UNIVERSITY OF CALIFORNIA,
IRVINE

Deep Latent Variable Modeling of Physiological Signals

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Computer Science

by

Khuong An Vo

Dissertation Committee:
Professor Nikil Dutt, Chair
Associate Professor Hung Cao
Professor Ramesh Srinivasan
Professor Erik Sudderth

2024

DEDICATION

This thesis is dedicated to my parents, sister, and brother for their unwavering love, support, and encouragement.

TABLE OF CONTENTS

Page

LIST OF FIGURES **v**

LIST OF TABLES **ix**

ACKNOWLEDGMENTS **x**

VITA **xii**

ABSTRACT OF THE DISSERTATION **xiv**

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 Outline and contributions . . . 4

2 Background **7**
- 2.1 Physiological Signals . . . 7
  - 2.1.1 Electrocardiogram (ECG or EKG) . . . 7
  - 2.1.2 Photoplethysmogram (PPG) . . . 8
  - 2.1.3 Electroencephalogram (EEG) . . . 9
  - 2.1.4 Frequency Domain Representation . . . 10
- 2.2 Probabilistic Graphical Models . . . 11
  - 2.2.1 Random Variables and Probabilities . . . 11
  - 2.2.2 Graphical Models . . . 13
- 2.3 Latent Variable Models . . . 17
  - 2.3.1 Posterior Inference . . . 20
  - 2.3.2 Parameter Learning . . . 24

3 PPG-to-ECG Signal Translation For Continuous Atrial Fibrillation Detection via Attention-based Deep State-Space Modeling **27**
- 3.1 Introduction . . . 27
- 3.2 Methodology . . . 30
  - 3.2.1 Probabilistic Modeling of ECG from PPG signals . . . 30
  - 3.2.2 State-Space Modeling of ECG from PPG Signals . . . 31
  - 3.2.3 Neural Network Architectures . . . 36
- 3.3 Experiments . . . 37

3.3.1 Dataset . . . 37
3.3.2 Evaluation Metrics . . . 38
3.3.3 Implementation and Results . . . 39
3.4 Conclusion . . . 43

4 Composing Graphical Models with Generative Adversarial Networks for EEG Signal Modeling 45
4.1 Introduction . . . 45
4.2 Methodology . . . 47
4.2.1 EEG Signal Synthesis with GANs . . . 47
4.2.2 Conjoining GANs with Bayesian Networks . . . 48
4.2.3 Network Architectures and Training Hyperparameters . . . 52
4.3 Experiments . . . 54
4.3.1 Dataset . . . 54
4.3.2 Evaluation Metrics . . . 54
4.3.3 Results and Discussion . . . 55
4.4 Conclusion . . . 58

5 Deep Latent Variable Joint Cognitive Modeling of Neural Signals and Human Behavior 60
5.1 Introduction . . . 60
5.2 Neurocognitive Variational Autoencoders . . . 62
5.2.1 Generative EEG Modeling with VAEs . . . 62
5.2.2 Disentangled Cognitive Latent Space of EEG . . . 65
5.2.3 Structured EEG Modeling from Behavior . . . 66
5.3 Experiments . . . 68
5.3.1 EEG and Behavioral Dataset . . . 68
5.3.2 Results . . . 68
5.4 Conclusion . . . 79
5.5 Supplementary Materials . . . 81
5.5.1 Neural Network Architectures and Training Hyperparameters . . . 81
5.5.2 Simulation Studies . . . 82
5.5.3 Experimental Tasks . . . 83
5.5.4 Decision-Making Models - The Drift-Diffusion Model (DDMs) . . . 85

6 Conclusions 88
6.1 Contributions . . . 88
6.2 Future Work . . . 89

Bibliography 91

LIST OF FIGURES

Page

2.1 Examples of different probabilistic graphical models. 13

3.1 A PPG-ECG waveform pair. PPG signals can often become contaminated by noise. 28
3.2 The graphical model for ECG translation from PPG. Shaded nodes represent observed variables. Clear nodes represent latent variables. Diamond nodes denote deterministic variables. Variables $\boldsymbol{x}_t, \boldsymbol{y}_t$, and $\boldsymbol{c}_t$ represent PP intervals, RR intervals, and context vectors, respectively. $\alpha_{t,i}$ are attention weights defines how well two intervals $\boldsymbol{x}_i$ and $\boldsymbol{y}_t$ are aligned. The attention mechanism is shown only at time step 2. 33
3.3 The graphical model at latent state inference time. Variables $\boldsymbol{y}_t, \boldsymbol{h}_t, \boldsymbol{g}_t$, and $\boldsymbol{z}_t$ represent respectively RR intervals, backward, forward recurrent states, and latent states. 34
3.4 Examples of the translated ECG signals. In each subfigure: the top panel shows the input PPG waveform and the bottom panel shows the reconstructed ECG waveform compared with the reference waveform. The average ECG waveform (dark blue) of all possible pulses overlaid on each individual pulse (light blue). 44

4.1 Directed graphical models for EEG signal modeling. Each time step corresponds to a δ-second multi-channel signal. Shaded nodes represent observed variables. Clear nodes represent latent variables. Directed edges indicate statistical dependencies between variables. 48
4.2 Last 10-second of a 30-second synthetic 23-channel EEG signal by the GMMarkov-GAN model, conditioned on patient 3. 5 channels with the highest standard deviations are shown. 56
4.3 t-SNE visualization of the static latent spaces. 58
4.4 ROC curve for epilepsy seisure detection. 59

5.1 The Neurocognitive VAE. After the generative process (a) learns the joint latent neurocognitive variables (Section 5.2.2), the regularized discriminative process (b) retrofits its hierarchical latent space to the joint latent space (Section 5.2.3). Inference networks q and Generation networks p contain neural network parameters θ and ϕ. Black arrows: flows of operations. Red arrows: loss functions. MSE and WFPT stand for Mean Squared Error and Wiener First Passage Time, respectively. The heatmaps represent the probability distributions in the latent spaces. Plasma color maps are for the drift-diffusion variables ($z_C \in \mathbb{R}^3$), while greenery color maps are for residual neural variables ($z_N \in \mathbb{R}^{32}$). Blue blocks contain μ and σ, which are the parameters of the multivariate Gaussian latent spaces. Gray blocks contain z sampled ($\sim$) from the distributions. The variables x and y represent EEG signals and choice-RTs, respectively. Each trapezoid represents a different convolutional neural network (see Table 5.2 for detailed architectures). 63

5.3 Drift-diffusion single-trial parameter estimations from correct responses of subject s1. The parameters are constrained by the subject priors resulting from a Bayesian MCMC modeling (without EEG data). Scatter plots illustrate the relationship between the parameters and the observed choice-RTs for each trial. The top two rows are posterior inferences from neural signals, while the bottom two are from behaviors. The left column shows the drift-rate (δ) estimates, the middle column shows boundary (α) estimates, and the right column presents non-decision time (ndt) estimates. The correlations between the choice-RTs and the inferred DDM parameters are consistent with what is expected. On top of each panel are the Spearman correlation coefficients (ρ). The covariances of the inferred parameters are indicated by circles, which correspond to contours having one standard deviation. For clarity, each circle is magnified 300 times. 74

5.5 Performance of the model in reconstructing 98 EEG channels of subject s1 by averaging $\approx$ 800 predicted EEG trials from $\approx$ 800 choice-RTs in the test set. Time point zero denotes the time point of stimulus onset. The first row displays the original (blue) and generated (orange) trial-averaged EEG data at the pooled electrodes. The x-axis denotes the time in milliseconds from stimulus onset, and the y-axis denotes the signal amplitude. The second, third, and fourth rows are (left) frequency spectra and (right) EEG signals averaged over all test choice-RT trials ($\approx$ 800/3 per condition). The signals on the right are low-pass filtered at 15 Hz for clarity of N200 peaks. Each colored line corresponds to one reconstructed EEG channel. In low-noise conditions, the spectra show a strong peak at the Gabor flicker frequency of 30 Hz, and the ERP waveform shows a shorter N200 latency and larger peak amplitude. Under high-noise conditions, the spectra show a strong peak at the noise flicker frequency of 40 Hz, and the ERP waveform shows a longer N200 latency and a smaller peak amplitude. 76

5.6 Performance of the model in reconstructing single-trial N200 peaks from choice-RTs in four subjects. The dotted lines are references to the original data. The distributions of (left) single-trial N200 peak latencies across three noise conditions and (right) the N200 peak amplitude statistics are shown. Single-trial observations of the peak latency of N200 are found using the SVD method (Nunez et al., 2019) for each subject and noise condition. 77
5.8 Sensitivity analysis of choice-RTs and latent drift-diffusion parameters on EEG signal generation in four subjects. The left column presents the effects of choice-RTs on the output neural signals. The blue bars represent the power at 30 Hz, while the red bars represent the power at 40 Hz. The orange bars show the N200 latencies. The middle column shows the changes in the single-trial N200 distribution w.r.t to *hypothetical* changes in the cognitive parameters. The yellow distribution represents the reference data, while the blue and red ones correspond to modified parameter settings that decrease or increase the N200 latencies, respectively. The modification in subject s4 (ndt $\pm$ 0.05, $\delta\pm$ 0.3) is different from other subjects. The right column characterizes the changes in 30 Hz and 40 Hz peaks w.r.t to the changes in the same cognitive parameters. 79
5.9 Drift-diffusion parameter estimates from neural signals in a simulation of trial-level choice RTs and EEG signals. The top panels show the overlap between the recovered and the original distributions of trial-specific drift-rate and NDT. The reference values for the drift rate and NDT are drawn from the normal distributions $\mathcal{N}(1.5, 0.2)$ and $\mathcal{N}(0.3, 0.05)$, respectively. The bottom scatter plots illustrate the relationship between the recovered parameters and the original parameters each trial. ρ are the Spearman correlation coefficients. 83
5.10 Example stimuli of the cue and response intervals of medium noise conditions (Nunez et al., 2019). In the response phase, participants identified the spatial-frequency target represented by each Gabor, using their left hand to press a button for a target with a low spatial frequency (2.4 cpd) and their right hand for a target with a high spatial frequency (2.6 cpd). N200 waveforms were calculated time-locked to the onset of the Gabor stimulus during the response intervals. The visual noise altered at a frequency of 40 Hz, while the Gabor signal modulated at 30 Hz, inducing 40 Hz and 30 Hz electrocortical responses that monitor attention to both noise and signal. 84

5.11 The DDM is illustrated in action during a two-choice task, with non-decision time shown in green. Following the visual encoding period, the decision variable (DV) begins evidence accumulation and reaches either the upper or lower limit for each trial. The black vector depicts the average rate of evidence accumulation. The blue curve depicts the distribution of response times when choice 1 is successfully picked, while the red curve depicts the distribution of reaction times when choice 2 is correctly selected. When the DV drifts towards the incorrect boundary owing to random noise, incorrect decisions are made. The distribution of reaction times for incorrect trials is depicted by the dotted curve. EEG data for each trial, processed using singular value decomposition to highlight N200, is shown on top that track the start of evidence accumulation. 86

LIST OF TABLES

Page

3.1 ECG translation performance of different models. The top three rows show models' performance on healthy subjects, while the fourth row shows the performance on both the healthy and AFib subjects. If not specified, healthy subjects and clean signals is the default setting. The LSTM model (Tang et al., 2022) is subject-dependent, while the P2E-WGAN (Vo et al., 2021) and our model are subject-independent. 40

3.2 AFib detection performance. The performance on the translated ECG is evaluated when the MINA model (Hong et al., 2019) is trained on real ECG but tested on synthetic ECG. The fusion performance is when the MINA model is extended to receive both real ECG and synthetic ECG inputs. x% random time samples are omitted, simulating intermittent ECG recording, while synthetic ECG is always available. 42

4.1 Network architectures. Models having similar architectures are grouped together. 53

4.2 Performances of different GAN models in interictal EEG signal synthesis and reconstruction tasks. 56

5.1 Comparison of the sum of Wiener negative log-likelihood ($-\sum \log$ Wiener $(\mathrm{RT}_i \mid \boldsymbol{\omega}_i)$) of four subjects on the test sets. $\bar{\boldsymbol{\omega}}$ represents the median fitted cognitive parameters from the training set. 69

5.2 Neural network parametrization 82

ACKNOWLEDGMENTS

This PhD journey has been the most challenging yet highly fulfilling experience of my life. Achieving this milestone would not have been possible without the individuals who have become part of it along the way.

First of all, I would like to thank the members of my thesis committee for their invaluable feedback, constructive criticism, and unwavering support throughout this process. I am grateful to Prof. Nikil Dutt for his warm welcome to UCI and for his guidance in research from day one. His advice on presentations, paper writing, and networking has significantly improved my skills as a researcher. I thoroughly enjoyed our discussions on energy-efficient deep learning. I extend my heartfelt thanks to Prof. Hung Cao for offering me numerous opportunities for both academic and personal growth. Prof. Cao inspired my work on AI in healthcare, involving me in multidiscplinary teams where I learned extensively, leading to many fruitful collaborations and ideas for commercializing research projects. I am thankful to Prof. Ramesh Srinivasan for his guidance on challenging yet intellectually stimulating neurocognitive projects. Working with him allowed me to delve deeply into the most complex problems in AI and neuroscience. His patience, amiable advising approach, and expert knowledge have been immensely helpful in overcoming numerous significant challenges in the research. I deeply appreciate Prof. Erik Sudderth for his insightful feedback and suggestions that have greatly influenced my research. His course on Learning in Graphical Models provided the essential groundwork for the main topic of this thesis.

I also would like to express my gratitude to my supervisors at Samsung Device Solution Research America, Dr. Mostafa El-Khamy and Dr. Yoojin Choi, for their extensive guidance and support, which extended beyond the duration of my internship. Portions of this dissertation were carried out during my time as an intern.

Before embarking on my PhD journey, I was privileged to have met and been inspired by several accomplished scholars, particularly Prof. Khanh Dang, Prof. Tho Quan, Mao Nguyen, and Minh Nguyen. Their influence motivated me to engage in research and pursue graduate studies in the United States.

A special thanks goes to my friends and fellow Ph.D. candidates at UCI, whose collaboration and camaraderie made this journey a memorable one. These include Michale Lee, Tai Le, Jenny Sun, Manoj Vishwanath, Emad Kasaeyan Naeini, Isaac Menchaca, Floranne Ellington, Amir Naderi, Alex Huang, Sadaf Sarafan, Ramses Torres, Anh Nguyen, Steven Cao, and many others.

I am grateful to the National Science Foundation (NSF) and the National Institutes of Health (NIH) for providing the financial support that enabled me to pursue and complete this research. Without their generosity, this work would not have been possible. Funding resources: NSF grant #1917105, #1658303, #1850849, #2051186, #2126976, and NIH SBIR grant #R44OD024874.

Chapter 3 of this dissertation is an adaptation of the material as it appears in "PPG-to-ECG Signal Translation for Continuous Atrial Fibrillation Detection via Attention-based Deep State-Space Modeling", International Conference of the IEEE Engineering in Medicine and Biology Society, used with permission from IEEE. The coauthors on this publication are Mostafa El-Khamy and Yoojin Choi.

Chapter 4 of this dissertation is a reprint of the material as it appears in "Composing Graphical Models with Generative Adversarial Networks for EEG Signal Modeling", IEEE International Conference on Acoustics, Speech and Signal Processing, used with permission from IEEE. The coauthors on this publication are Manoj Vishwanath, Ramesh Srinivasan, Nikil Dutt, and Hung Cao. Ramesh Srinivasan, Nikil Dutt, and Hung Cao directed and supervised research which forms the basis for the dissertation.

Chapter 5 of this dissertation is an adaptation of the material as it appears in "Deep Latent Variable Joint Cognitive Modeling of Neural Signals and Human Behavior", NeuroImage, used with permission from Elsevier. The coauthors on this publication are Qinhua Jenny Sun, Michael D. Nunez, Joachim Vandekerckhove, and Ramesh Srinivasan. Ramesh Srinivasan directed and supervised research which forms the basis for the dissertation.

