to the other comparable domain generalization techniques. For the standard baselines, we use a standard shallow CNN to represent DL models with lower data requirements, and ResNet-50 [1] to assess the performance of more complex DNNs. These are both single-objective models that aim only to minimize the error on the training set. As such, they can be used as an indication of the inclination of the datasets towards overfitting to the training distribution. We can expect most classic DNNs applied to the same data to exhibit similar limitations.

For the generalization approaches, we combine feature extraction procedures that optimize for subject generalization. Since there are currently few state-of-the-art generalization techniques applied to EEG that experience improvements without including samples from the test domain, we use label-agnostic and multi-label modeling techniques for comparison. The algorithms compared include KPCA [119] to represent nonlinear label-agnostic dimensionality reduction approaches, VMD [94] for the EEG signal decomposition approach, and MIDA [81] for multi-domain adaptation. For VMD, we apply feature extraction on the raw EEG similar to [93], although we apply the same 15 s window used in the other methods. The results are shown in Tables 4.3 and 4.4.

Label-agnostic approaches such as KPCA and VMD operate solely based on the distribution of the data they are modeling. For KPCA, a nonlinear mapping is generated based on the principal components of the data and how much each contributes to the overall variance. Thus, they are suitable for identifying strong latent factors. However, this also makes them vulnerable to noisy data and outliers. Conversely, VMD is a sig-

Table 4.3: SEED Comparison Results (3-class)

	CNN	**ResNet**	**KPCA-DNN**	**VMD-DNN**	**MIDA-DNN**	**BiVDANN**
Mean	49.78%	52.31%	51.66%	58.45%	60.14%	63.28%
SD	5.09%	4.51%	6.67%	8.17%	4.51%	6.45%

Table 4.4: DEAP Comparison Results (2-class)

	CNN	**ResNet**	**KPCA-DNN**	**VMD-DNN**	**MIDA-DNN**	**BiVDANN**
Mean	51.64%	52.14%	55.31%	57.78%	60.97%	63.52%
SD	5.20%	7.32%	9.14%	6.75%	5.57%	5.21%

nal decomposition method that is designed to be resistant to noise effects, particularly those in nonlinear and non-stationary signals. It decomposes a signal into a finite number of IMFs which are comparable to KPCA's principal components. Moreover, the sifting process generally leads to approximately Gaussian distributed modes which can minimize the effect of outlier noise. These two algorithms can indicate how techniques driven only by data distribution perform on EEG datasets. They can also indicate the strong impact of the signal noise and outliers on the datasets. Moreover, near the end of the section, we also include model ablation tests to investigate the performance of basic VAEs and DANNs on these datasets.

Table 4.3 shows the performance on the SEED dataset. Here, the models were classifying three different emotion classes (positive, negative, neutral);thus, the random point lies at $\sim$33%. All models could classify significantly better than random; however, our proposed BiVDANN achieved the best overall subject-independent performance. The MIDA-DNN model also achieved high relative accuracy, displaying the advantage of enforcing subject-invariance for generalized models. In contrast, the shallow CNN got the lowest result likely due to overfitting. This is also reflected by ResNet-50, which only exhibits marginally higher performance than the shallow CNN despite having a much higher number of parameters.

Next, Table 4.4 shows the results on the DEAP dataset. Despite being only a binary classification problem, all the models had greater difficulty with this task compared to the 3-class SEED dataset case. Once again, our BiVDANN achieved the best overall accuracy, followed by MIDA. Similarly, the CNN and ResNet-50 models fared the worst, this time showing more similar performances. However, this time, their performance

was not significantly better ($p > 0.05$) than that of a random classifier. This indicates that without proper controls, models trained on the DEAP dataset may have a higher probability of leaning toward overfitting due to some external factor. Similar results were found in a number of studies [11, 23], so there is something inherent to the DEAP dataset that limits the performance of ML models. Since our method already accounts for subjective difference, the problem must lie in some other characteristic of the data. It is even possible that the labels themselves were noisy. More study is needed to confirm the true source of the issue.

In contrast, the multiple sessions in the SEED dataset may be reducing the overfitting effect of these environmental variances since each subject's training data is composed of combined data distributions across different sessions. However, more study is required to confirm this hypothesis.

4.5.3 Qualitative Analysis of Embeddings

Next, we wish to evaluate the effects of the characteristics of the different embeddings generated by the different methods on the model performance. To visualize the embeddings, we apply t-Distributed Stochastic Neighbor Embedding (t-SNE) [120] and multidimensional scaling (MDS) [121]. t-SNE is a nonlinear dimensionality reduction method that focuses on preserving small pairwise distances or local similarities between neighbors. Conversely, MDS is a technique for plotting data in lower dimensions while preserving a global distance metric (e.g., Euclidean distance). It works by using pairwise dissimilarities to reconstruct a mapping that preserves the triangle distances between data points. t-SNE is effective at emphasizing data clustering, while MDS is more effective at showing the global patterns of the data distribution. Here, we use the 2D maps to qualitatively assess how the data points and classes are clustered in the embedding spaces for the different techniques. The t-SNE visualizations should easily expose any clustering behaviors in the embeddings. Meanwhile, the MDS visualizations can illustrate the overall distribution of the embeddings, thereby exposing any unbalanced distributions. In the case of KPCA, it already serves as a data visualization algorithm and will provide a baseline reference for the natural distribution of the dataset.

To generate the visualizations, we train the respective algorithms with a random 70% batch from the datasets. As for the inputs to the visualization algorithms, in the case of the deep models, we use the *embedding layer outputs* of the models, or in other words,

the data embeddings positioned right before the final set of fully-connected classification layers, as shown in the right side of Figure 4.2. For VMD and MIDA, we use the raw features generated by the algorithms. Finally, for KPCA, we visualize the embeddings using the first two principal components.