VITA

Khuong An Vo

EDUCATION

Doctor of Philosophy in Computer Science **2024**
University of California, Irvine Irvine, California

Master of Science in Computer Science **2021**
University of California, Irvine Irvine, California

Bachelor of Engineering in Computer Science **2017**
Vietnam National University HCMC, Vietnam

RESEARCH EXPERIENCE

Graduate Research Assistant **2019–2024**
University of California, Irvine Irvine, California

Research Scientist Intern **2022**
Samsung Electronics America San Diego, California

R&D Engineer **2016-2018**
YouNet Group HCMC, Vietnam

TEACHING EXPERIENCE

Teaching Assistant **2019-2020**
University of California, Irvine Irvine, California

REFEREED JOURNAL PUBLICATIONS

Deep latent variable joint cognitive modeling of neural signals and human behavior **2024**
NeuroImage

A novel wireless ECG system for prolonged monitoring of multiple zebrafish for heart disease and drug screening studies **2022**
Biosensors and Bioelectronics

Deep learning-based framework for cardiac function assessment in embryonic zebrafish from heart beating videos **2021**
Computers in Biology and Medicine

Continuous non-invasive blood pressure monitoring: a methodological review on measurement techniques **2020**
IEEE Access

An efficient and robust deep learning method with 1-D octave convolution to extract fetal electrocardiogram **2020**
Sensors

REFEREED CONFERENCE PUBLICATIONS

PPG-to-ECG signal translation for continuous atrial fibrillation detection via attention-based deep state-space modeling **2024**
International Conference of the IEEE Engineering in Medicine and Biology Society

Composing graphical models with generative adversarial networks for EEG signal modeling **2022**
IEEE International Conference on Acoustics, Speech and Signal Processing

Decision SincNet: Neurocognitive models of decision making that predict cognitive processes from neural signals **2022**
International Joint Conference on Neural Networks

P2E-WGAN: ECG waveform synthesis from PPG with conditional Wasserstein generative adversarial networks **2021**
ACM Symposium on Applied Computing

STINT: Selective transmission for low-energy physiological monitoring **2020**
ACM/IEEE International Symposium on Low Power Electronics and Design

ABSTRACT OF THE DISSERTATION

Deep Latent Variable Modeling of Physiological Signals

By

Khuong An Vo

Doctor of Philosophy in Computer Science

University of California, Irvine, 2024

Professor Nikil Dutt, Chair

A deep latent variable model is a powerful method for capturing complex distributions. These models assume that underlying structures, but unobserved, are present within the data. In this dissertation, we explore high-dimensional problems related to physiological monitoring using latent variable models. First, we present a novel deep state-space model to generate electrical waveforms of the heart using optically obtained signals as inputs. This can bring about clinical diagnoses of heart disease via simple assessment through wearable devices. Second, we present a brain signal modeling scheme that combines the strengths of probabilistic graphical models and deep adversarial learning. The structured representations can provide interpretability and encode inductive biases to reduce the data complexity of neural oscillations. The efficacy of the learned representations is further studied in epilepsy seizure detection formulated as an unsupervised learning problem. Third, we propose a framework for the joint modeling of physiological measures and behavior. Existing methods to combine multiple sources of brain data provided are limited. Direct analysis of the relationship between different types of physiological measures usually does not involve behavioral data. Our method can identify the unique and shared contributions of brain regions to behavior and can be used to discover new functions of brain regions. The success of these innovative computational methods would allow the translation of biomarker findings across species and provide insight into neurocognitive analysis in numerous biological studies and clinical diagnoses, as well as emerging consumer applications.

Chapter 1

Introduction

1.1 Motivation

The modeling of physiological signals stands as a cornerstone in contemporary neuroscience and cardiology research, providing a powerful framework for understanding the intricate processes underlying the generation of signals in the heart and brain. These models, often constructed using computational methods, offer invaluable insights into the complex interactions among biological variables and serve as invaluable tools for hypothesis testing and experimental design. At the heart of physiological modeling lie equations that describe the interrelations among variables governing the dynamics of synthetic time series. These equations, typically derived from existing data or informed by domain expertise, describe the temporal evolution of physiological phenomena. By systematically altering parameters within these equations, researchers can explore a wide range of scenarios, uncovering novel insights and validating theoretical predictions. Previous studies in neuroscience and cardiology have leveraged computational models (Niederer et al., 2019; Glomb et al., 2020) providing principal laws, empirically validated rules, or other domain expertise, typically

presented as general, time-dependent, and nonlinear partial differential equations.

With regard to electroencephalogram (EEG) signals, neural mass models represent a prominent paradigm for understanding the complex dynamics of large populations of neurons. One such influential model is the Jansen-Rit model (Jansen et al., 1993; Jansen and Rit, 1995). This model emerged as a lumped parameter representation of a cortical column, specifically designed to capture the dynamics of human EEG rhythms and visual evoked potentials. Building upon earlier work by Lopes da Silva and Katznelson, the Jansen-Rit model retains non-linearities within the cortical column, thus offering a more biologically realistic description of neuronal activity. At its core, the Jansen-Rit model is based on the interaction between two distinct populations of neurons within the cortical column: pyramidal cells and local excitatory/inhibitory interneurons. Each population contributes to the generation and modulation of EEG signals through intricate synaptic connections and feedback loops. The model describes the dynamics of postsynaptic potentials using a set of coupled differential equations, which are formulated to capture the complex interplay between excitatory and inhibitory neuronal populations. These equations are typically rewritten as a system of first-order differential equations, resulting in a six-dimensional dynamical system that encapsulates the behavior of the cortical column.

In terms of the electrocardiogram (ECG) signal, the McSharry et al. (2003) model is a significant contribution, providing a framework for understanding the complex dynamics of cardiac rhythms. This model provides a mathematical description of the electrical activity of the heart through a system of three ordinary differential equations. Central to the model is the recognition of the distinct components of the ECG waveform, each of which corresponds to specific cardiac events and functions. The typical sequence begins with the P wave, representing atrial depolarization, followed by the QRS complex, reflecting ventricular depolarization, and concludes with the T wave, indicating ventricular repolarization. By capturing the temporal relationships between these waveform components, the model provides

insights into the underlying physiological processes driving cardiac activity. Moreover, the constructed simulator offers the flexibility to manipulate various attributes of the produced ECG signals. Researchers can adjust parameters such as the interval between waves, the magnitude of P-waves and Q-waves, and the average and standard deviation of heart rate patterns. Additionally, the model allows for the exploration of frequency-domain aspects of heart rate variability, providing insights into the dynamic regulation of cardiac rhythm.

Although electrophysical models based on differential equations have been instrumental in understanding physiological signals, they come with inherent limitations. These models often rely on strong assumptions, can be computationally intensive, and may suffer from model misspecification issues. Consequently, there is a growing emphasis on developing more powerful and flexible models capable of handling diverse types of medical time series data. Medical time series data exhibit complex multidimensional dependencies, including spatio-temporal dependencies in biosignals and multimodal dependencies across physiological measures and behaviors. One of the key challenges is effectively modeling spatio-temporal dependencies in biosignals. Biosignals such as EEG and ECG are inherently dynamic and exhibit spatial variations across different body regions. Another challenge lies in integrating multiple modalities of medical data. In modern healthcare settings, patient data often comprise a diverse array of measurements from different sensors and modalities, including physiological signals, imaging data, and clinical observations. These presents significant challenges for modeling and analysis in biomedical research and clinical practice.

In recent decades, the exponential growth of data collection in various medical applications has paved the way for the emergence of data-driven approaches capable of unlocking valuable insights and addressing complex challenges in healthcare. These approaches leverage advanced computational techniques to analyze large volumes of data, uncover underlying structures, and utilize extracted information for tasks such as predictive modeling and pattern recognition. A significant breakthrough in this domain has been the integration of

probabilistic modeling and deep learning techniques. This fusion of methodologies combines the expressive power of deep neural networks with the probabilistic framework, enabling the parameterization of rich probabilistic distributions over latent variables. By incorporating established or desired inductive biases, these models can effectively capture the complex relationships and uncertainties inherent in medical data. Two prominent classes of models that have propelled recent progress are variational autoencoders (VAEs) and generative adversarial networks (GANs), both belonging to the broader category of deep generative models. These models offer scalable and efficient solutions for unsupervised learning of complex, high-dimensional data distributions.

This thesis aims to explore a diverse category of sequential and multimodal models, with a particular focus on their application to complex spatio-temporal physiological measures. By leveraging these models, we seek to unlock valuable insights from unlabeled datasets, paving the way for a deeper understanding of dynamic physiological processes.

1.2 Outline and contributions

Chapter 2 first presents the pertinent techniques employed in both clinical practice and research for physiological monitoring, offering crucial insights into the body's internal condition. We focus on biosignals obtained through sensors either on or inside the body, like surface ECG and EEG. Subsequently, we introduce probability theory, graphical models, and latent variable models, which underpin all the methods discussed later in this dissertation.

Chapter 3 presents a sequential modeling of ECG signals from photoplethysmography (PPG). PPG is a cost-effective and non-invasive technique that utilizes optical methods to measure cardiac physiology. PPG has become increasingly popular in health monitoring and is used in various commercial and clinical wearable devices. Compared to electrocardiog-

raphy (ECG), PPG does not provide substantial clinical diagnostic value, despite the strong correlation between the two. Here, we propose a subject-independent attention-based deep state-space model (ADSSM) to translate PPG signals to corresponding ECG waveforms. The model is not only robust to noise but also data-efficient by incorporating probabilistic prior knowledge. To evaluate our approach, 55 subjects' data from the MIMIC-III database were used in their original form, and then modified with noise, mimicking real-world scenarios. Our approach was proven effective as evidenced by the PR-AUC of 0.986 achieved when inputting the translated ECG signals into an existing atrial fibrillation (AFib) detector. ADSSM enables the integration of ECG's extensive knowledge base and PPG's continuous measurement for early diagnosis of cardiovascular disease.

Chapter 4 presents a deep generative model of EEG signals that provides not only a stochastic procedure that directly generates data but also insights to further understand the neurological mechanisms. Specifically, we propose a generative and inference approach that combines the complementary benefits of probabilistic graphical models and GANs for EEG signal modeling. We investigate the method's ability to jointly learn coherent generation and inverse inference models on the CHI-MIT epilepsy multi-channel EEG dataset. We further study the efficacy of the learned representations in epilepsy seizure detection formulated as an unsupervised learning problem.

Chapter 5 introduces a method for joint cognitive modeling of neural signals and human behavior. As the field of computational cognitive neuroscience continues to expand and generate new theories, there is a growing need for more advanced methods to test the hypothesis of brain-behavior relationships. Recent progress in Bayesian cognitive modeling has enabled the combination of neural and behavioral models into a single unifying framework. However, these approaches require manual feature extraction, and lack the capability to discover previously unknown neural features in more complex data. Consequently, this would hinder the expressiveness of the models. To address these challenges, we propose a Neurocogni-

tive Variational Autoencoder (NCVA) to conjoin high-dimensional EEG with a cognitive model in both generative and predictive modeling analyses. Importantly, our NCVA enables both the prediction of EEG signals given behavioral data and the estimation of cognitive model parameters from EEG signals. This novel approach can allow for a more comprehensive understanding of the triplet relationship between behavior, brain activity, and cognitive processes.

Chapter 6 concludes the main contributions of this dissertation and discusses some directions for future work.

Chapter 2

Background

2.1 Physiological Signals

2.1.1 Electrocardiogram (ECG or EKG)

Electrocardiography (ECG) is fundamental in diagnosing and managing cardiac health . By recording the heart's electrical activity, the ECG provides crucial insights into its rhythm, rate, and muscular functionality. This non-invasive tool quantifies voltage differences between points on the body's surface over time, enabling clinicians to assess the heart's performance with precision. At its core, an ECG captures the electrical signals generated by the heart with each beat. These signals, represented graphically as waves and complexes on the ECG tracing, reflect the coordinated sequence of events during the cardiac cycle, including atrial depolarization, ventricular depolarization, and ventricular repolarization. Key among its features is the identification of the R-wave, marking the onset of ventricular contraction that propels blood from the heart into the aorta, a pivotal event in the cardiac cycle. In clinical settings, traditional 12-lead ECG devices provide a comprehensive view of the heart's

electrical activity from multiple angles. By delivering 12 concurrent ECG signals, these devices offer invaluable insights into irregularities and pinpoint specific areas of concern within the heart. Recent advancements have seen the integration of single-lead ECG sensors into wrist-worn devices, bringing cardiac monitoring closer to everyday life. These compact gadgets, featuring electrodes on the wrist and an additional point of contact, offer a simplified yet effective means of evaluating heart rhythm and certain functional aspects.

2.1.2 Photoplethysmogram (PPG)

Photoplethysmogram (PPG) signals are obtained through the emission of light from a light-emitting diode (LED) onto the skin, followed by the assessment of the light either reflected back from the skin surface or transmitted through bodily tissues. This fundamental principle underpins its application in a variety of devices, from wrist-worn wearables to medical-grade pulse oximeters. The PPG signal tracks variations in blood volume over time, particularly in arterial blood. As the arterial pulse wave reaches the measurement site, typically at the fingertip or earlobe, it triggers detectable changes in blood volume, generating characteristic waveforms in the PPG signal. Each heartbeat manifests as a distinct peak in the PPG waveform, reflecting the pulsatile nature of blood flow and pressure within the arteries. A major benefit of PPG sensors is their capability to record physiological signals passively, without the need for user involvement, in contrast to wrist-worn ECG sensors which require active user engagement during the collection of signals. This ease of use, coupled with their non-invasive nature, has propelled the widespread adoption of PPG sensors in consumer wearable devices for tracking heart rate, heart rate variability, and cardiac rhythm. However, it is important to recognize that the PPG signal is susceptible to noise, stemming from factors such as motion artifacts, ambient light interference, and variations in skin perfusion.

2.1.3 Electroencephalogram (EEG)

Electroencephalography (EEG) provides invaluable insights into the electrical activity of the brain. By recording the synchronized post-synaptic currents primarily in cortical pyramidal neurons (Nunez and Srinivasan, 2006), EEG offers a unique window into cognitive processes, neural dynamics, and neurological disorders. Over the years, EEG has become a cornerstone in both clinical diagnostics and cognitive research, enabling researchers and clinicians to delve deep into the complexities of brain function. Broadly categorized into spontaneous potentials, such as sleep rhythms, and evoked potentials, which are time-locked responses to external stimuli, EEG captures brain activity with high temporal resolution. Operating on the scale of milliseconds, EEG is capable of detecting rapid changes in neural activity, making it a powerful tool for studying dynamic brain processes. However, despite its temporal precision, EEG possesses inherent limitations in spatial resolution. The electrical signals detected at the scalp originate from currents that propagate through head tissues via volume conduction, resulting in a low spatial resolution. While EEG can provide insights into broad patterns of brain activity, its ability to localize specific neural sources is limited. Nevertheless, EEG has found extensive applications in both clinical and research settings. In clinical practice, EEG is used to diagnose and manage various neurological conditions, including epilepsy, sleep disorders, stroke, and Alzheimer's disease. In the realm of cognitive sciences, EEG offers insights into sensorimotor pathways, memory, language processing, and general intelligence. One of the key advantages of EEG lies in its affordability, portability, and suitability for real-time observation. Unlike other brain imaging techniques that require specialized equipment and expertise, EEG can be deployed with minimal resources, making it accessible to a wide range of researchers and clinicians. However, EEG analysis is not without challenges. EEG signals are non-stationary, exhibit a poor signal-to-noise ratio, and exhibit high variability among individuals, posing significant obstacles to the development of generalized models for EEG analysis.

2.1.4 Frequency Domain Representation

A univariate time series signal of length T consists of a sequence of real-valued data points $\boldsymbol{x} = (x_0, \ldots, x_{T-1}) \in \mathbb{R}^T$, each representing observations of a specific phenomenon. Observations in a time series are generally interdependent, and grasping this dependency is crucial for recognizing various phenomena as they appear. In the frequency domain, a time series is analyzed by decomposing it into sinusoids that vary in amplitude and phase.

Assuming the periodic nature of the time series $\boldsymbol{x} = (x_0, \ldots, x_{T-1})$, we can represent each element x_t with an equation

$$x_t = \sum_{k=0}^{T-1} X_k e^{\frac{j2\pi kt}{T}} \tag{2.1}$$

where $X_k \in \mathbb{C}, k = 0, \ldots, T-1$, represent the Fourier coefficients. Each element x_t in the time series is broken down into T frequency components $e^{j2\pi kt/T}, k = 0, \ldots, T-1$, each scaled by the Fourier coefficients. In the discrete FT (DFT), the time series $\boldsymbol{x}$ is expressed via the Fourier coefficients $X_k = (1/T)\boldsymbol{v}_k^* \boldsymbol{x}, \quad k = 0, \ldots, T-1$, where $\boldsymbol{v}_k := \left[1, e^{j2\pi k(1)/T}, e^{j2\pi k(2)/T}, \ldots, e^{j2\pi k(T-1)/T}\right]$. The collection of vectors $\{\boldsymbol{v}_0, \ldots, \boldsymbol{v}_{T-1}\}$ forms an orthogonal basis for the T-dimensional complex vector space. Each Fourier coefficient thus serves as an independent representation of a subcomponent of the entire time series. The conversion of the time series into the frequency domain can be efficiently performed using the fast FT (FFT) algorithm, while the transformation from the frequency domain back to the time domain is accomplished using the inverse DFT.

2.2 Probabilistic Graphical Models

2.2.1 Random Variables and Probabilities

Definition 1 (Random Variable).

In a random experiment with a sample space $\mathcal{S}$, a function X maps each element $s \in \mathcal{S}$ to a single real number $X(s) = x$. This function X is known as a random variable (r.v.).

Typically, random variables are represented by capital letters, while the values they take are denoted by lowercase letters. The term univariate distribution will be used to describe the distributions of a single random variable, indicated by the non-bold x. The term multivariate distributions will apply to distributions involving multiple random variables, typically represented as a vector with bold $\boldsymbol{x}$.

Definition 2 (Discrete Random Variable and Probability Mass Function).

1. A random variable X is termed discrete if its range consists of countable values.
2. If X is a discrete r.v., the function $P(X = x)$ is called the probability mass function (PMF) of X.
3. The PMF of any discrete r.v. X must adhere to these two criteria:
   - Nonnegativity: $P(X = x) > 0$ if $x = x_i$ for some i, and $P(X = x) = 0$ otherwise.
   - Summation to 1: $\sum_{i=1}^{\infty} P(X = x) = 1$.

Definition 3 (Continuous Random Variable and Probability Density Function).

1. A r.v. X is defined as continuous if there is a function $p(.)$ such that for every real

number x, the cumulative distribution function (CDF) is given by $F_X(x) = P(X \leqslant x) = \int_{-\infty}^{x} p(x)\,dx$.

2. For a continuous r.v. X, the function $p(.)$ in $F_X(x) = \int p(x)\,dx$ is known as the probability density function (PDF).

3. The PDF of any continuous r.v. X must meet the following two criteria:

   - Nonnegativity: $p(x) \geq 0$.
   - Integration to 1: $\int_{-\infty}^{\infty} p(x)\,dx = 1$.

The term "probability distribution" will be applied to both discrete probability mass functions and continuous probability density functions in our discussions. The specific use of the term will be inferred based on the context.

Definition 4 (Exponential Family of Probability Distributions).

1. The exponential family of probability distributions is a set of PDFs or PMF's characterized by the following form:

$$p(\boldsymbol{x} \mid \boldsymbol{\eta}) = h(\boldsymbol{x}) \exp\left\{T(\boldsymbol{x})^T \boldsymbol{\eta} - A(\boldsymbol{\eta})\right\} \tag{2.2}$$

   Here, $\boldsymbol{x}$ represents a specific value of a r.v. $X, T(\boldsymbol{x})$ is the sufficient statistic, and $\boldsymbol{\eta}$ is the natural parameter. The function $A(\boldsymbol{\eta})$ is known as the log-partition function and $h(\boldsymbol{x})$ is the base measure.

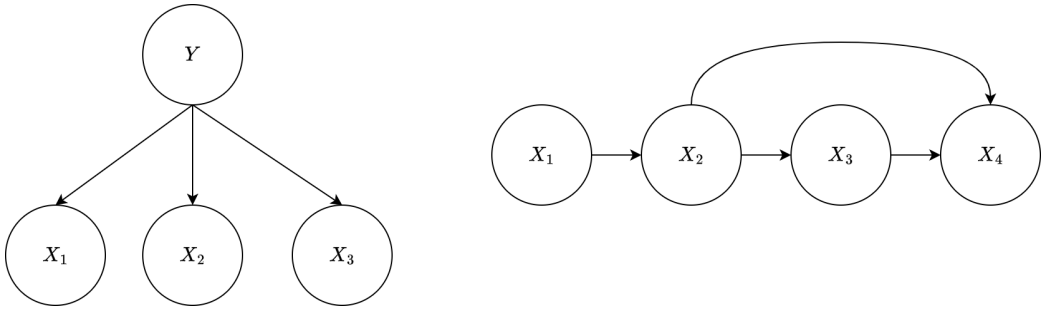
2. A fundamental property of PDFs or PMFs in the exponential family is expressed by:

$$\mathbb{E}[T(\boldsymbol{x})] = \nabla A(\boldsymbol{\eta}) \tag{2.3}$$

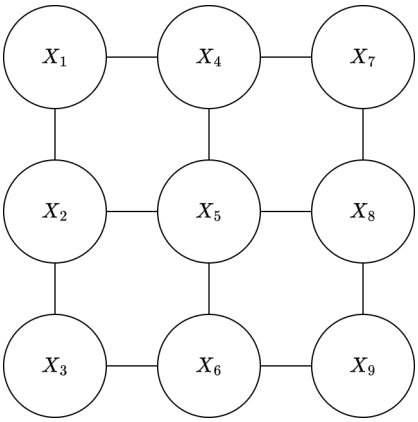
Here, $\nabla A(\boldsymbol{\eta})$ denotes the gradient of $A(\boldsymbol{\eta})$. This property highlights the relationship between the expected value of the sufficient statistic and the gradient of the log-partition function.

The exponential family includes many well-known distributions such as the normal, exponential, Poisson, and gamma distributions, among others (Wainwright et al., 2008). One of the key advantages of distributions in the exponential family is their mathematical tractability, which simplifies parameter estimation, hypothesis testing, and model interpretation.

2.2.2 Graphical Models



(a) Directed graphical models



(b) Undirected graphical models

Figure 2.1: Examples of different probabilistic graphical models.

Probabilistic Graphical Models (PGMs) (Koller and Friedman, 2009) serve as a structured

framework for representing and reasoning about complex probabilistic relationships among multiple variables. One of the key advantages of PGMs lies in their ability to handle uncertainty and variability efficiently. As the number of variables grows, managing probability distributions becomes increasingly complex. In many real-world scenarios, understanding these relationships and making informed decisions based on uncertain data is essential.

At their core, PGMs leverage graphs to visually depict probabilistic relationships between variables. These graphs consist of nodes, which represent random variables, and edges, which denote probabilistic dependencies between variables. Random variables can be classified as either observed, with their values determined by the problem, or latent, with their values remaining unknown.

There are two main types of PGMs: directed graphical models, also known as Bayesian networks, and undirected graphical models, also called Markov Random Fields. Bayesian networks utilize directed edges to represent influential relationships between variables, while Markov Random Fields capture the concept of local interactions among variables through undirected edges.

Representing domain knowledge through graphical structures

One notable aspect of PGMs is the inherent structure of graph representations, which directly implies a factorization of the joint distribution over random variables. In a graphical model, every graph structure entails a specific factorization of the joint distribution. For instance, in a directed graphical model, such as the one shown in Figure 2.1a (right), the joint distribution factors according to conditional probabilities associated with each node:

$$P(X_1, \ldots, X_4) = P(X_1)\, P(X_2 \mid X_1)\, P(X_3 \mid X_2)\, P(X_4 \mid X_2, X_3) \tag{2.4}$$

Similarly, considering the model depicted in Figure 2.1b (left), the joint distribution can be expressed as a product of clique potentials, where each clique represents a set of variables that are directly connected:

$$P\left(X_1, \ldots, X_9\right) = \frac{1}{Z} \prod_{c_i \in \mathcal{C}} \phi_i\left(x_c\right) \tag{2.5}$$

where $\mathcal{C}$ represents all cliques within the graph , specifically every pair of connected nodes, $\phi_i(x_c)$ refers to the clique potentials, which are scalar values given to each possible combination of variables within the clique x_c, and Z is the normalization constant.

This factorization has practical implications, as we need only track parameters associated with each conditional probability or clique potential. By exploiting this structured representation, significant parameter savings can be achieved compared to a full join distribution. Additionally, the development of a graphical model typically requires integration with domain experts. Such integration encourages careful deliberation of the choice and interconnections of random variables in the model. For example, the grid structure observed in Figure 2.1b suggests spatial correlations among variables, akin to those found in image pixels. Furthermore, graphical models enable the utilization of structural properties of the underlying data distribution to simplify the computation of probabilistic queries. One notable advantage is the simplification of independence properties, facilitated by the graph's topology.

Independence properties

By leveraging insights from graph theory and probability theory, graphical models offer a structured approach to probabilistic reasoning. An essential aspect of understanding these relationships lies in studying independence properties among the distributions over random variables within the graphical model:

- *Marginal Independence.* Marginal independence occurs when two random variables are independent of each other, irrespective of the values of other variables. In graphical models, this independence is evident when there is no direct connection (edge) between the variables. For instance, in a directed graphical model, if there is no directed path between two variables, they are considered marginally independent. Mathematically, marginal independence implies that the joint distribution factorizes into the product of marginal distributions for each variable.

- *Conditional Independence.* Conditional independence arises when two random variables become independent given the values of a third set of variables. In graphical models, conditional independence statements reveal when observing certain variables render others independent. This concept is crucial for understanding the influence and interaction between variables. Conditional independence relationships are often inferred from the graph's structure, where paths between variables indicate possible dependencies or influences.

In directed graphical models, a unique type of conditioning exists that causes variables, which are marginally independent, to become dependent. This concept of "explaining away" refers to a phenomenon where the evidence observed for one variable increases or decreases the probability of another variable being responsible for the same observation, depending on their shared dependencies. Consider a simple scenario represented by a DGM where variables A and B capture potential causes of chest pain (variable C): heart attack and indigestion, respectively, as shown in the graph $A \rightarrow C \leftarrow B$. As the doctors gather more information about the patient's symptoms and medical history, they can start to narrow down the potential causes. For example, if the patient also shows symptoms like sweating and shortness of breath, which are commonly associated with heart attacks, this evidence increases the likelihood of a heart attack being the cause of the chest pain. Consequently, the probability of indigestion decreases.

The Markov blanket of a random variable refers to a set of variables in a graphical model that, when conditioned on, renders the random variable independent of all other variables in the graph. Specifically, for any random variables X and Y in the graphical model G, the Markov blanket (MB) of X is defined as the minimal set of variables such that conditioning on this set renders the conditional independence of X and Y:

$$P(X \mid \mathrm{MB}(X), Y) = P(X \mid \mathrm{MB}(X)) \tag{2.6}$$

In undirected graphical models, the Markov blanket of a random variable consists of all its neighboring variables. In directed graphical models, the Markov blanket includes a node's parents, its children, and its children's co-parents.

2.3 Latent Variable Models

In the realm of machine learning, one of the fundamental challenges is to accurately model and understand the underlying probability distributions $p(\boldsymbol{x})$ governing high-dimensional data. High-dimensional data, such as images, text documents, and sensor readings, often exhibit complex structures and dependencies that are not easily captured by simple parametric models. Modeling these complex probability distributions is crucial for various tasks, including generative modeling, anomaly detection, and density estimation. Generative models aim to learn the underlying data distribution and generate new samples that resemble the original data. Anomaly detection algorithms rely on accurate probability estimates to identify deviations from normal behavior. Density estimation techniques seek to estimate the probability density function of the data, enabling various downstream tasks such as sampling and likelihood evaluation.

Introducing an unobserved latent variable $\boldsymbol{z}$ with lower dimensionality than observed vectors

and defining a conditional distribution $p(\boldsymbol{x} \mid \boldsymbol{z})$ for the data is a powerful approach in probabilistic modeling, particularly in scenarios involving high-dimensional data. This framework allows us to capture complex correlations in the observed variable $\boldsymbol{x}$ by leveraging the latent variable $\boldsymbol{z}$, which serves as a compact representation of underlying factors influencing the data. In this framework, the latent variable encodes meaningful information about the structure and content of the observed data. For example, in the context of modeling medical data, $\boldsymbol{z}$ could encapsulate latent representations of various attributes such as disease subtypes or phenotypes, biomarker signatures, and other relevant features. To formalize this probabilistic model, we introduce a prior distribution $p(\boldsymbol{z})$ over the latent variables, representing our beliefs about the likely configurations of $\boldsymbol{z}$. This prior distribution encodes any prior knowledge or assumptions about the latent space. We then compute the joint distribution over observed and latent variables, denoted as $p(\boldsymbol{x}, \boldsymbol{z})$, which describes the probability of observing a particular data point $\boldsymbol{x}$ along with its corresponding latent representation $\boldsymbol{z}$:

$$p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x} \mid \boldsymbol{z})p(\boldsymbol{z}) \tag{2.7}$$

Introducing a latent variable in the model enables us to express the complex marginal distribution $p(\boldsymbol{x})$ as a more tractable joint distribution, which consists of the conditional distribution $p(\boldsymbol{x} \mid \boldsymbol{z})$ and the prior distribution $p(\boldsymbol{z})$. Typically, simpler distributions such as exponential family distributions are used to define the conditional distribution $p(\boldsymbol{x} \mid \boldsymbol{z})$ and the prior distribution $p(\boldsymbol{z})$. Exponential family distributions have desirable properties, including tractable normalization constants, which make them computationally efficient for modeling purposes. Once we have the joint distribution $p(\boldsymbol{x}, \boldsymbol{z})$, we can obtain the desired data distribution $p(\boldsymbol{x})$ by marginalizing over the latent variables:

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{z}) \mathrm{d}\boldsymbol{z} = \int p(\boldsymbol{x} \mid \boldsymbol{z}) p(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} \tag{2.8}$$

By applying Bayes' theorem, we can calculate the posterior distribution $p(\boldsymbol{z} \mid \boldsymbol{x})$ as

$$p(\boldsymbol{z} \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \boldsymbol{z}) p(\boldsymbol{z})}{p(\boldsymbol{x})} \tag{2.9}$$

which allows inference of the latent variable given the observation.

Latent variable models offer a framework to describe the generative process behind the observed data. This process can be interpreted as follows:

1. Sampling Latent Variables: To generate a new data point, we first sample a latent variable $\boldsymbol{z}^{(s)}$ from the prior distribution $p(\boldsymbol{z})$. This latent variable captures unobserved factors or features that influence the generation of the data.

2. Generating Observations: Once we have sampled $\boldsymbol{z}^{(s)}$, we use it to sample a new observation $\boldsymbol{x}^{(s)}$ from the conditional distribution $p(\boldsymbol{x} \mid \boldsymbol{z}^{(s)})$. This conditional distribution captures the relationship between the latent variables and the observed data, allowing us to generate realistic data points.

Latent variable models (LVMs) are particularly effective when the data lie in a manifold, a lower-dimensional structure embedded within the higher-dimensional data space. By capturing the essential characteristics of the data manifold, LVMs can effectively model the underlying data distribution while reducing the dimensionality of the representation. LVMs serve not only as black-box density models but also as interpretable frameworks for incor-

porating prior knowledge about the generative process underlying the data. Probabilistic graphical models, such as Bayesian networks or Markov random fields, provide a principled way to encode dependencies among variables and incorporate domain knowledge into the joint distribution $p(\boldsymbol{x}, \boldsymbol{z})$.

This dissertation is centered on non-linear LVMs, specifically those utilizing deep neural networks, termed Deep Latent Variable Models (DLVMs). DLVMs are adept at modeling complex, high-dimensional data distributions, yet they necessitate approximate inference due to the intractability of the integral in Equation (2.8), which lacks an analytic solution. Subsequent chapters will explore variational auto-encoders (VAEs) and generative adversarial networks (GANs), which combine principles from deep learning and latent variable models to create highly flexible distributions using deep neural networks.

2.3.1 Posterior Inference

In latent variable models, the posterior distribution updates our understanding of the latent variables based on the data observed. It is essential for probabilistic reasoning, facilitating prediction, inference, and the learning of model parameters. To approximate the complex posterior distribution, two main types of methods are utilized, each balancing accuracy with computational efficiency:

1. *Sampling.* Sampling techniques, including Markov Chain Monte Carlo (MCMC) approaches, offer an approximation of the posterior distribution through sample generation. These techniques produce samples from the posterior distribution, enabling the estimation of expectations and the execution of inference via Monte Carlo integration. A notable advantage of sampling techniques is their ability to provide precise outcomes with unlimited computational resources. Nevertheless, they often require significant computational effort and may not efficiently handle large datasets. Moreover, assessing

convergence and verifying the quality of the samples can pose difficulties.

2. *Deterministic Approximation.* Deterministic approximation methods approximate the posterior distribution analytically by employing parametric distribution families or specific factorizations. Techniques such as variational inference and expectation propagation are examples. These approaches are scalable and efficient, thus appropriate for handling large datasets. Nonetheless, they cannot ensure precise outcomes, even with unlimited computational resources, because of the fundamental approximations involved. Despite these constraints, deterministic approximation methods remain popular due to their computational manageability and ability to scale.