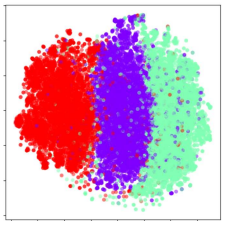
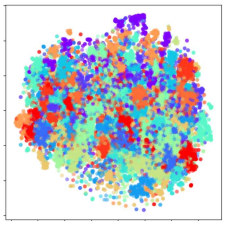
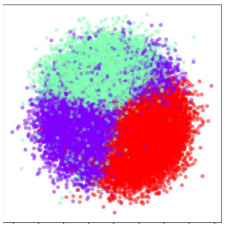
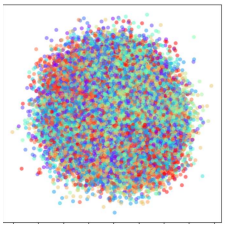
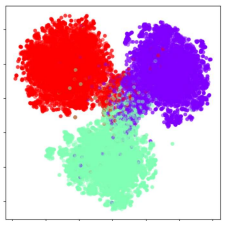
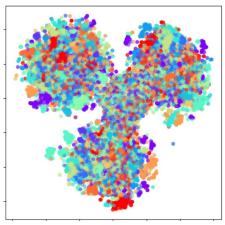
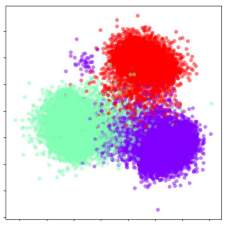
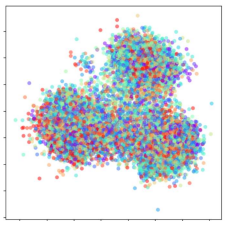
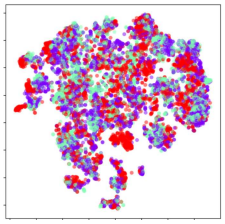
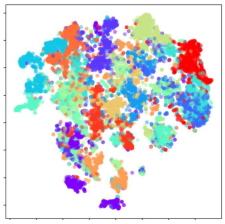
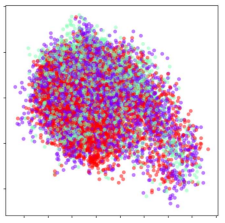
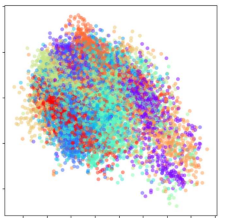
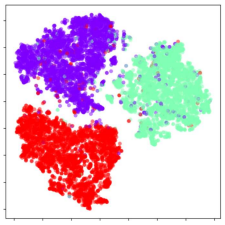
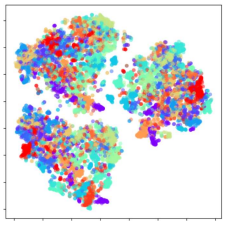
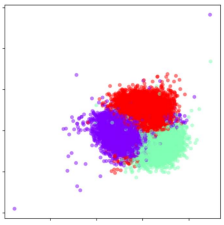
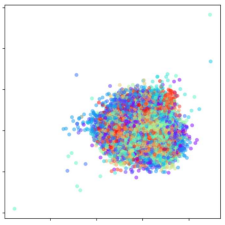
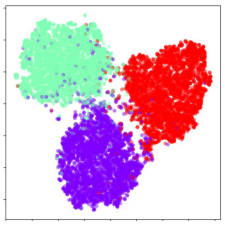
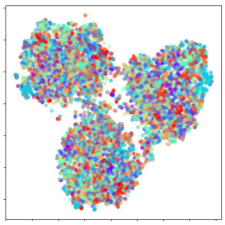
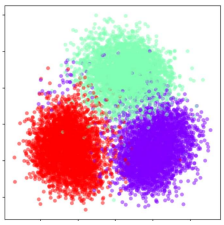
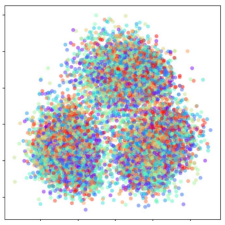
SEED Embeddings

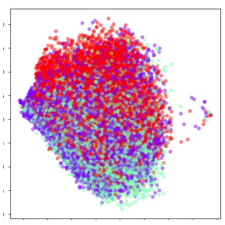
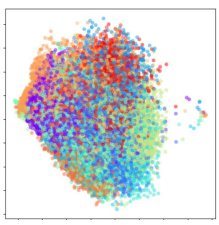
From the embeddings of the SEED dataset shown in Table 4.5, we observe that all the DL models can generate features that are clearly optimized for emotion separation. However, for the CNN and ResNet-50 models, we also observe some subject-based clusters leaking into the distributions. This indicates the presence of strong subject-dependent effects in the dataset and could indicate that features found by these models are also somewhat subject-dependent. Judging by their weak classification performance in the subject-independent tests, this inherent dependence may significantly limit their generalization performance. Meanwhile, for KPCA, we also observe some class-based separation of clusters, although subject clusters are much more dominant here than in the other models. This indicates that the variances of the features themselves are strongly influenced by the different subjects. This typically leads to poor generalization performance, as indicated in Table 4.3. VMD seems to be an exception to this since it shows heavy subjective clustering effects despite exhibiting better performance than the other baseline models. This could be due to its relatively normal distribution of features as shown in the MDS visualization. Meanwhile, for MIDA, the HSIC-based regularization exhibits the best mixing of the subject clusters, however this comes at the expense of the class clusters which show many incorrectly mapped data points. Finally, for the BiVDANN embeddings, we see strong class separation and minimal subjective influence in the feature distributions. The model is also able to generate a much more Gaussian feature distribution that emulates a more comprehensive dataset while prioritizing class separation.

DEAP Embeddings

Next, we analyze the embeddings for the DEAP dataset shown in Table 4.6. Again, we observe a similar pattern in the case of the DL models, where the embeddings show clear emotion class separation. However, there is also much stronger subject-dependent clustering compared to the results from SEED. This is most evident for the CNN model,

Table 4.5: SEED embedding visualizations for the tested models. Different colors correspond to their respective emotion class (**negative**, **neutral**, **positive**) or unique subject. The axis values represent the arbitrary 2D feature mapping performed by the dimension reduction algorithms.

Model	**t-SNE**		**MDS**	
	Classes	Subjects	Classes	Subjects
CNN				
ResNet-50				
VMD				
MIDA				
BiVDANN				

KPCA	
Classes	Subjects
	

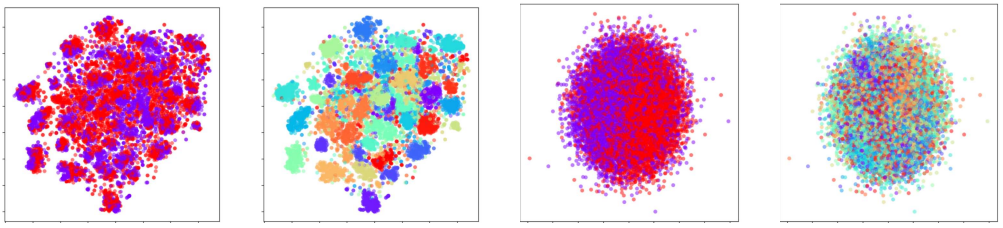
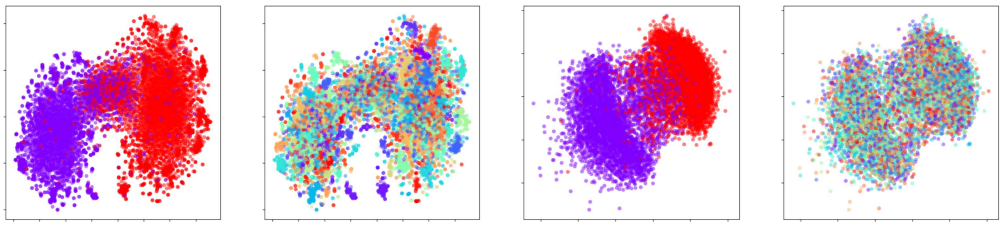
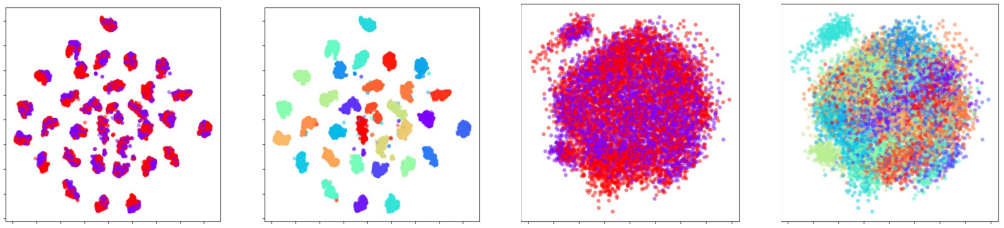
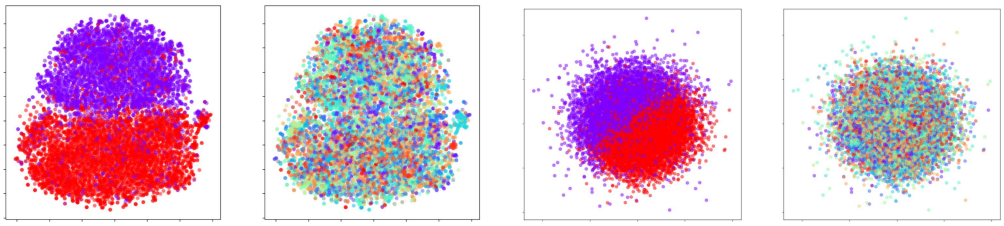
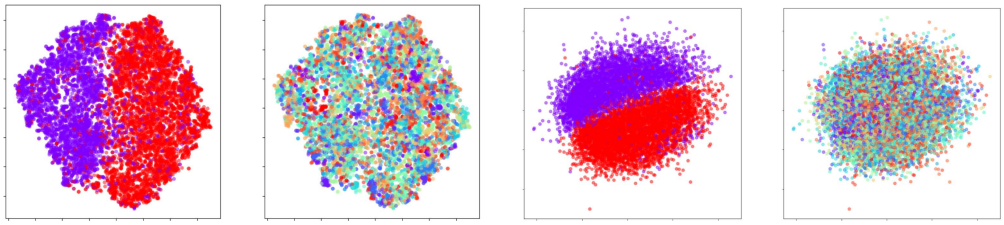
where subject clusters are even more dominant than the class clusters. This could explain the poor performance of the CNN model, which was only slight better than a random classifier at just 51.64% binary accuracy. ResNet-50 was better able to avoid subject-dependent effects; however, it also generated separate and lopsided clusters for the negative and positive classes. In this case, most of the testing samples are seen between the two clusters, leading to poor classification performance. The results for KPCA are once again similar to SEED, where subject-based clusters dominate the feature distribution. Therefore, it is unsurprising that its classification performance was similarly poor. Meanwhile, the BiVDANN was once again able to generate a feature distribution that separates the emotion classes while minimizing overfitting to subject-based variances.