Variational Inference

Variational inference (VI) (Jordan et al., 1999) leverages the calculus of variations to approximate the posterior distribution $p(\boldsymbol{z} \mid \boldsymbol{x})$ by finding an approximate distribution $q(\boldsymbol{z})$ that minimizes the Kullback-Leibler (KL) divergence between the variational distribution and the true posterior. The KL divergence between $q(\boldsymbol{z})$ and $p(\boldsymbol{z} \mid \boldsymbol{x})$ is defined as:

$$KL[q(\boldsymbol{z}) \| p(\boldsymbol{z} \mid \boldsymbol{x})] = -\mathbb{E}_{q(\boldsymbol{z})}\left[\log \frac{p(\boldsymbol{z} \mid \boldsymbol{x})}{q(\boldsymbol{z})}\right] \tag{2.10}$$

The goal of variational inference is to find a good approximation $q(\boldsymbol{z})$ that minimizes this KL divergence. However, the intractability of the posterior $p(\boldsymbol{z} \mid \boldsymbol{x})$ makes direct optimization challenging. To address this, we introduce the evidence lower bound (ELBO), denoted as $\mathcal{F}(q)$, which is defined as:

$$-\mathbb{E}_{q(\boldsymbol{z})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})}\right] \tag{2.11}$$

The ELBO serves as a lower bound on the marginal likelihood $\log p(\boldsymbol{x})$. By maximizing the ELBO with respect to $q(\boldsymbol{z})$, we indirectly minimize the KL divergence, as the KL divergence is equal to $\log p(\boldsymbol{x}) - \mathcal{F}(q)$.

In practice, the variational distribution $q(\boldsymbol{z})$ is often constrained to a parametric family (e.g., Gaussian distribution) to make the optimization tractable. The parameters of this distribution are then optimized to maximize the ELBO. This trade-off between flexibility and tractability ensures that the variational approximation $q(\boldsymbol{z})$ is both expressive enough to capture the posterior distribution and computationally feasible to work with.

In essence, VI offers a principled approach to approximate complex posterior distributions, facilitating efficient and scalable inference within probabilistic models. By framing the inference problem as an optimization task, VI inference reduces the complexity of inference to a simpler optimization problem.

Expectation Propagation

Unlike variational inference, which minimizes the KL divergence from a chosen approximation to the true posterior, expectation propagation (EP) (Minka, 2013) minimizes the reverse KL divergence, which is defined as:

$$KL[p(\boldsymbol{z} \mid \boldsymbol{x}) \| q(\boldsymbol{z})] = \mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{z} \mid \boldsymbol{x})}{q(\boldsymbol{z})}\right] \tag{2.12}$$

It is important to note that EP does not necessarily minimize the KL divergence but can be freely implemented with any divergence measure. When $q(\boldsymbol{z})$ belongs to the exponential family, which is commonly the case, the minimization of the reverse KL divergence simplifies to the alignment of natural parameters. In EP, the goal is to approximate the posterior distribution $p(\boldsymbol{z})$ with a set of factors $Q(\boldsymbol{z})$ such that each factor $q_i(\boldsymbol{z})$ approximates the corresponding factor $f_i(\boldsymbol{z})$ in the exact posterior factorization. This approximation is achieved by minimizing a sequence of local Kullback-Leibler (KL) divergences between the exact factors and the EP approximating factors.

The EP approximation for $p(\boldsymbol{z})$ is given by:

$$Q(\boldsymbol{z}) = \prod_{i=0}^{n} q_i(\boldsymbol{z}) \tag{2.13}$$

where $q_i(\boldsymbol{z})$ represents the EP approximating factors and $Q(\boldsymbol{z})$ is the global approximation.

To find the parameters for determining the approximate factors $q_i(\boldsymbol{z})$, EP minimizes a sequence of local KL divergences:

$$\begin{aligned}
q_0(\boldsymbol{z}) &= \underset{q_0(\boldsymbol{z})\in\mathcal{Q}}{\arg\min} KL\left(\tilde{f}_0(\boldsymbol{z}) \| q_0(\boldsymbol{z}) Q^{\backslash 0}(\boldsymbol{z})\right) \\
q_1(\boldsymbol{z}) &= \underset{q_1(\boldsymbol{z})\in\mathcal{Q}}{\arg\min} KL\left(\tilde{f}_1(\boldsymbol{z}) \| q_1(\boldsymbol{z}) Q^{\backslash 1}(\boldsymbol{z})\right) \\
&\vdots \\
q_n(\boldsymbol{z}) &= \underset{q_n(\boldsymbol{z})\in\mathcal{Q}}{\arg\min} KL\left(\tilde{f}_n(\boldsymbol{z}) \| q_n(\boldsymbol{z}) Q^{\backslash n}(\boldsymbol{z})\right)
\end{aligned} \tag{2.14}$$

where $\mathcal{Q}$ denotes the space of possible approximate factors, $\tilde{f}_i(\boldsymbol{z}) = f_i(\boldsymbol{z}) Q^{\backslash i}(\boldsymbol{z})$ is the tilted

distribution, and $Q^{\setminus i}(\boldsymbol{z})$ is the cavity distribution obtained by removing the current KL minimizer $q_i(\boldsymbol{z})$ from $Q(\boldsymbol{z})$.

Each KL divergence minimization problem in the above equations is solved by exclusively optimizing $q_i(\boldsymbol{z})$ instead of the EP global approximation. Therefore, expectation propagation is considered a local approximation algorithm as each KL divergence is minimized locally with respect to a selected EP approximating factor.

In EP, convergence is not guaranteed because the local KL divergence minimization does not necessarily ensure that the KL divergence is minimized from the exact posterior distribution to the EP global posterior approximation. Despite the absence of formal convergence guarantees, EP remains a widely used and effective approximate inference algorithm in probabilistic modeling due to its flexibility and applicability to a variety of models.

2.3.2 Parameter Learning

In parameter learning for latent variable models, we aim to estimate the optimal parameters $\theta^\star$ of the model given a training set comprising N data points $\{\boldsymbol{x}^i\}_{i=1}^N$. The likelihood $p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$ and the prior $p_\theta(\boldsymbol{z})$ are assumed to belong to families of distributions parameterized by unknown parameters θ.

The optimal parameters $\theta^\star$ can be learned using Maximum Likelihood Estimation (MLE), which involves maximizing the log-likelihood function $\mathcal{L}(\theta)$:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log p_\theta(\boldsymbol{x}^i) \tag{2.15}$$

Given that the latent variable $\boldsymbol{z}^i$ is different for each data point $\boldsymbol{x}^i$, but the parameters θ

are shared across all data points, we can express the log-likelihood as the sum of individual terms $\mathcal{L}_i(\theta)$:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \underbrace{\log \int p_\theta(\boldsymbol{x}^i, \boldsymbol{z}^i) \mathrm{d}\boldsymbol{z}^i}_{\mathcal{L}_i(\theta)} \tag{2.16}$$

In practice, the marginal density of the observations $p_\theta(\boldsymbol{x})$ is often intractable and needs to be approximated. One approach is to use the evidence lower bound (ELBO), denoted as $\mathcal{F}_i(\theta, q)$, which provides a lower bound to $\log p_\theta(\boldsymbol{x})$ for any distribution $q(\boldsymbol{z})$ over the latent variables:

$$\mathcal{L}_i(\theta) \geq \mathcal{F}_i(\theta, q) = \mathbb{E}_{q(\boldsymbol{z})} \left[\log \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} \right] \tag{2.17}$$

By maximizing the total ELBO $\mathcal{F}(\theta, q) = \sum_{i=1}^{N} \mathcal{F}_i(\theta, q)$ with respect to θ and $q(\boldsymbol{z})$, we can learn the parameters of the model. The variational distribution $q(\boldsymbol{z})$ can be interpreted as an approximation to the posterior distribution $p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$, and the ELBO coincides with the log-likelihood only when $q(\boldsymbol{z})$ equals the true posterior distribution.

In practice, the variational distribution $q(\boldsymbol{z})$ is often constrained to a particular parametric family to make the optimization tractable, and the parameters of this distribution are optimized along with the parameters θ of the model. Therefore, by maximizing the ELBO, we indirectly maximize the log-likelihood, enabling parameter learning in latent variable models.

Expectation Maximization

The Expectation Maximization (EM) algorithm (Dempster et al., 1977) provides a systematic approach for maximizing the likelihood function in models with latent variables. It iteratively alternates between two steps: the E-step, where the posterior distribution over latent variables is estimated, and the M-step, where the model parameters are updated based on the estimated posteriors.

Starting with initial parameters θ_0, the EM algorithm iterates until convergence as follows:

1. E-step (Expectation): Given the current parameters θ_k, estimate the posterior distribution over latent variables, denoted as $q_{k+1}(\boldsymbol{z})$, by maximizing the ELBO $\mathcal{F}_i(\theta_k, q)$ with respect to $q(\boldsymbol{z})$. In many cases, this step involves solving a posterior inference problem, aiming to find an approximation to the true posterior distribution $p_{\theta_k}(\boldsymbol{z} \mid \boldsymbol{x})$. If the posterior is intractable, approximate inference methods can be employed.

2. M-step (Maximization): Fixing the estimated distribution over latent variables $q_{k+1}(\boldsymbol{z})$, update the parameters θ_{k+1} by maximizing the ELBO $\mathcal{F}_i(\theta, q_{k+1})$ with respect to θ. This step involves optimizing the model parameters using techniques such as gradient ascent.

For simpler classes of models where exact inference is possible, each EM iteration guarantees not to decrease the marginal likelihood after each combined step. Specifically, after the E-step, where θ_k is held fixed, the ELBO equals the log-likelihood. Subsequently, maximizing the ELBO in the M-step does not decrease the log-likelihood.

In summary, the EM algorithm offers a systematic and effective approach for optimizing model parameters when latent variables are involved. By iteratively refining the estimates of the latent variables and updating the model parameters, EM enables efficient learning in latent variable models even when exact inference is not feasible.

Chapter 3

PPG-to-ECG Signal Translation For Continuous Atrial Fibrillation Detection via Attention-based Deep State-Space Modeling

3.1 Introduction

The measurement of the electrical activity generated by an individual's heart, known as an electrocardiogram (ECG), typically requires the placement of several electrodes on the body. ECG is considered the preferred method for monitoring vital signs and for the diagnosis, management, and prevention of cardiovascular diseases (CVDs), which are a leading cause of death globally, accounting for approximately 32% of all deaths in 2017 according to Global Burden of Disease reports (Allen, 2007). It has also been demonstrated that sudden cardiac arrests are becoming more prevalent in young individuals, including athletes (sudden, 2020).

Regular ECG monitoring has been found to be beneficial for the early identification of CVDs (Rosiek and Leksowski, 2016). Among heart diseases, atrial fibrillation (AFib) is adults' most common rhythm disorder. Identifying AFib at an early stage is crucial for the primary and secondary prevention of cardioembolic stroke, as it is the leading risk factor for this type of stroke (Olier et al., 2021). Advancements in electronics, wearable technologies, and machine learning have made it possible to record ECGs more easily and accurately, and to analyze large amounts of data more efficiently. Despite these developments, there are still challenges associated with continuously collecting high-quality ECG data over an extended period, particularly in everyday life situations. The 12-lead ECG, considered the clinical gold standard, and simpler versions, such as the Holter ECG, can be inconvenient and bulky due to the need to place multiple electrodes on the body, which can cause discomfort. Additionally, the signals may degrade over time as the impedance between the skin and electrodes changes. Consumer-grade products such as smartwatches have developed solutions to address these issues. However, these products require users to place their fingers on the watch to form a closed circuit, making continuous monitoring impossible.

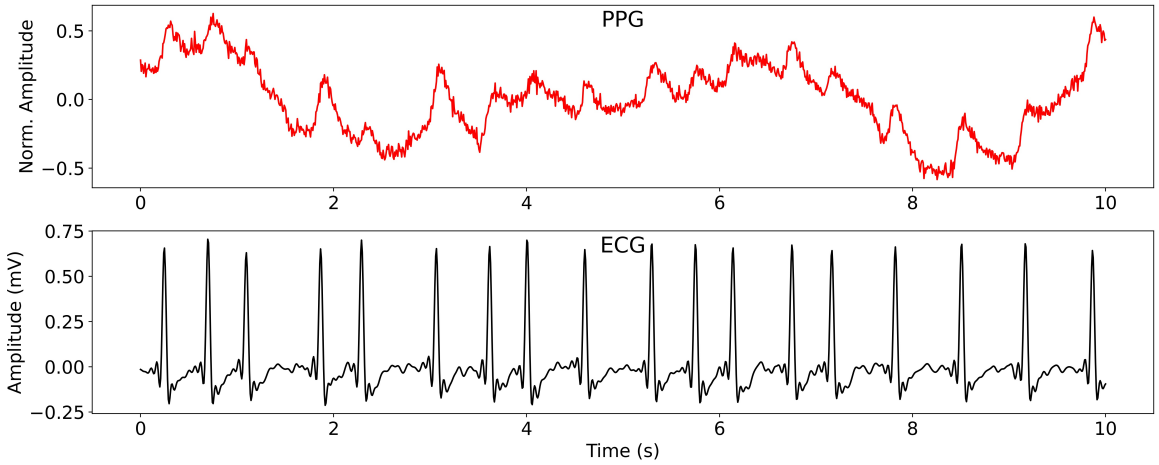


Figure 3.1: A PPG-ECG waveform pair. PPG signals can often become contaminated by noise.

One potential solution to these issues is to use a mathematical method to derive ECG data from an alternative, highly correlated, non-invasive signal, such as the photoplethysmogram

(PPG), which can be easily acquired using various wearable devices, including smartwatches. PPG is more convenient, cost-effective, and user-friendly. PPG has been increasingly adopted in consumer-grade devices. This technique involves the use of a light source, usually an LED, and a photodetector to measure the changes in light absorption or reflection as blood flows through the tissue. ECG and PPG signals are inherently correlated as both are influenced by the same underlying cardiac activity, namely the depolarization and repolarization of the heart. These contractions lead to changes in peripheral blood volume, which are measured by PPG. Figure 3.1 shows the relationship between ECG and PPG waveforms. Although there are established standards for interpreting ECG for clinical diagnosis, the use of PPG is still mostly limited to measuring heart rate and oxygen saturation (Reisner et al., 2008). By translating PPG to ECG signals, clinical diagnoses of cardiac diseases and anomalies could be made in real-time.

Few research works attempted to synthesize ECG from PPG signals. In (Banerjee et al., 2014), a machine learning-based approach was proposed to estimate the ECG parameters, including the RR, PR, QRS, and QT intervals, using features from the time and frequency domain extracted from a fingertip PPG signal. Additionally, Zhu et al. (2019); Tian et al. (2020) proposed models to reconstruct the entire ECG signal from PPG in the frequency domain. However, the performance of these approaches relied on cumbersome algorithms for feature crafting. With recent advances in deep learning, Vo et al. (2021); Sarkar and Etemad (2021); Chiu et al. (2020) leveraged the expressiveness and structural flexibility of neural networks to build end-to-end PPG-to-ECG algorithms. However, the models suffer from data-hungry problems as they do not explicitly model the underlying sequential structures of the data. In addition, complex deep learning models cannot run efficiently on resource-constrained devices (e.g., wearables) due to their high computational intensity, which poses a critical challenge for real-world deployment (Lee et al., 2020). Furthermore, deterministic models face difficulties in effectively generalizing to noisy data.

To address these challenges, we propose a deep probabilistic model to accurately estimate ECG waveforms from raw PPG. The contributions of this work are three-fold:

- We present a deep generative model incorporating prior knowledge about the data structures that enable learning on small datasets. Specifically, we develop a deep latent state-space model augmented by an attention mechanism.
- The probabilistic nature of the model enhances its robustness to noise. We demonstrate this by evaluating the model on data corrupted with Gaussian and baseline wandering noise, replicating real-life situations.
- Our method is effective not only in healthy subjects but also in subjects with AFib. It is orthogonal and complementary to existing AFib detection methods (Hong et al., 2020) by simply providing the translated ECG to any pre-trained models. This would enhance the performance of existing models by enabling uninterrupted monitoring, thereby facilitating the early detection of cardiovascular disease.

3.2 Methodology

3.2.1 Probabilistic Modeling of ECG from PPG signals

We are given a dataset $\mathcal{D} := \left\{(\boldsymbol{x}^1, \boldsymbol{y}^1), \ldots, \left(\boldsymbol{x}^N, \boldsymbol{y}^N\right)\right\}$ with the i-th observation $\boldsymbol{y}^i \in \mathbb{R}^{n_y}$, i.e., ECG signals of n_y time samples, depending on $\boldsymbol{x}^i \in \mathbb{R}^{n_x}$, i.e., PPG signals of n_x time samples. Throughout the paper, superscript i is omitted when we refer to only one sequence or when it is clear from the context.

We aim to learn a generative process with a latent-variable model comprising of a parametric non-linear Gaussian prior over latents $p_{\theta_z}(\boldsymbol{z} \mid \boldsymbol{x})$ and likelihood $p_{\theta_y}(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{x})$. The learning

process minimizes a divergence between the true data-generating distribution and the model w.r.t θ:

$$\begin{aligned} &\arg\min_{\theta} KL\left(p_{\mathcal{D}}(\boldsymbol{y} \mid \boldsymbol{x}) \| p_{\theta}(\boldsymbol{y} \mid \boldsymbol{x})\right) \\ &= \arg\max_{\theta} \mathbb{E}_{p_{\mathcal{D}}(\boldsymbol{y}|\boldsymbol{x})}\left[\log p_{\theta}(\boldsymbol{y} \mid \boldsymbol{x})\right] \end{aligned} \tag{3.1}$$

where $p_{\theta}(\boldsymbol{y} \mid \boldsymbol{x}) = \int p_{\theta_y}(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{x}) \, p_{\theta_z}(\boldsymbol{z} \mid \boldsymbol{x}) \, d\boldsymbol{z}$ is the conditional likelihood/evidence of data point $\boldsymbol{y}$ given condition $\boldsymbol{x}$, approximated by averaging over the latent $\boldsymbol{z}$.

Nevertheless, estimating $p_{\theta}(\boldsymbol{y} \mid \boldsymbol{x})$ is typically intractable. This issue can be mitigated by introducing a parametric inference model $q_{\phi}(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y})$ to construct a conditional variational evidence lower bound on the conditional log-likelihood $\log p_{\theta}(\boldsymbol{y} \mid \boldsymbol{x})$ as follows

$$\begin{aligned} &\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}; \theta, \phi) \\ &\triangleq \log p_{\theta}(\boldsymbol{y} \mid \boldsymbol{x}) - KL\left(q_{\phi}(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y}) \| p_{\theta}(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y})\right) \\ &= \mathbb{E}_{q_{\phi}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})}\left[\log p_{\theta_y}(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{x})\right] - KL\left(q_{\phi}(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y}) \| p_{\theta_z}(\boldsymbol{z} \mid \boldsymbol{x})\right) \end{aligned} \tag{3.2}$$

Taking the likelihood model $p_{\theta_y}(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{x})$ to be a decoder, the latent inference model $q_{\phi}(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y})$ to be an encoder, and the prior model $p_{\theta_z}(\boldsymbol{z} \mid \boldsymbol{x})$, a conditional variational autoencoder (CVAE) (Kingma and Welling, 2013; Sohn et al., 2015) considers this objective from a deep probabilistic autoencoder perspective. Here θ and ϕ are neural network parameters, and learning takes place via stochastic gradient ascent using unbiased estimates of $\nabla_{\theta,\phi} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(\boldsymbol{x}^i, \boldsymbol{y}^i; \theta_z, \theta_y, \phi\right)$.

3.2.2 State-Space Modeling of ECG from PPG Signals

In the previous section, we consider the networks that process the entire time series as a whole, which do not explicitly model the underlying sequential natures of the data. This may

lead to resource-inefficient learning. Here, propose to address the problems by leveraging the *quasi-periodic nature* of the physiological signals.

ECG Generative (Decoding) Process from PPG

We consider nonlinear dynamical systems with observations $\boldsymbol{y}_t \in \mathbb{R}^{n_{rr}}$, i.e., RR intervals or the time elapsed between two successive R peaks on the ECG, depending on control inputs $\boldsymbol{x}_t \in \mathbb{R}^{n_{pp}}$, i.e., PP intervals or the time elapsed between two successive systolic peaks on the PPG. We choose the peaks to segment the signals as they are the most robust features. Corresponding discrete-time sequences of length T are denoted as $\boldsymbol{y}_{1:T} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T)$ and $\boldsymbol{x}_{1:T} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)$.

Given an input PPG $\boldsymbol{x}_{1:T}$, we are interested in a probabilistic model $p(\boldsymbol{y}_{1:T} \mid \boldsymbol{x}_{1:T})$. Formally, we consider

$$p(\boldsymbol{y}_{1:T} \mid \boldsymbol{x}_{1:T}) = \int p(\boldsymbol{y}_{1:T} \mid \boldsymbol{z}_{1:T}, \boldsymbol{x}_{1:T})\, p(\boldsymbol{z}_{1:T} \mid \boldsymbol{x}_{1:T})\, \mathrm{d}\boldsymbol{z}_{1:T} \tag{3.3}$$

where $\boldsymbol{z}_{1:T}$ represents the latent sequence associated with the given model. This implies that we are considering a generative model that incorporates a latent dynamical system with an emission model $p(\boldsymbol{y}_{1:T} \mid \boldsymbol{z}_{1:T}, \boldsymbol{x}_{1:T})$ and transition model $p(\boldsymbol{z}_{1:T} \mid \boldsymbol{x}_{1:T})$.

To derive state-space models, we make certain assumptions regarding the state transition and emission models, as shown in Figure 3.2:

$$p(\boldsymbol{z}_{1:T} \mid \boldsymbol{x}_{1:T}) = \prod_{t=0}^{T-1} p(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t, \boldsymbol{x}_{1:T}) \tag{3.4}$$

$$p(\boldsymbol{y}_{1:T} \mid \boldsymbol{z}_{1:T}, \boldsymbol{x}_{1:T}) = \prod_{t=1}^{T} p(\boldsymbol{y}_t \mid \boldsymbol{z}_t) \tag{3.5}$$

Equations 3.4 and 3.5 make the assumption that the current state $\boldsymbol{z}_t$ includes all the relevant

information about both the current observation $\boldsymbol{y}_t$ and the next state $\boldsymbol{z}_{t+1}$, given the current control input $\boldsymbol{x}_t$.

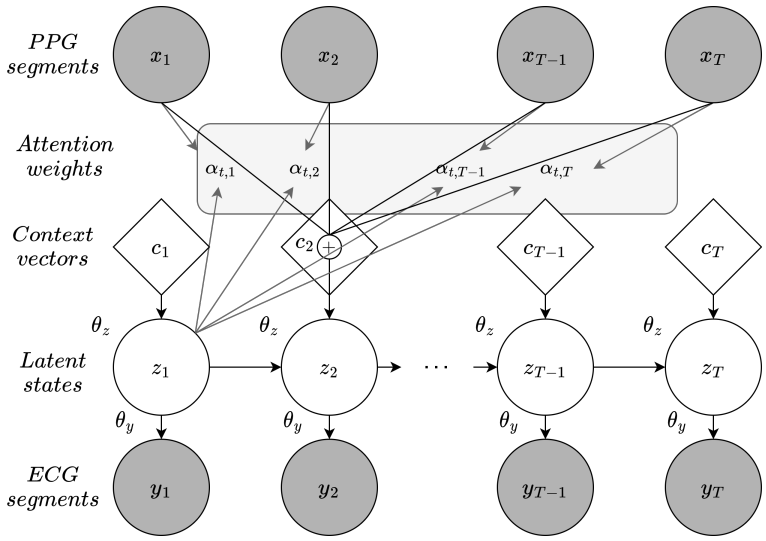


Figure 3.2: The graphical model for ECG translation from PPG. Shaded nodes represent observed variables. Clear nodes represent latent variables. Diamond nodes denote deterministic variables. Variables $\boldsymbol{x}_t, \boldsymbol{y}_t$, and $\boldsymbol{c}_t$ represent PP intervals, RR intervals, and context vectors, respectively. $\alpha_{t,i}$ are attention weights defines how well two intervals $\boldsymbol{x}_i$ and $\boldsymbol{y}_t$ are aligned. The attention mechanism is shown only at time step 2.

In contrast to the DKF model of (Krishnan et al., 2015, 2017), our model takes into account the entire input signal $\boldsymbol{x}_{1:T}$ for each output $\boldsymbol{y}_t$ via an attention mechanism (Bahdanau et al., 2014). Note that there are usually misalignments between the PPG and ECG cycles. Therefore, it is difficult to construct optimal and exact sample pairs. This attention mechanism not only helps to add more context to generate ECG segments, but also helps to address the problem of misalignment.

Let us define $\boldsymbol{c}_t$ a sum of features of the input sequence (PP intervals), weighted by the alignment scores:

$$\boldsymbol{c}_t = \sum_{i=1}^{T} \alpha_{t,i} \boldsymbol{x}_i \tag{3.6}$$

$$\alpha_{t,i} = \frac{\exp\left(\boldsymbol{s}\left(\boldsymbol{z}_{t-1}, \boldsymbol{x}_i\right)\right)}{\sum_{i'=1}^{n} \exp\left(\boldsymbol{s}\left(\boldsymbol{z}_{t-1}, \boldsymbol{x}_{i'}\right)\right)} \tag{3.7}$$

The alignment function $\boldsymbol{s}$ assigns a score $\alpha_{t,i}$ to the pair of input at position i and output at position t, $(\boldsymbol{x}_i, \boldsymbol{y}_t)$, based on how well they match. The set of $\alpha_{t,i}$ are weights defining how much of each source segment should be considered for each output interval.

Both state transition (prior) and emission models are non-linear Gaussian transformations parametrized by neural networks θ_z and θ_y :

$$p_{\theta_z}(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t, \boldsymbol{x}_{1:T}) = \mathcal{N}(\boldsymbol{z}_{t+1} \mid \boldsymbol{\mu}_{\theta_z}(\boldsymbol{z}_t, \boldsymbol{c}_{t+1}), \boldsymbol{\sigma}^2_{\theta_z}(\boldsymbol{z}_t, \boldsymbol{c}_{t+1})); \tag{3.8}$$

$$p_{\theta_y}(\boldsymbol{y}_t \mid \boldsymbol{z}_t) = \mathcal{N}(\boldsymbol{y}_t \mid \boldsymbol{\mu}_{\theta_y}(\boldsymbol{z}_t), \boldsymbol{I}) \tag{3.9}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ are the means and diagonal covariance matrices of the normal distributions $\mathcal{N}$, $\boldsymbol{I}$ is the identity covariance matrix.

Latent State Inference (Posterior Encoding) Process

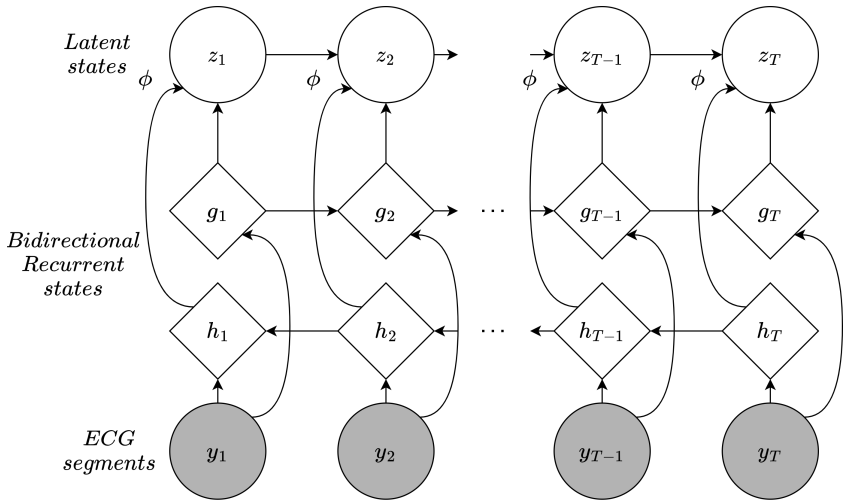


Figure 3.3: The graphical model at latent state inference time. Variables $\boldsymbol{y}_t, \boldsymbol{h}_t, \boldsymbol{g}_t$, and $\boldsymbol{z}_t$ represent respectively RR intervals, backward, forward recurrent states, and latent states.

Unlike a deterministic translation model, the process needs to find meaningful probabilistic embeddings of ECG segments in the latent space. We want to identify the structure of the parametetrized posterior distribution $q_\phi(\boldsymbol{z}_{1:T} \mid \boldsymbol{y}_{1:T})$. Notice that we made a design choice

to perform inference using only $\boldsymbol{y}_{1:T}$. We chose this with the conditional independence assumption that PPG segments do not provide more information than ECG segments alone. The graphical model in Figure 3.2 shows that the $\boldsymbol{z}_t$ node blocks all information coming from the past and flowing to $\boldsymbol{z}_{t+1}$ (i.e., $\boldsymbol{z}_{1:t-1}$ and $\boldsymbol{y}_{1:t}$), leading to the following structure as in Figure 3.3:

$$q_\phi\left(\boldsymbol{z}_{1:T} \mid \boldsymbol{y}_{1:T}\right) = q_\phi\left(\boldsymbol{z}_1 \mid \boldsymbol{y}_{1:T}\right) \prod_{t=1}^{T-1} q_\phi\left(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t, \boldsymbol{y}_{t+1:T}\right) \tag{3.10}$$

where

$$q_\phi(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t, \boldsymbol{y}_{t+1:T}) = \mathcal{N}(\boldsymbol{z}_{t+1} \mid \boldsymbol{\mu}_\phi(\boldsymbol{z}_t, \boldsymbol{y}_{t+1:T}), \boldsymbol{\sigma}^2_\phi(\boldsymbol{z}_t, \boldsymbol{y}_{t+1:T})) \tag{3.11}$$

Training Process

The objective function becomes a timestep-wise conditional variational lower bound (Kingma and Welling, 2013; Sohn et al., 2015; Krishnan et al., 2017):

$$\begin{aligned}
&\log p_\theta(\boldsymbol{y} \mid \boldsymbol{x}) \geq \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}; \theta_y, \theta_z, \phi) \triangleq \\
&\sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_\phi(\boldsymbol{z}_t \mid \boldsymbol{y}_{t:T})} [\overbrace{\log \underbrace{p_{\theta_y}\left(\boldsymbol{y}_t \mid \boldsymbol{z}_t\right)}_{\text{emission model}}}^{\text{reconstruction}}] \\
&- \beta \overbrace{KL\left(q_\phi\left(\boldsymbol{z}_1 \mid \boldsymbol{y}_{1:T}\right) \| p_{\theta_z}\left(\boldsymbol{z}_1 \mid \boldsymbol{x}_{1:T}\right)\right)}^{\text{regularization}} \\
&- \beta \sum_{t=1}^{T-1} \mathop{\mathbb{E}}_{q_\phi(\boldsymbol{z}_t \mid \boldsymbol{y}_{t:T})} [\overbrace{KL(\underbrace{q_\phi\left(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t, \boldsymbol{y}_{t:T}\right)}_{\text{posterior inference model}} || \underbrace{p_{\theta_z}\left(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t, \boldsymbol{x}_{1:T}\right)}_{\text{prior transition model}})}^{\text{regularization}}]
\end{aligned} \tag{3.12}$$

where β controls the regularization strength. During training, the Kullback–Leibler (KL) losses in the regularization terms "pull" the posterior distributions (which encode EEG segments) and the prior distributions (which embed PPG segments) towards each other. We

learn the generative and inference models jointly by maximizing the conditional variational lower bound with respect to their parameters.

3.2.3 Neural Network Architectures

Let us denote $\boldsymbol{W}$, $\boldsymbol{v}$, and $\boldsymbol{b}$ the weight matrices.

Score Model: The alignment score α in Equation 3.7 is parametrized by a feedforward network with a single hidden layer, and this network is jointly trained with other parts of the model. The score function $\boldsymbol{s}$ is in the following form:

$$\boldsymbol{s}\left(\boldsymbol{z}_{t-1}, \boldsymbol{x}_i\right) = \boldsymbol{v}_s^\top \tanh\left(\boldsymbol{W}_s\left[\boldsymbol{z}_{t-1}; \boldsymbol{W}_x \boldsymbol{x}_i\right] + \boldsymbol{b}_s\right) \tag{3.13}$$

Prior Transition Model: We parametrize the transition function in Equation 3.8 from z_t to z_{t+1} using a Gated Transition Function as in (Krishnan et al., 2017). The model is flexible in choosing a non-linear transition for some dimensions while having linear transitions for others. The function is parametrized as follows:

$$\begin{aligned}
\boldsymbol{g}_t &= \text{sigmoid}(\boldsymbol{W}_{g_3}\,\text{ReLU}\left(\boldsymbol{W}_{g_2}\,\text{ReLU}\left(\boldsymbol{W}_{g_1}\left[\boldsymbol{z}_t; \boldsymbol{c}_{t+1}\right] + \boldsymbol{b}_{g_1}\right) + \boldsymbol{b}_{g_2}\right) + \boldsymbol{b}_{g_3}\right) \\
\boldsymbol{d}_t &= \boldsymbol{W}_{d_3}\,\text{ReLU}\left(\boldsymbol{W}_{d_2}\,\text{ReLU}\left(\boldsymbol{W}_{d_1}\left[\boldsymbol{z}_t; \boldsymbol{c}_{t+1}\right] + \boldsymbol{b}_{d_1}\right) + \boldsymbol{b}_{d_2}\right) + \boldsymbol{b}_{d_3} \\
\boldsymbol{\mu}_{\theta_z}(\boldsymbol{z}_t, \boldsymbol{c}_{t+1}) &= (1 - \boldsymbol{g}_t) \odot \left(\boldsymbol{W}_{\mu_z}\left[\boldsymbol{z}_t; \boldsymbol{c}_{t+1}\right] + \boldsymbol{b}_{\mu_z}\right) + \boldsymbol{g}_t \odot \boldsymbol{d}_t \\
\boldsymbol{\sigma}^2_{\theta_z}(\boldsymbol{z}_t, \boldsymbol{c}_{t+1}) &= \text{softplus}\left(\boldsymbol{W}_{\sigma_z^2}\,\text{ReLU}\left(\boldsymbol{d}_t\right) + \boldsymbol{b}_{\sigma_z^2}\right)
\end{aligned} \tag{3.14}$$

where $\mathbb{I}$ denotes the identity function, and $\odot$ denotes element-wise multiplication.