General Observations

Based on the visualization results, we made several observations. First, we see that while class-based clustering is generally favorable, the generalization performance is also strongly influenced by other effects, such as subjective clustering. It may be better to minimize subjective clustering, particularly for datasets where distributions naturally lean towards subjective variance, as shown by the KPCA tests. Second, even in cases where subjective clustering is avoided, other unlabeled variances can also contribute to distribution imbalances to reduce generalization performance, such as in the ResNet-50 embeddings. Conversely, the VMD results demonstrate that well-distributed embeddings, even when strongly influenced by subjective variances, can still produce well-performing models. In summary, the best results seem to come from embeddings that favor class-based clustering and avoid subjective variances, while maintaining a balanced (i.e., Normal) data distribution.

With the qualitative analysis, we showed that our proposed model successfully achieved its main objectives of generating an effective generalized feature embedding for EEG emotion classification. Unlike label-agnostic approaches, such as KPCA, it can leverage subject label data to avoid the effects of subjective variance. Additionally, unlike single-objective methods like unregularized DL techniques, it generates a more contiguous distribution, which seems to better accommodate unseen data samples. This is best demonstrated in Figure 4.6, where both CNN and ResNet-50 generate good class-based clusters, but also have distributions with a high tendency of test sam-

Table 4.6: DEAP embedding visualizations for the tested models. Different colors correspond to their respective emotion class (**negative**, **positive**) or unique subject. The axis values represent the arbitrary 2D feature mapping performed by the dimension reduction algorithms.

Model	t-SNE		MDS	
	Classes	**Subjects**	**Classes**	**Subjects**
CNN				
ResNet-50				
VMD				
MIDA				
BiVDANN				

KPCA	
Classes	Subjects
	

ples landing between or outside of the known clusters. This could explain the poor classification performance and showcases the unpredictability introduced when treating emotion as discrete classes. Intuitively, negative and positive emotions should not be clearly separable classes; rather, they are different ends of the same spectrum. In effect, we expect that their distributions should overlap possibly at the point where neutral emotions should be located. By generating models that violate this smooth continuity, we may end up with overfitting representations that perform poorly on new samples. This is generally avoided by employing large comprehensive datasets. However, in cases where such a dataset is not available, such as in EEG-emotion analysis, regularization constraints, such as in our proposed BiVDANN, can provide a good alternative. In the next section, we test the contributions of the different components of BiVDANN with model ablation tests.

4.5.4 Model Ablation Study

For the model ablation study, we are interested in learning the contribution of the different components of the system to the overall model performance. The BiVDANN model features the following main components: bilateral inputs, variational regularizer, domain-independent regularizer, and emotion classifier. To test each of these components independently, the following models are trained and tested:

- BiCNN—includes bilateral inputs, but no variational regularization or domain-independent regularization.
- BiVAE—includes bilateral inputs, includes variational regularization, but no domain-independent regularization.
- BiDANN—includes bilateral inputs, includes domain-independent regularization, but no variational regularization.
- L-VDANN—only left-hemisphere inputs, includes both variational regularization and domain-independent regularization.
- R-VDANN—only right-hemisphere inputs, includes both variational regularization and domain-independent regularization.

The subject-independent test results for the ablation tests are summarized in Tables 4.7 and 4.8. The tests show that without the regularizers, the performance of

Table 4.7: SEED Ablation Results (3-class)

	BiCNN	**BiVAE**	**BiDANN**	**L-VDANN**	**R-VDANN**	**BiVDANN**
Mean	52.02%	53.41%	60.71%	57.49%	58.54%	63.28%
SD	5.29%	3.80%	5.33%	6.11%	5.66%	6.45%

Table 4.8: DEAP Ablation Results (2-class)

	BiCNN	**BiVAE**	**BiDANN**	**L-VDANN**	**R-VDANN**	**BiVDANN**
Mean	52.15%	58.08%	61.59%	60.69%	62.10%	63.52%
SD	6.42%	6.32%	5.09%	4.11%	5.45%	5.21%

the BiCNN model is equivalent to that of ResNet-50 shown in Tables 4.3 and 4.4. As such, we can expect other complex unregularized deep learning models to show similarly overfit performance. Next, the BiVAE model shows improved results with the aid of variational regularization. By jointly optimizing for emotion classes and a normal distribution, it achieves better performance than the VMD model, but worse than the MIDA model. The BiDANN model shows mostly equivalent performance to the MIDA, indicating that the domain-independent regularization is roughly equivalent to MIDA. Compared to BiVAE, BiDANN also shows comparable though slightly weaker performance. For the unilateral input tests, both models perform better than most other ablation models except for the BiDANN. This showcases the generalization ability of models that utilize the DANN regularization. Additionally, the R-VDANN model shows better performance than the L-VDANN model, though neither performs better than the bi-lateral BiVDANN. These results generally support the bilateral theory of emotion that posit that different hemispheres may be responsible for different types of emotions [122]. Although, the slight differences between the unilateral input models may also lend some credence to the right hemisphere hypothesis [123] that hypothesizes that the right hemisphere is primarily responsible for emotional perception, however its effects here are not statistically significant.

4.6 Discussion

Among generalized models that do not require additional fine-tuning, BiVDANN showed superior performance, beating comparable state-of-the-art techniques including KPCA, VMD and MIDA, even when these were supplemented with DNN classifiers. It consis-

tently achieved the highest subject-independent accuracy on both the SEED and DEAP datasets and showed promising performance even when constrained to only right hemisphere features. The qualitative embedding analysis also revealed clues regarding the advantages of the proposed model. Unlike label-agnostic approaches, such as VMD and KPCA, BiVDANN directly addresses the primary source of signal noise, which is subject variability. This results in features that are less prone to overfitting to certain subjects. Meanwhile, compared to purely context-focused approaches, such as DANN and MIDA, the proposed model can discover a well-distributed Gaussian feature space even when given a relatively small training set for the large number of model parameters. The VAE pretraining step also allows the model to learn salient features even when emotion labels are not yet available. As more EEG data become publicly available, it will be interesting to see at what point databases can achieve generalized distributions by virtue of their sample sizes. At that point, Gaussian regularization may no longer be necessary and regularization can be geared more toward other sources of variability.