Emission Model: We parameterize the emission function in Equation 3.9 using a two-hidden layer network as:

$$\boldsymbol{\mu}_{\theta_y}\left(\boldsymbol{z}_t\right) = \boldsymbol{W}_{e_3}\,\text{ReLU}\left(\boldsymbol{W}_{e_2}\,\text{ReLU}\left(\boldsymbol{W}_{e_1}\boldsymbol{z}_t + \boldsymbol{b}_{e_1}\right) + \boldsymbol{b}_{e_2}\right) + \boldsymbol{b}_{e_3} \tag{3.15}$$

Posterior Inference Model: We use a Bi-directional Gated Recurrent Unit network (Chung et al., 2014) (GRU) to process the sequential order of RR intervals backward from $\boldsymbol{y}_T$ to $\boldsymbol{y}_{t+1}$ and forward from $\boldsymbol{y}_{t+1}$ to $\boldsymbol{y}_T$. The GRUs are denoted here as $\boldsymbol{h}_t = \text{GRU}\left(\boldsymbol{W}_y\boldsymbol{y}_T, \ldots, \boldsymbol{W}_y\boldsymbol{y}_{t+1}\right)$ and $\boldsymbol{g}_t = \text{GRU}\left(\boldsymbol{W}_y\boldsymbol{y}_{t+1}, \ldots, \boldsymbol{W}_y\boldsymbol{y}_T\right)$, respectively. The hidden states of the GRUs parametrize the variational distribution, which are combined with the previous latent states for the inference in Equation 3.11 as follows:

$$\begin{aligned}
\tilde{\boldsymbol{h}}_t &= \frac{1}{3}\left(\tanh\left(\boldsymbol{W}_h\boldsymbol{z}_t + \boldsymbol{b}_h\right) + \boldsymbol{h}_t + \boldsymbol{g}_t\right) \\
\boldsymbol{\mu}_\phi(\boldsymbol{z}_t, \boldsymbol{y}_{t+1:T}) &= \boldsymbol{W}_\mu\tilde{\boldsymbol{h}}_t + \boldsymbol{b}_\mu \\
\boldsymbol{\sigma}^2_\phi(\boldsymbol{z}_t, \boldsymbol{y}_{t+1:T}) &= \text{softplus}\left(\boldsymbol{W}_{\sigma^2}\tilde{\boldsymbol{h}}_t + \boldsymbol{b}_{\sigma^2}\right)
\end{aligned} \tag{3.16}$$

All the hidden layer sizes are 256, and the latent space sizes are 128. Input and output segments at each timestep are of size 90. We use Adam (Kingma and Ba, 2014) for optimization, with a learning rate of 0.0008, exponential decay rates $\beta_1 = 0.9$, and $\beta_2 = 0.999$. We train the models for 5000 epochs, with a minibatch size 128. We set the regularization hyperparameter $\beta = 0$ at the beginning of training and gradually increase it until $\beta = 1$ is reached at epoch 1250.

3.3 Experiments

3.3.1 Dataset

The MIMIC-III Waveform Database Matched Subset (Moody et al., 2020; Johnson et al., 2016a) was used for the experiments. The database contains recordings collected from patients at various hospitals. Each session has multiple physiological signals, including PPG and ECG signals, sampled at a frequency of 125 Hz. We used the records of 43 healthy sub-

jects and 12 subjects having AFib, including 30 males and 25 females, 23-84 years old. The dataset is made publicly available [1]. Each record duration is 5 minutes. The first 48 s of each record were used as the training set, the next 12 s as the validation set, and the remaining 228 s as the test set. The preprocessing steps, including filtering, alignment, and normalization, were performed as described in (Tang et al., 2022). We applied HeartPy (Van Gent et al., 2019; van Gent et al., 2019) to identify peaks in PPG signals. Each long signal is split into 4-s chunks. Each peak-to-peak interval was linearly interpolated to a length of 90 during training, which is the mean length of the intervals in the training set. The original interval length information can be preserved by making it an additional feature along with each normalized interval. Alternatively, we can apply padding instead of interpolation. However, we found that these did not contribute to improving the performance under the experimental setting. Original PP interval lengths were used as RR interval lengths in translated ECG signals during testing. This can be justified, as PPG recordings are used to analyze heart rate variability as an alternative to ECG (Lu et al., 2009; Aschbacher et al., 2020). Noise was added to the signals for robustness evaluation. The amplitudes of the baseline noise signals are 0.3, 0.4, and 0.1, and the frequencies are 0.3, 0.2 and 0.9 Hz, respectively. Gaussian noise of standard deviation 0.3.

3.3.2 Evaluation Metrics

ECG Translation from PPG

Pearson's correlation coefficient (ρ) measures how much an original ECG signal $\boldsymbol{y}_{1:T}$ and its reconstruction $\hat{\boldsymbol{y}}_{1:T}$ co-vary:

$$\rho = \frac{\left(\boldsymbol{y}_{1:T} - \bar{y}_{1:T}\right)^{\top} \left(\hat{\boldsymbol{y}}_{1:T} - \bar{\hat{y}}_{1:T}\right)}{\left\|\boldsymbol{y}_{1:T} - \bar{y}_{1:T}\right\|_2 \left\|\hat{\boldsymbol{y}}_{1:T} - \bar{\hat{y}}_{1:T}\right\|_2} \tag{3.17}$$

[1] https://github.com/khuongav/dvae_ppg_ecg

Root Mean Squared Error (RMSE) measures the differences between the values of the original signal and its reconstruction:

$$\text{RMSE} = \frac{\|\boldsymbol{y}_{1:T} - \hat{\boldsymbol{y}}_{1:T}\|_2}{\sqrt{n_y}} \tag{3.18}$$

Signal-to-Noise Ratio (SNR) compares the level of the desired signal to the level of undesired noise:

$$\text{SNR} = 20 \log \frac{\|\boldsymbol{y}_{1:T}\|_2^2}{\|\boldsymbol{y}_{1:T} - \hat{\boldsymbol{y}}_{1:T}\|_2^2} \tag{3.19}$$

AFib Detection

Performance was measured by the Area under the Receiver Operating Characteristic (ROC-AUC), the Area under the Precision-Recall Curve (PR-AUC), and the F1 score. The PR-AUC is considered a better measure for imbalanced data.

3.3.3 Implementation and Results

ECG Translation from PPG

Table 3.1 shows the performance of our model and compares it with other models in terms of means and standard deviations of ρ, RMSE and SNR. The correlation between the signals generated by our model and the reference signals is statistically strong, with a value ρ of 0.858. Also, low values of RMSE (0.07) and high SNR (15.365) show strong similarities between them and reference ECG signals. When the attention mechanism is not applied on

Table 3.1: ECG translation performance of different models. The top three rows show models' performance on healthy subjects, while the fourth row shows the performance on both the healthy and AFib subjects. If not specified, healthy subjects and clean signals is the default setting. The LSTM model (Tang et al., 2022) is subject-dependent, while the P2E-WGAN (Vo et al., 2021) and our model are subject-independent.

	Correlation	RMSE (mV)	SNR (dB)
ADSSM	0.858 ± 0.174	0.07 ± 0.047	15.365 ± 11.053
ADSSM w/o attention	0.823 ± 0.194	0.08 ± 0.047	13.013 ± 10.537
ADSSM (healthy sub., noisy sig.)	0.847 ± 0.174	0.076 ± 0.049	13.887 ± 10.58
ADSSM (healthy & AFib sub.)	0.804 ± 0.22	0.078 ± 0.05	12.261 ± 11.328
P2E-WGAN	0.773 ± 0.242	0.091 ± 0.052	9.616 ± 9.252
LSTM (sub. dependent)	0.766 ± 0.234	0.093 ± 0.053	8.189 ± 9.560

the input PPG, there is a notable decline in performance, with ρ falling to 0.823, RMSE increasing to 0.08, and SNR decreasing to 13.013. This underscores the importance of the mechanism in providing relevant contexts for translation. The third row shows our model's performance on the noisy dataset. The negligible drop in metrics from 0.858 to 0.847 (ρ), 0.07 to 0.76 (RMSE), and 15.365 to 13.887 (SNR) demonstrates the robustness of our model. We attribute this to the probabilistic nature of the model, which better handles the measurement noise. As expected, the model performed worse on subjects with AFib due to the erratic patterns of the AFib signals (no visible P waves and an irregularly irregular QRS complex). In the next section, we show that the synthetic AFib signals are beneficial to the downstream detection task.

The P2E-WGAN model (Vo et al., 2021), a 1D deep convolutional generative adversarial network (4,064,769 parameters) for signal-to-signal translation, was recently proposed to translate PPG into ECG signals from a large number of subjects. P2E-WGAN achieved

significantly lower performance than our model (645,466), requiring almost six times the parameters. Our model is less affected when data is scarce, which is common in healthcare. On the other hand, the LSTM model (Tang et al., 2022) is a deep recurrent neural network that was also recently proposed and built separately for each subject. The performance of our model, trained in a cross-subject setting, surpassed that of the LSTM model trained separately for each subject. These results prove the effectiveness and efficiency of our proposed sequential data structure. Further work with a larger number of subjects having AFib is needed to demonstrate that we can extend the model to new individuals. In addition, exploring strategies to manage the class imbalance problem (Johnson and Khoshgoftaar, 2019), which arises from the fewer AFib records compared to the healthy ones, would be beneficial.

In Figure 3.4, translated ECG waveforms are plotted with respect to the reference ECG waveforms of different heart rates. We can see that the model closely reconstructed the waveforms and maintained their essential properties, such as the missing P waves of the AFib ECG. In addition, we can be informed of the translation uncertainty by using a posterior on the latent embedding to propagate uncertainty from the embedding to the data. More specifically, with a distribution $p(\boldsymbol{z})$ on the latent feature our predictions will be $p_\theta(\boldsymbol{y} \mid \boldsymbol{x}) = \int p_{\theta_y}(\boldsymbol{y} \mid \boldsymbol{z})\, p_{\theta_z}(\boldsymbol{z} \mid \boldsymbol{x})\, d\boldsymbol{z}$. This would make the model more trustworthy and give patients and clinicians greater confidence in using it for medical diagnosis (Begoli et al., 2019). Future studies are expected to investigate methods to develop a fully Bayesian model and introduce a more flexible latent space (Tran et al., 2023; Bendekgey et al., 2024). Such advancements are advantageous in the medical field, particularly when data availability is limited or when uncertainty quantification and learning interpretable representations are essential.

Table 3.2: AFib detection performance. The performance on the translated ECG is evaluated when the MINA model (Hong et al., 2019) is trained on real ECG but tested on synthetic ECG. The fusion performance is when the MINA model is extended to receive both real ECG and synthetic ECG inputs. x% random time samples are omitted, simulating intermittent ECG recording, while synthetic ECG is always available.

	Real ECG	Translated ECG	
ROC-AUC	0.995 ± 0.006	0.99 ± 0.004	
PR-AUC	0.987 ± 0.013	0.986 ± 0.007	
F1	0.985 ± 0.009	0.944 ± 0.014	
Fusion	30% missing	50% missing	70% missing
ROC-AUC	0.992 ± 0.006	0.99 ± 0.006	0.99 ± 0.009
PR-AUC	0.986 ± 0.011	0.982 ± 0.012	0.981 ± 0.016
F1	0.971 ± 0.01	0.969 ± 0.012	0.956 ± 0.046

AFib Detection

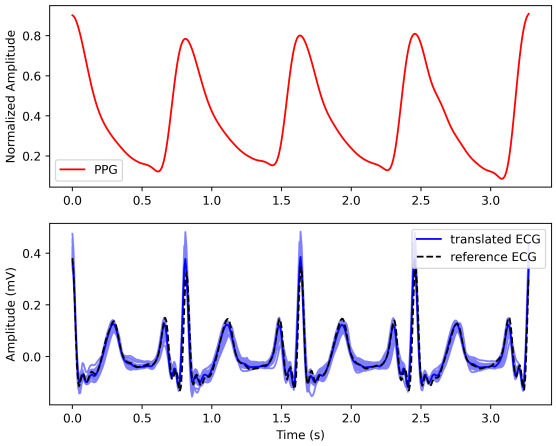
We evaluated the performance of our model on the benefits of the translated ECG for the AFib detection task. To do so, we used a state-of-the-art AFib detection model, Multilevel Knowledge-Guided Attention (MINA) (Hong et al., 2019), trained on real ECG signals, each of 10 s, and tested against synthetic. It should be noted that any pre-trained AFib detection model can be used in our pipeline. Table 3.2 reveals the mean detection performance of the model in the translated ECG that is close to that of the real ECG, ROC-AUC of 0.99 vs. 0.995, PR-AUC of 0.986 vs. 0.987, and F1 of 0.944 vs. 0.985. This implies that our model allows for the combined advantages of ECG's rich knowledge base and PPG's continuous measurement.

Furthermore, we extended the ability of the MINA model to receive real and translated ECG signals by incorporating the translated frequency channels into the model. In this scenario, both ECG and PPG signals can be measured simultaneously. This setting requires retraining of the MINA model on the fused real and synthetic ECG signal data set. To simulate the real-life setting where ECG measurement is intermittent while PPG input is continuous, we randomly zeroed out time samples with different probabilities: 30%, 50%, and 70%. As

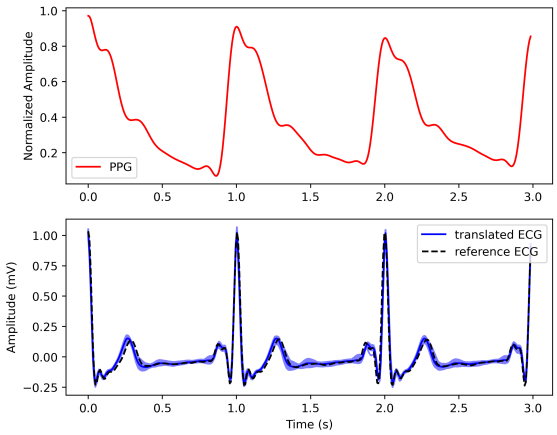
shown in the bottom results of Table 3.2, the performance remains almost unchanged in the fusion mode across the omission thresholds. Additionally, the model learns to utilize the sparse real ECG to marginally improve performance against only the translated ECG.

3.4 Conclusion

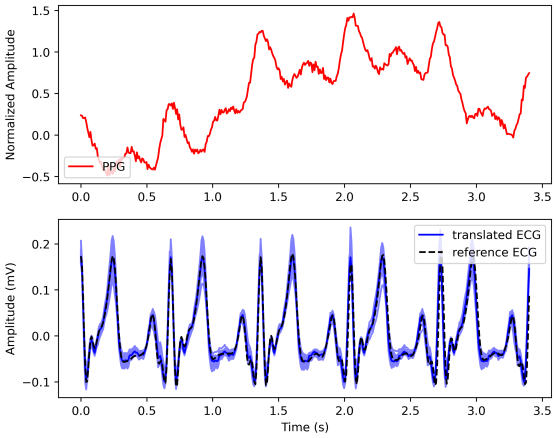
In this work, we present a novel attention-based deep state-space model to generate ECG waveforms with PPG signals as input. The results demonstrate that our model has the potential to provide a paradigm shift in telemedicine by bringing about ECG-based clinical diagnoses of heart disease via simple PPG assessment through wearable devices. Our model, trained on a small and noisy dataset, achieves an average Pearson's correlation of 0.847, RMSE of 0.076 mV, and SNR of 13.887 dB, demonstrating the efficacy of our approach. Significantly, our model enables the AFib monitoring capability in a continuous setting, assisting a state-of-the-art AFib detection model to achieve a PR-AUC of 0.986. Being a lightweight method also facilitates its deployment on resource-constrained devices. In our future work, we aim to validate the generalizability of the model with other pairs of physiological signals. Our method allows for the screening and early detection of cardiovascular diseases in the home environment, saving money and labor, while supporting society in unusual pandemic situations.



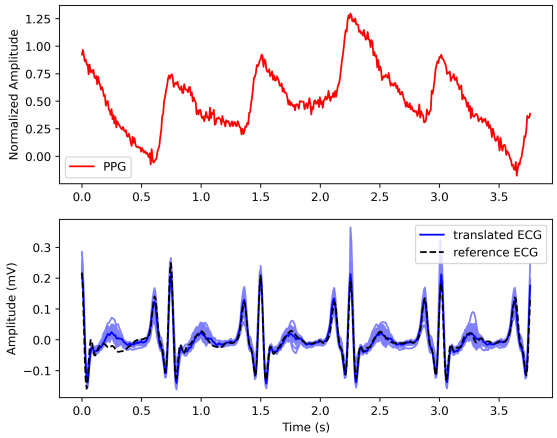
(a) Clean input PPG



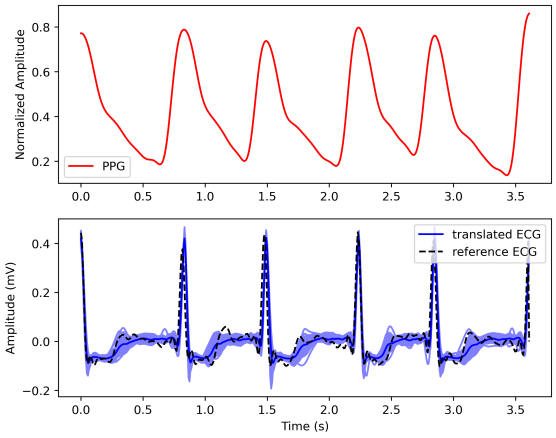
(b) Clean input PPG



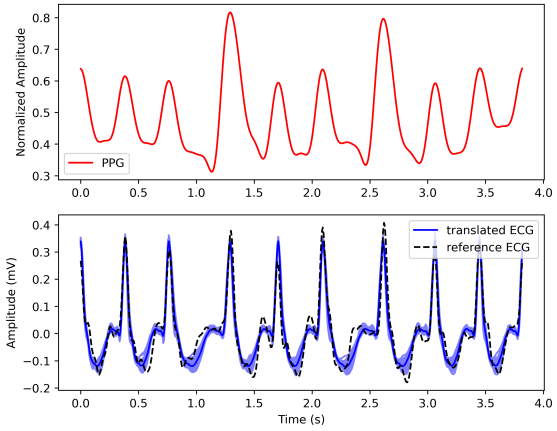
(c) Noisy input PPG



(d) Noisy input PPG



(e) AFib input PPG



(f) AFib input PPG

Figure 3.4: Examples of the translated ECG signals. In each subfigure: the top panel shows the input PPG waveform and the bottom panel shows the reconstructed ECG waveform compared with the reference waveform. The average ECG waveform (dark blue) of all possible pulses overlaid on each individual pulse (light blue).

Chapter 4

Composing Graphical Models with Generative Adversarial Networks for EEG Signal Modeling

4.1 Introduction

Electroencephalogram (EEG) is a non-invasive technique that measures the spontaneous electrical activity of the brain. EEG has been a driver of studies from basic neurological research to clinical applications. EEG modeling is essential to understanding the underlying mechanisms that generate brain signals and serve to design experiments and test hypotheses *in silico*. There exist extensive prior works on EEG computational models (Glomb et al., 2020) that derived principled neuroscience laws, empirically validated rules, or other domain expertise. Those are often in the form of general time-dependent and nonlinear partial differential equations. Nevertheless, they rely on strong assumptions which are not always generalizable. Further, those are slow to simulate and often suffer from model misspecifica-

tions.

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) provide a powerful framework and tools for machine learning, especially for deep representation learning and generative models. Over the past few years, GANs have witnessed tremendous advancements and achieved state-of-the-art performance in a variety of prominent tasks, including photo editing, video prediction, text generation, and signal synthesis (Jabbar et al., 2020; Vo et al., 2021). As a data-driven method, GANs are flexible and do not depend on rigid assumptions. Therefore, GANs hold great potential in modeling the inherent stochasticity and extrinsic uncertainty of EEG signals.

Recent work (Hartmann et al., 2018; Aznan et al., 2019; Pascual et al., 2019) applying GANs in EEG synthesis tend to simply characterize the spatio-temporal characteristics of EEG data subject to latent spaces of basic distributions, e.g., Gaussian or uniform distributions. Such assumptions impose limitations in capturing the intrinsic dependence among latent variables. Also, the GANs require deeper networks to synthesize longer sequences, which are computationally expensive and challenging to train, e.g., vanishing or exploding gradient problems. Moreover, the lack of inference capability in vanilla GANs hinder insight into structural information of EEG signals. On the other hand, probabilistic graphical models (Koller and Friedman, 2009; Wu et al., 2015) enable inference through structured representations but often lack the capability to model arbitrarily complex distributions.

To address these challenges, we propose a novel GAN-based approach for EEG signal modeling that couples deep implicit likelihoods (Mohamed and Lakshminarayanan, 2016) with structured latent variable representations to combine their complementary strengths. Our method uses graphical models for representing underlying structures of the signals, and applies ideas from the Graphical-GAN (Li et al., 2018) for effectively learning not only a generative model mapping from latent distributions to complex high-dimensional EEG data space but also an inverse inference model mapping from the data space to the latent space.

Our study paves the way for leveraging implicit probabilistic models to comprehensively investigate the mechanisms that generate brain waves.

4.2 Methodology

4.2.1 EEG Signal Synthesis with GANs

A GAN is a generative model trained by a pair of neural networks in a game-theoretic approach (Goodfellow et al., 2014). In GANs, a discriminator neural network D is trained to distinguish real from synthetic EEG signals, while a neural generator network G is trained to generate EEG signals from a latent space to make them indistinguishable by the discriminator. With EEG signal x drawn from data generating distribution $q(x)$, z drawn from noise prior p_z, and $p(x)$ is the generator's distribution over synthetic data, G and D jointly optimize the following objective:

$$\begin{aligned} \mathcal{L}_{GAN}(G, D) &= \mathbb{E}_{x\sim q(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z))] \\ &= \mathbb{E}_{x\sim q(x)}[\log D(x)] + \mathbb{E}_{x\sim p(x)}[\log(1 - D(x))] \end{aligned} \tag{4.1}$$

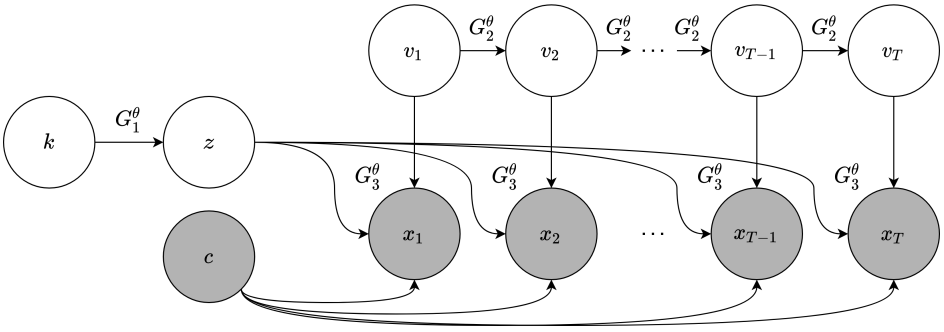
The discriminator is expected to output a high probability for a valid EEG signal and a low probability for a synthesized one, corresponding to the values of $\log D(x)$ and $\log(1-D(G(z))$, respectively. G and D are trained simultaneously until G is able to successfully fool D.

Following the proofs in (Goodfellow et al., 2014), given a fixed generator G, the optimal discriminator is given by $D^*(x) = \frac{q(x)}{q(x)+p(x)}$

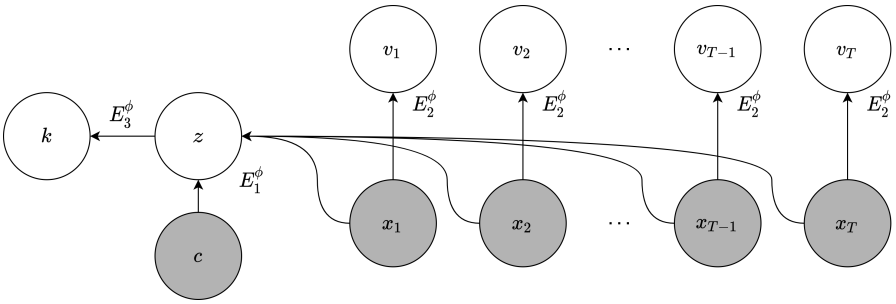
Under an optimal discriminator D^*, the generator minimizes the Jensen-Shannon (JS) divergence, which attains its minimum if and only if $p(x) = q(x)$.

4.2.2 Conjoining GANs with Bayesian Networks

Generative and Inverse Inference Process



(a) Generative model - p



(b) Inverse inference model - q

Figure 4.1: Directed graphical models for EEG signal modeling. Each time step corresponds to a δ-second multi-channel signal. Shaded nodes represent observed variables. Clear nodes represent latent variables. Directed edges indicate statistical dependencies between variables.

As shown in Figure 4.1, we model the generative process and the inverse inference process by a generative model and an inverse inference model in the Bayesian network. The framework exploits a Gaussian mixture model (GMM) to characterize the static latent variable structure with its capability to approximate arbitrary distributions, and a Markov model for the dynamic latent characterization. We use the notations p and q to denote the generative and inverse inference models.

The joint distribution of the generative model p is

$$\begin{aligned} & p(x_{1:T}, v_{1:T}, z, k, c) \\ & = p(k)p(z \mid k)p(c) \prod_{t=1}^{T} p(v_t \mid v_{t-1})p(x_t \mid z, v_t, c) \end{aligned} \tag{4.2}$$

where $p(k)$ and $p(c)$ are simple prior distributions for Gaussian mixture indicator k and condition c, e.g., a categorical distribution and a uniform distribution, $p(z \mid k)$ models a component selecting procedure for sampling noise z which encodes the temporal-spatial relationships invariant across time, v_t's form a first-order Markov chain, with $p(v_1|v_0) \sim \mathcal{N}(0, I)$, to encodes the temporal relationships variant across time, $p(x_t \mid z, v_t, c)$ specifies the conditional probability of the data at each time step t given noise z, state v_t, and condition c, and is of interest for the final generation.

The distribution function $p(x_{1:T}, v_{1:T}, z, k, c)$ is parametrized as generator neural networks. It consists of three parts: $z^p = G_1(k^p)$, $v_{t+1}^p = G_2(v_t^p, \epsilon_t)$, $\epsilon_t \sim \mathcal{N}(0, I)$, and $x_t^p = G_3(z, v_t^p, c)$. G_1 is responsible for a mapping from the input prior to a mixed Gaussian distribution with respect to k_p. G_2 transitions to a new state v_t^p given the previous state. G_3 uses noise z^p, state v_t^p, and condition c to generate the synthetic δ-second EEG signal x_t^p.

The joint distribution of the inverse inference model q is

$$\begin{aligned} & q(x_{1:T}, v_{1:T}, z, k, c) \\ & = q(x_{1:T})q(z \mid x_{1:T}, c)q(k \mid z) \prod_{t=1}^{T} q(v_t \mid x_t) \end{aligned} \tag{4.3}$$

where each latent variable of the Markov structure is assumed to be independent using the mean-field approximation (Jordan et al., 1999). $q(x_{1:T})$ is the empirical data distribution, $q(z \mid x_{1:T}, c)$, $q(v_t \mid x_t)$, and $q(k \mid z)$ are of interest for the inference. Contrary to $p(v_{t+1} \mid v_t)$, $q(v_t \mid x_t)$ models a dynamic tracing procedure for reconstructing the hidden features v_t. In

contrast to $p(z \mid k), q(k \mid z)$ models a component tracing procedure for reconstructing the Gaussian mixture indicator k.

The distribution function $q(x_{1:T}, v_{1:T}, z, k, c)$ is parametrized as extractor neural networks. It consists of three parts: $z^q = E_1(x^q_{1:T}, c)$, $v^q_t = E_2(x^q_t)$, and $k^q = E_3(z^q)$. E_1 and E_2 are responsible for a mapping from original signals to noise z^q and state v^q_t, respectively. E_3 infers within the latent space from z^q to k^q.

Learning Process

Our goal is to learn the parameters of the generative model p and the inverse inference model q by jointly minimizing the Jensen-Shannon (JS) divergence

$$JS(q(x_{1:T}, v_{1:T}, z, k, c) \| p(x_{1:T}, v_{1:T}, z, k, c)) \tag{4.4}$$

Expectation Propagation (EP) (Minka, 2013), a deterministic approximation algorithm, is proposed to utilize the locally structured data following (Li et al., 2018). The joint distributions can be factorized in terms of a set of factors $F_{\mathcal{G}} = \{(\mathrm{k}, \mathrm{z}), (\mathrm{v}_t, \mathrm{v}_{t-1}), (\mathrm{x}_t, \mathrm{v}_t, \mathrm{z}, c)\}$. For a factor a, the divergence of interest is

$$JS(q(a) \prod_{b \neq a} q(b) \| p(a) \prod_{b \neq a} p(b)) \tag{4.5}$$

EP iteratively minimize a local divergence in terms of each factor individually with the assumption that $\prod_{b \neq a} q(b) \approx \prod_{b \neq a} p(b)$. The divergence becomes

$$JS(q(a) \prod_{b \neq a} q(b) \| p(a) \prod_{b \neq a} q(b)) \tag{4.6}$$

Using the same proof sketch as in (Li et al., 2018), the divergence for factor a is approximated as

$$\begin{aligned} &JS(q(x_{1:T}, v_{1:T}, z, k, c) \| p(x_{1:T}, v_{1:T}, z, k, c)) \\ &\approx \mathbb{E}_q \left[\log \frac{2q(a)}{p(a) + q(a)} \right] + \mathbb{E}_p \left[\log \frac{2p(a)}{p(a) + q(a)} \right] \end{aligned} \tag{4.7}$$

The divergences are further averaged over all local factors as

$$\frac{1}{|F_{\mathcal{G}}|} \left[\mathbb{E}_q \left[\sum_{a \in F_{\mathcal{G}}} \log \frac{2q(a)}{p(a) + q(a)} \right] + \mathbb{E}_p \left[\sum_{a \in F_{\mathcal{G}}} \log \frac{2p(a)}{p(a) + q(a)} \right] \right] \tag{4.8}$$

Individual parametric discriminators D_a can be employed to estimate the local divergences as follows

$$\max_{\psi} \frac{1}{|F_{\mathcal{G}}|} \mathbb{E}_q \left[\sum_{a \in F_{\mathcal{G}}} \log \left(D_a(a) \right) \right] + \frac{1}{|F_{\mathcal{G}}|} \mathbb{E}_p \left[\sum_{a \in F_{\mathcal{G}}} \log \left(1 - D_a(a) \right) \right] \tag{4.9}$$

where ψ denotes the parameters in all discriminators. The discriminative models distinguish between the variables from the generative model p and those from the inverse inference model q as synthetic and original, respectively.

Optimization Objective

Three discriminators D_3, D_2 and D_1 receive local variable pairs, i.e., (k, z), (v_t, v_{t-1}), (x_t, v_t, z, c), from either the generative model p or the inverse inference model q, separately. The adversarial loss is as follows

$$
\begin{aligned}
&\mathcal{L}_{GAN}(G_*, E_*, D_*) \\
&= \mathbb{E}_q \left[\log D_3 \left(\mathrm{k}^q, \mathrm{z}^q \right) + \log D_2 \left(\mathrm{v}_t^q, \mathrm{v}_{t-1}^q \right) + \log D_1 \left(\mathrm{x}_t^q, \mathrm{v}_t^q, \mathrm{z}^q, c \right) \right] \\
&\quad + \mathbb{E}_p \left[\log \left(1 - D_3 \left(\mathrm{k}^p, \mathrm{z}^p \right) \right) + \log \left(1 - D_2 \left(\mathrm{v}_t^p, \mathrm{v}_{t-1}^p \right) \right) \right. \\
&\qquad \left. + \log \left(1 - D_1 \left(\mathrm{x}_t^p, \mathrm{v}_t^p, \mathrm{z}^p, c \right) \right) \right]
\end{aligned}
\tag{4.10}
$$

All components are trained simultaneously in an adversarial process. Let θ and ϕ denote the parameters of G_* and E_*, respectively. Iteratively, D_* learn to maximize Equation 4.10 by updating ψ, while G_* and E_* learn to minimize Equation 4.10 by updating corresponding parameters θ and ϕ, respectively.