Ablation testing allowed us to assess how each component of the model contributed to the improved performance. Among the ablated models, subject adversarial training (BiDANN) showed the best overall improvement. This supports the initial hypothesis that subject variability is a primary limiting factor for EEG-emotion models. Meanwhile, the VAE regularized model seemed to allow unseen samples to better fit within the primary distribution as opposed to the periphery, which may be overfit to a single emotion label. With non-regularized distributions, there is an increased chance that unseen users form their own feature clusters regardless of the emotion label. Another interesting finding was that a right-hemisphere model outperformed the left-hemisphere model, lending itself to the right-hemisphere hypothesis of emotion. Nevertheless, neither model could outperform the bi-lateral model; thus, lateralization may still play an important role in detecting human emotion.

Overall, it appears that the proposed model achieves its goal of providing a good generalized representation without the need for retraining. However, this comes with a trade-off in overall peak performance which was well below that of fine-tuned models adjusted for the target subject's data [84, 86, 88]. Among these works, some [84] have claimed over 90% accuracy under certain conditions; however, it is unclear what affects this and how much manual tuning is required to get these results consistently for each new subject. It is also uncertain if this is a reasonable expectation across datasets

when considering the inherent unpredictability of human emotions and the dataset labels themselves. Investigating this could be a challenge for future works and may be essential to proving if machine learning models are actually detecting emotion or if there is some other neurological phenomenon guiding the models. In any case, our approach provides a good starting point for general-use EEG-based systems while reserving the possibility for additional fine tuning when the necessary data requirements for new subjects are met.

Chapter 5

5. Summary and Discussion

5.1 Summary

In this study, we highlighted some of the common types of signal noise inherent to naturalistic ECG and EEG data, and proposed a number of optimizations for training DL models on these noisy physiological datasets. The primary objective of the approaches was to improve the robustness of the DL models to naturalistic data and noise-sensitive sensors data. The proposed solutions not only demonstrated how to efficiently train DL models on small, noisy psycho-physiological datasets, but also offered solutions for building generalized models that can be extended to other signal-based studies with similar noise characteristics. This not only increases the reliability of the models, but also reduces the costs of acquiring the necessary training data. Ultimately, this involved conducting the three main studies outlined in the following sections.

In the first study (ECG-Stress), we introduced a pipeline for building deep models of work stress using temporal-spectral visualizations of heart rate variability. Specifically, we used a short-time, sliding window implementation of the Lomb-Scargle transform to produce HR spectrograms. By extracting these spectrograms from ECG data obtained from subjects naturalistic work activities, and training DL models on stress level annotations, we built robust models of multi-level stress. Experiments on deep and shallow CNN architectures found that the proposed method could produce models that outperformed baseline models trained with gold-standard HRV features traditionally used in stress analysis. Additionally, it was discovered that relatively shallow models could be optimized to perform at a comparable level to more advanced deep architectures like ResNet-50. Based on these results, we learned that shallow convolutional architectures were generally sufficient for learning from the relatively simple single-channel heart rate

data even when transformed into a 2-dimensional spectral-temporal form. Therefore, in the next studies we attempted to apply a similar approach to multi-channel EEG data.

In the second study (EEG-Meditation), we built a small naturalistic dataset for mindfulness meditation and trained various DL architectures to detect meditative states. Once again a spectral approach was adopted to generate an intermediate representation that was robust to common sources of noise in the dataset. In this case, it was noticed that EEG samples would often feature a few electrodes that suffered from extreme noise, typically this was from loose contact with the scalp. Rather than dispose entire samples that had these relatively minor issues, a mask was applied to these channels and spatially-interpolated EEG topomaps were generated from the data. This enabled encoding of the spatial distribution of brainwave power while also making the representations robust to missing or noisy EEG channels. In a scheme similar to the previous ECG-Stress study, the topomaps were used to train deep and shallow DL models. At first, ResNet-50 models were trained to test the efficacy of 2D ML models on this form of data. After confirming that DL could indeed learn to predict meditation from the data, we analyzed the activations maps of the input layers of the network. It was discovered that the model favored inter-channel differences rather than the high frequency edge-like features usually found in image models. Based on these findings, we designed an optimized shallow-wide architecture which achieved an average accuracy of 88.91% for detecting meditative segments. Considering that out of the 34 usable sessions, 14 actually contained at least one but not more than 5 missing data channels, this result was promising. The study was successfully able to show that using the proposed approach, meditation models could be built from noisy data samples. Where the model failed was in subject-independent generalization. Efforts to solve the generalization problem eventually culminated in the final study featured in this dissertation.

In the third and final study, we proposed BiVDANN, a DL architecture designed specifically to learn a subject-generalized model from EEG data. By using the same spectral topography data in the previous proposed method, the model immediately benefited from the reduced the effects of electrode noise. In addition to this, the architecture also segmented the topomaps into left and right hemispheres to encourage the model to learn lateralized features, which are often cited in neuroscience studies as effective indicators of emotion. However, the core of the model was comprised of beta-VAEs tuned to discover Gaussian-distributed features, followed by a subject-adversarial

network to enforced discovery of subject-invariant features. We tested this approach using two of the most popular EEG-emotion datasets, SEED and DEAP. The results from the embedding comparisons showed that BiVDANN was more effective than comparable state-of-the-art EEG feature extractors, including KPCA, VMD and MIDA, at subject-independent tests. Quantitative and qualitative analysis of the outputs of the embedding layers confirmed that the technique was successfully able to learn a Gaussian-distributed subject-invariant embedding. Finally, component ablation experiments showed that the individual components of the architecture contributed significantly to the overall performance of the model, and the left-right ablations even provided some insights into the consequences of lateralization.

5.1.1 Primary Contributions

Through the course of the dissertation, the studies performed revealed the limitations of pre-built DL models when applied to psycho-physiological modeling and proposed a number of solutions. In general, the proposed solutions addressed the problems either through robust intermediate feature representations, or through applying regularization to discover effective deep feature embeddings. The primary contributions of the study are summarized as follows:

- Proposed a spectral-temporal ECG-stress modeling architecture that enabled learning of features which could outperform gold-standard HRV for detecting stress. Future works can benefit from ECG models which are more resistant to short bursts of noise compared to end-to-end models, but are also more flexible than HRV-based models.

- Proposed a novel ML approach for modeling meditation of novice and intermediate-level meditators using spectral-spatial EEG topographs. Modern ML studies of meditation are relatively rare, and to our knowledge ours is the first to apply deep learning for predicting meditative states from EEG. In addition, studies on novices meditators are even rarer, so the field also benefits from the easily replicable method presented in this work. Finally, the success of the method at training from noisy EEG samples means that future works can benefit from reduced costs of building usable EEG datasets.

- Introduced BiVDANN, a novel DL architecture for discovering generalized subject-

independent EEG features for emotion classification. The unique combination of regularizers enables BiVDANN to learn embeddings that exhibit both strong subject-invariance and a robust Gaussian distribution, allowing the model to better accommodate new subjects. In general, EEG modeling studies can benefit from the reduced data requirements for building generalized EEG models. In addition, the modular nature of the architectures allows for possible extensions in the future.

5.2 Conclusion

To conclude, this research addressed some of the common problems encountered when training psycho-physiological ML models through a combination of one or more of the following: 1) robust intermediate signal representations, 2) data-efficient compact CNN architectures, 3) model architecture optimizations based on analysis of NN activations and 4) architecture design that encourages learning of generalized subject-invariant features. Overall, all models successfully illustrated how to efficiently train DL models on datasets with different forms of signal noise, from morphological and intermittent spectral noise common to ECG data, to persistent channel noise seen in EEG data. Additionally, the experiments demonstrated how DL models could be optimized to learn from small psycho-physiological datasets, in the absence of gold-standard end-to-end deep feature extractors, by using efficient intermediate data representations. Comparative and ablative experiments showed that the proposed methods were effective in their respective target applications. Specifically, the spectral-temporal ECG-Stress model was more resistant to intermittent signal noise and outperformed HRV-based models, the EEG-Meditation model demonstrated how effective predictive models could be built from small EEG datasets with noisy channels, and finally, the proposed BiVDANN model showed how subject generalization could be achieved by learning regularized feature embeddings from the data. As the models themselves are linked only to their respective intermediate representations and not to the raw signals, the techniques explored in this work could potentially be applied to other signal processing problems that use comparable spectral representations and experience similar noise characteristics. Moreover, when standard end-to-end feature extractors do become available for the featured physiological signals, it should be possible to replace the intermediate feature extractors with those new components and merge them with the most effective architectures tested in this work.

5.3 Limitations

Although most of the proposed methods were shown to be generally successful, we recognize that there remain several key avenues for improvement. In this section, we discuss some of the more prominent weaknesses of this work.

In the case of the ECG-Stress model, the primary limitation is the lack of subject-independent testing. This was mainly due to the size constraints of the dataset. A larger dataset of samples gathered from a large number of subjects would allow the construction and testing of subject-independent approaches. In addition, a limitation of the architecture itself is its reliance on purely convolutional operations. While these were appropriate given the small dataset constraints and the spatial-spectral form of input data, CNNs are usually only effective for learning localized features. An ideal model architecture for non-stationary temporal signals like ECG should consider both short and long temporal patterns in the data. Various improvements could have been added, however most would only increase the data requirements even further.

In the case of the proposed EEG-Meditation model, one obvious limitation was the lack of direct comparisons to similar work. The scarcity of EEG-meditation studies couple with the lack of a standard definition of meditative states makes it difficult to accurately gauge the improvements in the technology. Unlike with many other psychological phenomena, there is currently no standard dataset for testing meditation-based ML models. Another limitation, similar to the ECG-Stress model, was the lack of substantial temporal feature learning. While a spectral topomap can embody short-term temporal patterns to a degree, a set of sliding window spectra would be much more effective at capturing non-stationary signal patterns.

Next, in the case of BiVDANN, it shares the lack of temporal feature learning of the other two approaches. A number of works have already been successful at applying LSTMs to EEG modeling, however these are primarily for applications with larger training sets or simpler problems like detecting event-related potentials (ERP). Presently, the few successful LSTM-based emotion models rely on domain adaptation to achieve reasonable subject-independent performance. With regards to the BiVDANN approach itself, a limitation of the current implementation is that it can be difficult to tune the model's regularization hyper-parameters. Coupled with the large size of the model itself, reaching an optimized state can be very costly in both time and processing power. And finally, while BiVDANN shows better performance when classifying unseen subjects,

the ceiling for performance is severely limited. In cases where some data from the new subject can be used to retrain a model, modern domain adaptation methods may still be preferable. However, a hybrid approach of the two may also be viable.

Finally, in general, a major limitation across all the methods is the requirement of manual feature extraction. The goal of DL approaches is generally to construct end-to-end models which can encapsulate the full range of useful features that can be learned from the raw data. However, the absence of prominent deep feature extractors for physiological data limits the capability of most existing models.

5.4 Recommendations for Future Work

Future work should concentrate on addressing the many limitations presented in this study. Of great priority would be addressing the need for constructing reusable feature extractors for building end-to-end models. This would require the combined use of physiological datasets from various disciplines, including emotion recognition. While there are relatively few psychological datasets for EEG, there are a lot more physiological datasets designed for medical purposes. The challenge for future works would be to design models that can effectively learn from these mixed datasets and possibly also dynamically adapt feature embeddings depending on the target application. Whatever the final form may be, successful feature extractors have the potential to revolutionize ECG and EEG-based ML works. It should then be easy to adapt those feature extractors and combine them with the core ML architectures presented in this work. For example, if a gold-standard spatial-spectral EEG feature extractor is developed in the future, it should be possible to append their outputs to the input layers of the BiVDANN to build a subject-generalized end-to-end model. Depending on the output format of the feature extractor, it may be necessary to replace the initial CNN layers with either dense layers or recurrent layers. The same could apply to the proposed ECG-Stress model and EEG-Meditation model. However, while waiting for such models to become available, other studies could focus on improving some of the limitations of the approaches presented in this work.

For ECG-Stress models, future work could focus on adapting the proposed temporal-spectral model for subject-independent stress detection. However, this would require larger multi-user datasets of naturalistic stress. We must emphasize naturalistic datasets since typical induced stressors such as the Stroop test may not convey the same subtle

features as the chronic stress we experience in everyday life. In the case of model architectures, future work could explore extensions of this proposed model towards temporal featuring learning through the inclusion of structures like LSTMs or Transformers[124]. In case lack of data is still an issue, pre-training certain components of the architecture combined with strong regularization could allow for training an effective model.

In the case of EEG models, a lot of progress is constantly being made due to the proliferation of EEG in HCI applications. Of course, many of these advancements could also be applied to meditation in an attempt to address the scarcity of studies on this field. However, a broader topic for future works could be the exploration of the potential of domain adaptation for feature-adjacent EEG applications. For example, psychological concepts like working memory, attention and meditation share commonalities in brain region and brain wave frequencies. One could imagine that in a map of ideal features for all EEG applications, these would be somewhat adjacent to each other. So, theoretically it should be possible to adapt effective models in one of these fields to the other, and vice-versa. This is currently one of the main goals of many domain adaptation works. BiVDANN also offers similar capabilities, and could easily be extended to enforce generalization across different fields through its adversarial embedding regularizer. For example, it may be possible to adapt the BiVDANN emotion model to other contexts, such as working memory, by combing the feature outputs of both models and introducing an additional context adaptation component to align their feature spaces. This should result in a joint feature space that encodes the most effective features in each context, as well as overlapping features that apply to both. This could make it possible to benefit from the noise rejection and subject-generalization learned from both contexts. However, the Gaussian distribution enforced by the variational regularizer may be too strict for this type of task and may require a relaxation of the variational loss to allow for a configurable multimodal distribution. The same concept may also apply to studies like stress and emotion. Adjacent physiological studies could have the potential to simultaneously advance each other, much in the same way that object recognition and facial recognition studies have done in recent years. Of course, domain adaptation is a generic task that can be applied to almost any data science context, so the most effective methods in these psycho-physiological studies could also be used for other signal processing problems.

And finally, future works in emotion modeling have the option to either add to, or

incorporate portions of, BiVDANN in future architectures. An obvious improvement would be to incorporate temporal learning. This has already been achieved by state-of-the-art works, but normally for domain adaptation. Future works could merge both approaches by creating a domain generalized model that more easily adapts to new users, similar to few-shot learning models. Alternatively, works could focus on training BiVDANN on multiple datasets while adding regularization to account for the change in training context. Theoretically, a model that is designed to understand the context of differing training sets can also be trained to identify features that are invariant across training sets. If the model is trained on a wide range of datasets, it has a higher chance of learning generalizable features. This then loops back to the ultimate goal which is to build models that can be reused across different physiological sensing applications.

Bibliography

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[2] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

[3] Pranav Rajpurkar, Awni Y Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y Ng. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv preprint arXiv:1707.01836*, 2017.

[4] Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 25(1):65–69, 2019.

[5] Beanbonyka Rim, Nak-Jun Sung, Sedong Min, and Min Hong. Deep learning in physiological signal data: A survey. *Sensors*, 20(4):969, 2020.

[6] Soraia M Alarcao and Manuel J Fonseca. Emotions recognition using eeg signals: A survey. *IEEE Transactions on Affective Computing*, 10(3):374–393, 2017.

[7] Gil Keren, Tobias Kirschstein, Erik Marchi, Fabien Ringeval, and Björn Schuller. End-to-end learning for dimensional emotion recognition from physiological signals. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 985–990. IEEE, 2017.

[8] Pouya Bashivan, Irina Rish, Mohammed Yeasin, and Noel Codella. Learning representations from EEG with deep recurrent-convolutional neural networks. In

Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL `http://arxiv.org/abs/1511.06448`.

[9] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017.

[10] Linda J Levine and Stewart L Burgess. Beyond general arousal: Effects of specific emotions on memory. *Social Cognition*, 15(3):157–181, 1997.

[11] Jason Matthew Harley. Measuring emotions: a survey of cutting edge methodologies used in computer-based learning environment research. *Emotions, technology, design, and learning*, pages 89–114, 2016.

[12] Russell Beale and Christian Peter. The role of affect and emotion in hci. In *Affect and emotion in human-computer interaction*, pages 1–11. Springer, 2008.

[13] Rosalind W Picard. *Affective computing*. MIT press, 2000.

[14] H. Iso, C. Date, A. Yamamoto, H. Toyoshima, N. Tanabe, S. Kikuchi, T. Kondo, Y. Watanabe, Y. Wada, T. Ishibashi, H. Suzuki, A. Koizumi, Y. Inaba, A. Tamakoshi, and Y. Ohno. Perceived mental stress and mortality from cardiovascular disease among Japanese men and women: the Japan Collaborative Cohort Study for Evaluation of Cancer Risk Sponsored by Monbusho (JACC Study). *Circulation*, 106(10):1229–1236, Sep 2002.

[15] E. Kowata, A. Hozawa, M. Kakizaki, Y. Tomata, M. Nagai, Y. Sugawara, S. Kuriyama, and I. Tsuji. [Perceived stress and cardiovascular disease mortality. The Ohsaki Cohort Study]. *Nihon Koshu Eisei Zasshi*, 59(2):82–91, Feb 2012.

[16] M. Marcus, M.T. Yasamy, M. van Ommeren, and D. Chisholm. Depression: A global public health concern. Technical report, World Health Organization, 2012.

[17] G. Acampora, D.J. Cook, P. Rashidi, and A.V. Vasilakos. A survey on ambient intelligence in healthcare. *Proceedings of the IEEE*, 101(12):2470–2494, Dec 2013. ISSN 0018-9219. doi: 10.1109/JPROC.2013.2262913.

[18] Adnan Ghaderi, Javad Frounchi, and Alireza Farnam. Machine learning-based signal processing using physiological signals for stress detection. In *2015 22nd Iranian Conference on Biomedical Engineering (ICBME)*, pages 93–98. IEEE, 2015.

[19] J.A. Healey and R.W. Picard. Detecting stress during real-world driving tasks using physiological sensors. *Intelligent Transportation Systems, IEEE Transactions on*, 6(2):156–166, June 2005. ISSN 1524-9050. doi: 10.1109/TITS.2005.848368.

[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1409.1556`.

[21] Gari D Clifford and Lionel Tarassenko. Quantifying errors in spectral estimates of hrv due to beat replacement and resampling. *IEEE transactions on biomedical engineering*, 52(4):630–638, 2005.

[22] Shweta H Jambukia, Vipul K Dabhi, and Harshadkumar B Prajapati. Classification of ecg signals using machine learning techniques: A survey. In *2015 International Conference on Advances in Computer Engineering and Applications*, pages 714–721. IEEE, 2015.

[23] Edgar P Torres P, Edgar A Torres, Myriam Hernández-Álvarez, Sang Guun Yoo, et al. Eeg-based bci emotion recognition: A survey. *Sensors*, 20(18):5083, 2020.

[24] Muhammad Ghifary. *Domain adaptation and domain generalization with representation learning*. PhD thesis, Victoria University of Wellington, 2016.

[25] Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5): 1–46, 2020.

[26] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pages 8093–8104. PMLR, 2020.

[27] Paolo Melillo, Marcello Bracale, and Leandro Pecchia. Nonlinear heart rate variability features for real-life stress detection. case study: students under stress due

to university examination. *BioMedical Engineering OnLine*, 10:96+, November 2011. ISSN 1475-925X. doi: 10.1186/1475-925x-10-96. URL `http://dx.doi.org/10.1186/1475-925x-10-96`.

[28] Gülce N Dikecligil and Lilianne R Mujica-Parodi. Ambulatory and challenge-associated heart rate variability measures predict cardiac responses to real-world acute emotional stress. *Biological psychiatry*, 67(12):1185–1190, 2010.

[29] Ahmad Rauf Subhani, Likun Xia, Aamir Saeed Malik, and Zahiruddin Othman. Quantification of physiological disparities and task performance in stress and control conditions. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2060–2063. IEEE, 2013.

[30] Nagesh Jadhav and Rekha Sugandhi. Survey on human behavior recognition using affective computing. In *2018 IEEE Global Conference on Wireless Computing and Networking (GCWCN)*, pages 98–103. IEEE, 2018.

[31] M Shamim Hossain and Ghulam Muhammad. Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion*, 49: 69–78, 2019.

[32] Juan Lorenzo Mutia Hagad, Ken-ichi Fukui, and Masayuki Numao. Modelling naturalistic work stress using spectral hrv representations and deep learning. In *Proceedings of the Annual Conference of JSAI 33rd Annual Conference, 2019*, pages 2H5E204–2H5E204. The Japanese Society for Artificial Intelligence, 2019.

[33] Markad V Kamath, Dhanjoo N Ghista, Ernest L Fallen, David Fitchett, Donald Miller, and Robert McKelvie. Heart rate variability power spectrogram as a potential noninvasive signature of cardiac regulatory system response, mechanisms, and disorders. *Heart and vessels*, 3(1):33–41, 1987.

[34] Julian F Thayer, Fredrik Åhs, Mats Fredrikson, John J Sollers III, and Tor D Wager. A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health. *Neuroscience & Biobehavioral Reviews*, 36(2):747–756, 2012.

[35] Hye-Geum Kim, Eun-Jin Cheon, Dai-Seg Bai, Young Hwan Lee, and Bon-Hoon

Koo. Stress and heart rate variability: A meta-analysis and review of the literature. *Psychiatry investigation*, 15(3):235, 2018.

[36] Yuan Shi, Minh Hoai Nguyen, Patrick Blitz, Brian French, Scott Fisk, Fernando De la Torre, Asim Smailagic, Daniel P Siewiorek, Mustafa al'Absi, Emre Ertin, et al. Personalized stress detection from physiological measurements. *International symposium on quality of life technology*, pages 28–29, 2010.

[37] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1301–1309, 2017.

[38] Selcan Kaplan Berkaya, Alper Kursat Uysal, Efnan Sora Gunal, Semih Ergin, Serkan Gunal, and M Bilginer Gulmezoglu. A survey on ecg analysis. *Biomedical Signal Processing and Control*, 43:216–235, 2018.

[39] JD Schipke, G Arnold, and M Pelzer. Effect of respiration rate on short-term heart rate variability. *Journal of Clinical and Basic Cardiology*, 2(1):92–95, 1999.

[40] Peter Welch. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, 15(2):70–73, 1967.

[41] T Jr Deeming. Fourier analysis with unequally-spaced data. *Astrophysics and Space Science*, 36(1):137–158, 1975.

[42] Jacob T VanderPlas. Understanding the lomb–scargle periodogram. *The Astrophysical Journal Supplement Series*, 236(1):16, 2018.

[43] G.B. Moody. Spectral analysis of heart rate without resampling. In *Computers in Cardiology 1993, Proceedings.*, pages 715–718, Sep 1993. doi: 10.1109/CIC.1993.378302.

[44] Jeffrey D Scargle. Studies in astronomical time series analysis. ii-statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal*, 263:835–853, 1982.

[45] HPA Van Dongen, E Olofsen, JH Van Hartevelt, and EW Kruyt. A procedure of multiple period searching in unequally spaced time-series with the lomb–scargle method. *Biological Rhythm Research*, 30(2):149–177, 1999.

[46] Costel Munteanu, Catalin Negrea, Marius Echim, and Kalevi Mursula. Effect of data gaps: comparison of different spectral analysis methods. In *Annales Geophysicae*, volume 34, pages 437–449. Copernicus GmbH, 2016.

[47] Xinglei Zhu, Gerald T Beauregard, and Lonce L Wyse. Real-time signal estimation from modified short-time fourier transform magnitude spectra. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1645–1653, 2007.

[48] Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116, 1998.

[49] Juan Lorenzo Hagad, Ken-ichi Fukui, and Masayuki Numao. A hierarchical model of authentic work stress using physiological signals and stress coping profiles. In *Proceedings of 30th Annual Conference of the Japanese Society for Artificial Intelligence*, pages 3E34–3E34. The Japanese Society for Artificial Intelligence, 2016.

[50] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. 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 (June 13).

[51] Andreas Voss, Rico Schroeder, Andreas Heitmann, Annette Peters, and Siegfried Perz. Short-term heart rate variability-influence of gender and age in healthy subjects. *PloS one*, 10(3):e0118308, 2015.

[52] Fred Shaffer and JP Ginsberg. An overview of heart rate variability metrics and norms. *Frontiers in public health*, 5:258, 2017.

[53] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

[54] Alberto Chiesa and Alessandro Serretti. Mindfulness-based stress reduction for stress management in healthy people: a review and meta-analysis. *The journal of alternative and complementary medicine*, 15(5):593–600, 2009.

[55] Dawn Bazarko, Rebecca A Cate, Francisca Azocar, and Mary Jo Kreitzer. The impact of an innovative mindfulness-based stress reduction program on the health and well-being of nurses employed in a corporate setting. *Journal of workplace behavioral health*, 28(2):107–133, 2013.

[56] Paul Grossman. On measuring mindfulness in psychosomatic and psychological research. *Journal of Psychosomatic Research, 64*, 2008.

[57] Han-Jeong Hwang, Soyoun Kim, Soobeom Choi, and Chang-Hwan Im. Eeg-based brain-computer interfaces: a thorough literature survey. *International Journal of Human-Computer Interaction*, 29(12):814–826, 2013.

[58] Phattarapong Sawangjai, Supanida Hompoonsup, Pitshaporn Leelaarporn, Supavit Kongwudhikunakorn, and Theerawit Wilaiprasitporn. Consumer grade eeg measuring sensors as research tools: A review. *IEEE Sensors Journal*, 20(8): 3996–4024, 2019.

[59] Amar Kachenoura, Laurent Albera, Lotfi Senhadji, and Pierre Comon. Ica: a potential tool for bci systems. *IEEE Signal Processing Magazine*, 25(1):57–68, 2007.

[60] Chi-Yuan Chang, Sheng-Hsiou Hsu, Luca Pion-Tonachini, and Tzyy-Ping Jung. Evaluation of artifact subspace reconstruction for automatic eeg artifact removal. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1242–1245. IEEE, 2018.

[61] M Rajya Lakshmi, TV Prasad, and Dr V Chandra Prakash. Survey on eeg signal processing methods. *International Journal of Advanced Research in Computer Science and Software Engineering*, 4(1), 2014.

[62] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, 16(5):051001, 2019.

[63] Juan Lorenzo Hagad, Kenichi Fukui, and Masayuki Numao. Deep visual models for eeg of mindfulness meditation in a workplace setting. In *International Workshop on Health Intelligence*, pages 129–137. Springer, 2019.

[64] Pierre Thodoroff, Joelle Pineau, and Andrew Lim. Learning robust features using deep learning for automatic seizure detection. In *Machine learning for healthcare conference*, pages 178–190, 2016.

[65] Albert Vilamala, Kristoffer H Madsen, and Lars K Hansen. Deep convolutional neural networks for interpretable analysis of eeg sleep stage scoring. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2017.

[66] Zhicheng Jiao, Xinbo Gao, Ying Wang, Jie Li, and Haojun Xu. Deep convolutional neural networks for mental load classification based on eeg data. *Pattern Recognition*, 76:582–595, 2018.

[67] Rohan Dixit. Meditation training and neurofeedback using a personal eeg device. In *2012 AAAI Spring Symposium Series*, 2012.

[68] Asieh Ahani, Helane Wahbeh, Hooman Nezamfar, Meghan Miller, Deniz Erdogmus, and Barry Oken. Quantitative change of eeg and respiration signals during mindfulness meditation. *Journal of neuroengineering and rehabilitation*, 11(1): 1–11, 2014.

[69] Esmeralda C Djamal, Dewi P Pangestu, and Dea A Dewi. Eeg-based recognition of attention state using wavelet and support vector machine. In *2016 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, pages 139–144. IEEE, 2016.

[70] Hong Lin and Yuezhe Li. Using eeg data analytics to measure meditation. In *International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management*, pages 270–280. Springer, 2017.

[71] Tim R Mullen, Christian AE Kothe, Yu Mike Chi, Alejandro Ojeda, Trevor Kerth, Scott Makeig, Tzyy-Ping Jung, and Gert Cauwenberghs. Real-time neuroimaging and cognitive monitoring using wearable dry eeg. *IEEE Transactions on Biomedical Engineering*, 62(11):2553–2567, 2015.

[72] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. Meg and eeg data analysis with mne-python. *Frontiers in neuroscience*, 7:267, 2013.

[73] Fang Wang, Sheng-hua Zhong, Jianfeng Peng, Jianmin Jiang, and Yan Liu. Data augmentation for eeg-based emotion recognition with deep convolutional neural networks. In *International Conference on Multimedia Modeling*, pages 82–93. Springer, 2018.

[74] Openbci ultracortex mark iv. `https://docs.openbci.com/docs/04AddOns/01-Headwear/MarkIV`, 2021. Accessed: 2021-05-24.

[75] Diana Winston. Guided meditations - UCLA mindfulness awareness research center. `https://www.uclahealth.org/marc/mindful-meditations`, 2018. Accessed: 2018-11-02.

[76] Juan Lorenzo Hagad, Tsukasa Kimura, Ken-ichi Fukui, and Masayuki Numao. Learning subject-generalized topographical eeg embeddings using deep variational autoencoders and domain-adversarial regularization. *Sensors*, 21(5):1792, 2021.

[77] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[78] Zirui Lan, Olga Sourina, Lipo Wang, Reinhold Scherer, and Gernot R Müller-Putz. Domain adaptation techniques for eeg-based emotion recognition: a comparative study on two public datasets. *IEEE Transactions on Cognitive and Developmental Systems*, 11(1):85–94, 2018.

[79] Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *The annals of statistics*, pages 1171–1220, 2008.

[80] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2010.

[81] Ke Yan, Lu Kou, and David Zhang. Learning domain-invariant subspace using domain features and independence maximization. *IEEE transactions on cybernetics*, 48(1):288–299, 2017.

[82] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

[83] Yucel Cimtay and Erhan Ekmekcioglu. Investigating the use of pretrained convolutional neural network on cross-subject and cross-dataset eeg emotion recognition. *Sensors*, 20(7):2034, 2020.

[84] Yun Luo, Si-Yang Zhang, Wei-Long Zheng, and Bao-Liang Lu. Wgan domain adaptation for eeg-based emotion recognition. In *International Conference on Neural Information Processing*, pages 275–286. Springer, 2018.

[85] Enver Sangineto, Gloria Zen, Elisa Ricci, and Nicu Sebe. We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 357–366, 2014.

[86] Yang Li, Wenming Zheng, Yuan Zong, Zhen Cui, Tong Zhang, and Xiaoyan Zhou. A bi-hemisphere domain adversarial neural network model for eeg emotion recognition. *IEEE Transactions on Affective Computing*, 2018.

[87] Anh Viet Phan, Minh Le Nguyen, Yen Lam Hoang Nguyen, and Lam Thu Bui. Dgcnn: A convolutional neural network over large-scale labeled graphs. *Neural Networks*, 108:533–543, 2018.

[88] Jinpeng Li, Shuang Qiu, Changde Du, Yixin Wang, and Huiguang He. Domain adaptation for eeg emotion recognition based on latent representation similarity. *IEEE Transactions on Cognitive and Developmental Systems*, 2019.

[89] Tor D Wager, K Luan Phan, Israel Liberzon, and Stephan F Taylor. Valence, gender, and lateralization of functional brain anatomy in emotion: a meta-analysis of findings from neuroimaging. *Neuroimage*, 19(3):513–531, 2003.

[90] Philip Haeusser, Thomas Frerix, Alexander Mordvintsev, and Daniel Cremers. Associative domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2765–2773, 2017.

[91] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader

and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[92] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

[93] Pallavi Pandey and KR Seeja. Subject independent emotion recognition from eeg using vmd and deep learning. *Journal of King Saud University-Computer and Information Sciences*, 2019.

[94] Konstantin Dragomiretskiy and Dominique Zosso. Variational mode decomposition. *IEEE transactions on signal processing*, 62(3):531–544, 2013.

[95] Mengxi Dai, Dezhi Zheng, Rui Na, Shuai Wang, and Shuailei Zhang. Eeg classification of motor imagery using a novel deep learning framework. *Sensors*, 19(3): 551, 2019.

[96] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[97] Xin Chai, Qisong Wang, Yongping Zhao, Xin Liu, Ou Bai, and Yongqiang Li. Unsupervised domain adaptation techniques based on auto-encoder for non-stationary eeg-based emotion recognition. *Computers in biology and medicine*, 79: 205–214, 2016.

[98] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[99] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Lauri Parkkonen, and Matti S Hämäläinen. Mne software for processing meg and eeg data. *Neuroimage*, 86:446–460, 2014.

[100] Donald B Percival, Andrew T Walden, et al. *Spectral analysis for physical applications*. cambridge university press, 1993.

[101] Xiao-Wei Wang, Dan Nie, and Bao-Liang Lu. Emotional state classification from eeg data using machine learning approach. *Neurocomputing*, 129:94–106, 2014.

[102] Ning Zhuang, Ying Zeng, Li Tong, Chi Zhang, Hanming Zhang, and Bin Yan. Emotion recognition from eeg signals using multidimensional information in emd domain. *BioMed research international*, 2017, 2017.

[103] Noppadon Jatupaiboon, Setha Pan-ngum, and Pasin Israsena. Emotion classification using minimal eeg channels and frequency bands. In *The 2013 10th international joint conference on Computer Science and Software Engineering (JCSSE)*, pages 21–24. IEEE, 2013.

[104] Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. Differential entropy feature for eeg-based emotion classification. In *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 81–84. IEEE, 2013.

[105] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=Sy2fzU9gl`.

[106] Xianxu Hou, Linlin Shen, Ke Sun, and Guoping Qiu. Deep feature consistent variational autoencoder. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1133–1141. IEEE, 2017.

[107] Hang Li, Haozheng Wang, Zhenglu Yang, and Masato Odagaki. Variation autoencoder based network representation learning for classification. In *Proceedings of ACL 2017, Student Research Workshop*, pages 56–61, 2017.

[108] Yang Li, Quan Pan, Suhang Wang, Haiyun Peng, Tao Yang, and Erik Cambria. Disentangled variational auto-encoder for semi-supervised learning. *Information Sciences*, 482:73–85, 2019.

[109] Nik Khadijah Nik Aznan, Amir Atapour-Abarghouei, Stephen Bonner, Jason D Connolly, Noura Al Moubayed, and Toby P Breckon. Simulating brain signals: Creating synthetic eeg data via neural-based generative models for improved

ssvep classification. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[110] Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. ISSN 1935-8237. doi: 10.1561/2200000056. URL `http://dx.doi.org/10.1561/2200000056`.

[111] Christopher P. Burgess, Irina Higgins, Arka Pal, Loïc Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *CoRR*, abs/1804.03599, 2018. URL `http://arxiv.org/abs/1804.03599`.

[112] Simao Eduardo, Alfredo Nazábal, Christopher KI Williams, and Charles Sutton. Robust variational autoencoders for outlier detection and repair of mixed-type data. In *International Conference on Artificial Intelligence and Statistics*, pages 4056–4066. PMLR, 2020.

[113] Youzhi Tu, Man-Wai Mak, and Jen-Tzung Chien. Variational domain adversarial learning for speaker verification. In *Interspeech*, pages 4315–4319, 2019.

[114] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.

[115] Henry Candra, Mitchell Yuwono, Rifai Chai, Ardi Handojoseno, Irraivan Elamvazuthi, Hung T Nguyen, and Steven Su. Investigation of window size in classification of eeg-emotion signal with wavelet entropy and support vector machine. In *2015 37th Annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 7250–7253. IEEE, 2015.

[116] David J Thomson. Jackknifing multitaper spectrum estimates. *IEEE Signal Processing Magazine*, 24(4):20–30, 2007.

[117] François Chollet et al. Keras: The python deep learning library. *Astrophysics Source Code Library*, pages ascl–1806, 2018.

[118] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference*

on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL `http://arxiv.org/abs/1412.6980`.

[119] Sebastian Mika, Bernhard Schölkopf, Alex J Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch. Kernel pca and de-noising in feature spaces. In *Advances in neural information processing systems*, pages 536–542, 1999.

[120] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[121] Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[122] Guido Gainotti. Emotional behavior and hemispheric side of the lesion. *Cortex*, 8 (1):41–55, 1972.

[123] Joan C Borod, Barbara A Cicero, Loraine K Obler, Joan Welkowitz, Hulya M Erhan, Cornelia Santschi, Ilana S Grunwald, Reto M Agosti, and John R Whalen. Right hemisphere emotional perception: evidence across multiple channels. *Neuropsychology*, 12(3):446, 1998.

[124] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.