In order to ensure the global consistency of an entire signal across time steps, a frequency domain loss is added as

$$
\mathcal{L}_f(G_*) = \|\bar{r}(x_{i,1:T}^q) - \bar{r}(x_{i,1:T}^p)\|_1 + \|\bar{\varphi}(x_{i,1:T}^q) - \bar{\varphi}(x_{i,1:T}^p)\|_1 \tag{4.11}
$$

where $\bar{r}$ and $\bar{\varphi}$ refer to the average magnitude and phase across signals i in a batch, respectively. They are computed by a fast Fourier transform (FFT). Hence, the total objective is

$$
\min_{G_*, E_*} \max_{D_*} \mathcal{L}_{GAN} + \lambda \mathcal{L}_f \tag{4.12}
$$

4.2.3 Network Architectures and Training Hyperparameters

Table 4.1 presents the architectures of the deep neural networks. Each time step corresponds to a 1-second EEG signal ($\delta = 1$). All the feature maps have 96 channels. Leaky ReLU activation functions are applied to all layers, with the slope 0.1 to stimulate easier gradient

Table 4.1: Network architectures. Models having similar architectures are grouped together.

$\mathbf{G_2}, \mathbf{D_3}, \mathbf{D_2}$	$\mathbf{G_3}$	
Linear 512, (SN), lReLU	Linear 1536, lReLU	
Linear 512, (SN), lReLU	Reshape 96x16	
Linear 256, (SN), lReLU	Upsample	
G_2 Linear 32	Conv 6, BN, lReLU	X 4
D_3 Linear 1, SN, Sigmoid	Conv 6, BN, lReLU	
D_2 Linear 1, SN, Sigmoid	Conv 1, Tanh	

$\mathbf{D_1}$	$\mathbf{E_2}, \mathbf{E_1}$	
Get x_t or $x_{[1,T]}$ (concatenated along channels)		
Conv 1, lReLU - 96x256		
Conv 6, BN/SN, lReLU		X 4
Conv 6, Stride 2, BN/SN, lReLU		
Reshape 1536		
Get v_t, z, c	E_2 Linear 32	
Linear 256, SN, lReLU	E_1 Linear 128	
Join features of x_t, v_t, z, c		
Linear 512, SN, lReLU		
Linear 1, SN, Sigmoid		

flow. Batch normalizations (BN) (Ioffe and Szegedy, 2015) are used at each convolutional layer of the generators and extractors. Spectral normalizations (SN) (Miyato et al., 2018) are applied to the discriminators to constrain their Lipschitz constants. c are subject embeddings as one-hot vectors. The sizes of z, k, and v_t, and ϵ_t are set at 128, 6, 32, and 16 respectively.

G_1 and E_2 are single-layer neural networks. We use the reparameterization trick (Kingma and Welling, 2013) to estimate the gradients with the continuous variable z, and the Gumbel-Softmax trick (Jang et al., 2016) (the temperature of 0.1) to estimate the gradients with the discrete variable k.

λ is set at 0.1 to have the training process driven mainly by the adversarial loss. In order to mitigate the issue of slow learning in regularized discriminators, a higher learning rate is provided to the discriminators than the generators and extractors by the Two Time-scale Update Rule (TTUR) (Heusel et al., 2017). The models are trained with the Adam

optimizer with the initial learning rate of 0.0004 for D_*, the learning rate of 0.0001 for G_* and E_*, and the exponential decay rates $\beta_1 = 0.5$ and $\beta_2 = 0.999$. All weights are initialized using a zero-centered Gaussian distribution with a standard deviation of 0.02. We make the implementation publicly available [1].

4.3 Experiments

4.3.1 Dataset

The 23-channel interictal EEG recordings from the CHB-MIT epilepsy dataset (Shoeb, 2009) are used for the experiments. The dataset consists of scalp EEG from pediatric subjects with intractable seizures. We select a subset of 6 patients (chb01-03, chb05-06, chb10) having the same measurement setup, including males and females, 1.5-14 years old. Interictal periods are extracted at least 4-hour away before a seizure onset and after the seizure ends. The signals are low-pass filtered with a cut-off frequency at 50 Hz and scaled to the range $[-1, 1]$. Overall, the dataset contains 43593 signals, from which 70% are used for training and validation, and the other 30% are used as the test set. Each signal is 10-second long (T=10), at a sampling rate of 256 Hz. Additionally, 339 ictal EEG signals are extracted for evaluating epilepsy seizure detection performance.

4.3.2 Evaluation Metrics

Sliced 2-Wasserstein distance (SWD) (Bonneel et al., 2015; Flamary et al., 2021) quantifies the cost of transforming one distribution to another. It is an approximation to the 2-

[1] https://github.com/khuongav/Graphical-Adversarial-Modeling-of-EEG

Wasserstein distance using 1D projections for a closed-form solution and is defined as

$$SWD_2(\mu,\nu) = \mathop{\mathbb{E}}_{\theta\sim\mathcal{U}\left(\mathbb{S}^{d-1}\right)} \left[\mathcal{W}_2^2\left(\theta_{\#}\mu, \theta_{\#}\nu\right)\right]^{\frac{1}{2}} \tag{4.13}$$

where μ and ν are two probability measures, $\theta_{\#}\mu$ stands for the pushforwards of the projection $\mathbb{R}^d \ni X \mapsto \langle\theta, X\rangle$, and $\mathcal{U}\left(\mathbb{S}^{d-1}\right)$ is the uniform distribution on the hypersphere of d dimensions.

Spectral entropy (SEN) measures the uniformity the of signal energy distribution in the frequency-domain. It is given by

$$\mathbb{H}(x) = -\sum_{f=0}^{f_s/2} P(f)\log_2[P(f)] \tag{4.14}$$

where P is the normalised power spectral density, and f_s is the sampling frequency of signal x.

Reconstruction error (REC) measures the differences between the values of an original signal and its reconstruction $\tilde{x}$ as

$$REC = \|x^q_{1:T} - \tilde{x}^q_{1:T}\|_1 \tag{4.15}$$

4.3.3 Results and Discussion

Table 4.2 presents the performance of our proposed approaches and the comparison with the BiGAN/ALI model (Dumoulin et al., 2016; Donahue et al., 2016). We denote its conditional version as C-BiGAN/ALI. GMMarkov-GAN is our model characterized by Gaussian mixture and Markov latent structures, while Markov-GAN is only with the Markov structure. C-BiGAN/ALI is the GAN with an inference capability but without a latent variable structure,

Table 4.2: Performances of different GAN models in interictal EEG signal synthesis and reconstruction tasks.

	SWD	REC	SEN
Original data			0.620 ± 0.070
GMMarkov-GAN	**1.16e-2**	**0.0474 ± 0.0392**	**0.608 ± 0.063**
Markov-GAN	1.34e-2	0.0494 ± 0.0413	0.636 ± 0.070
GMMarkov-GAN (w/o FFT)	1.70e-2	0.0519 ± 0.0438	0.585 ± 0.074
Markov-GAN (w/o FFT)	1.78e-2	0.0530 ± 0.0391	0.583 ± 0.069
C-BiGAN/ALI	2.13e-2	0.0562 ± 0.0415	0.539 ± 0.066

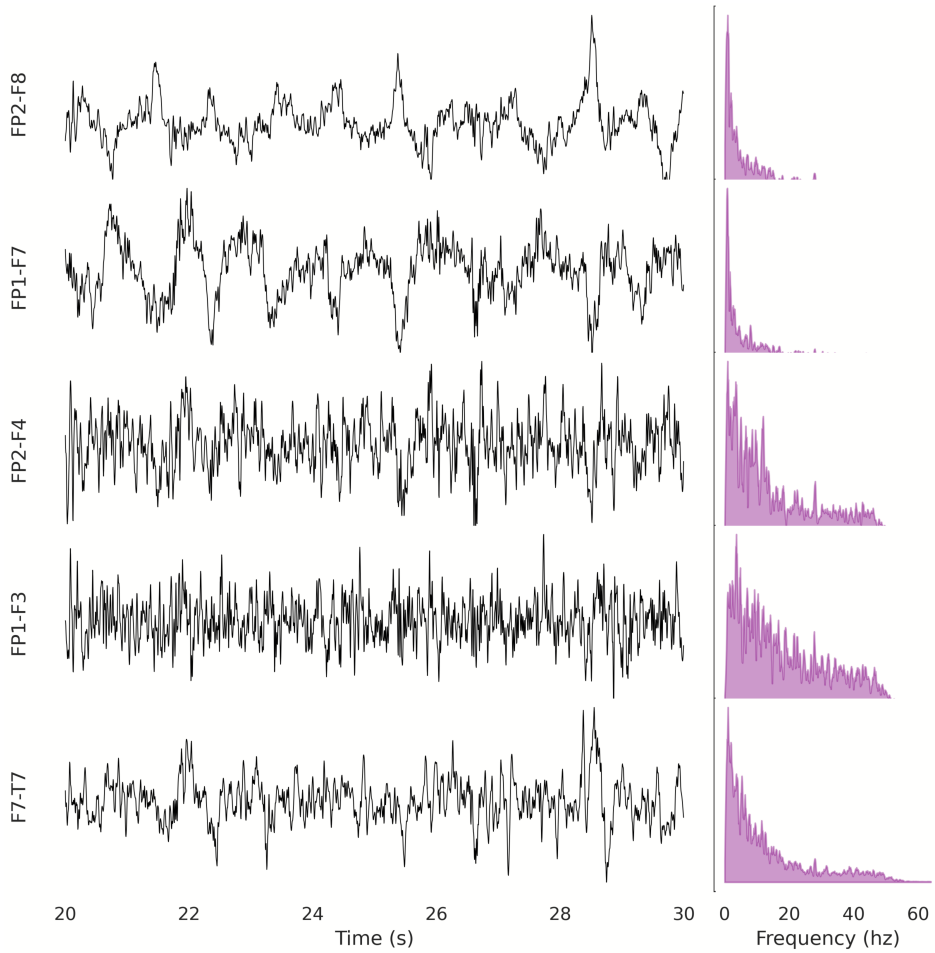


Figure 4.2: Last 10-second of a 30-second synthetic 23-channel EEG signal by the GMMarkov-GAN model, conditioned on patient 3. 5 channels with the highest standard deviations are shown.

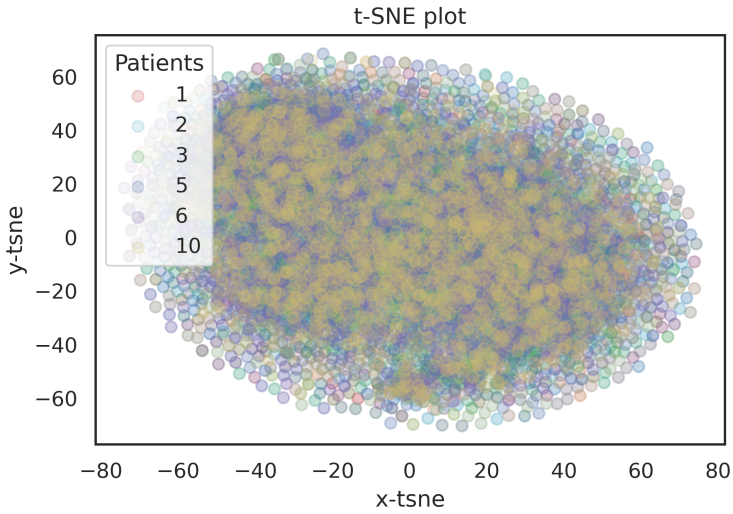
in which the latent space is a simple Gaussian, and data at each timestep are generated independently.

Both the graphical GANs achieve significantly lower SWD, REC, and SEN differences than C-BiGAN/ALI, indicating that they are better at capturing the characteristics of EEG in both time and frequency domains. Besides, by encoding the invariant spatial-temporal features of EEG signals subject to the flexibility of a Gaussian mixture, GMMarkov-GAN enjoys better performance (SWD of 0.0173, REC of 0.0519, and SEN difference of 0.035) than the Markov-GAN. We attribute this to GMMarkov-GAN being able to learn a structured clustering of the latent space as shown in Figure 4.3. These results prove the effectiveness of our proposed data structures and confirm our inverse inference strategy.

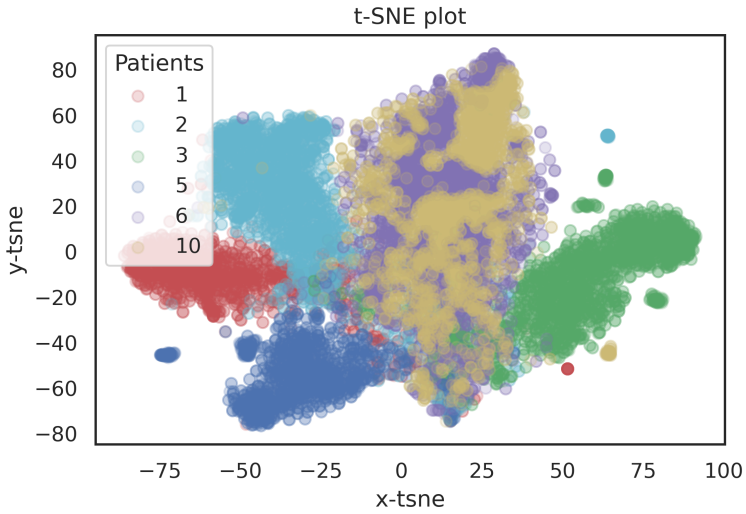
By training with the additional FFT loss, GMMarkov-GAN enjoys the highest performance (SWD of 0.0116, REC of 0.0474, and SEN difference of 0.012). It should be noted that the frequency-domain loss added little time for training, yet it noticeably improved the results.

In Figure 4.2, synthetic multi-channel EEG signals are plotted. The signals are naturally realistic across channels and show good fits in different frequency bands. Although our model is trained on 10 second-long signals, it can generate much longer sequences of 30 seconds, thanks to the Markov structure.

To demonstrate the efficacy of our generative and inverse mapping approach for auxiliary tasks, we further evaluated our approach in epilepsy seizure detection. As the model is trained on the interictal EEG signals, seizure segments are detected with reconstruction error thresholds in an anomaly detection framework. Figure 4.4 shows a high detection performance from our model by the ROC curve with the area under the curve of 0.92, competitive with contemporary approaches in supervised learning (Siddiqui et al., 2020). We plan to build on these results in our future work for interpreting more encoded features in the low-dimensional manifolds and further investigate the partial mode collapse issue of GANs (Bau et al., 2019).



(a) C-BiGAN/ALI



(b) GMMarkov-GAN

Figure 4.3: t-SNE visualization of the static latent spaces.

4.4 Conclusion

In this work, we proposed an EEG modeling scheme that combines the strengths of probabilistic graphical models and generative adversarial networks. Our experimental results demonstrate that our method effectively characterized EEG latent variable structure via a Gaussian mixture and a Markov model. The structured representations can provide interpretability and encode inductive biases to reduce the data complexity of neural oscillations.

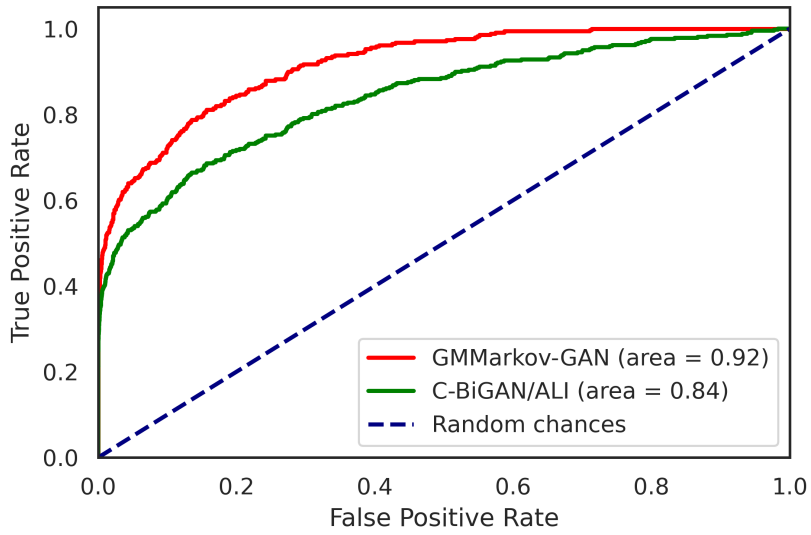


Figure 4.4: ROC curve for epilepsy seisure detection.

Our approach holds promise to new generative applications in neuroscience and neurology. Future directions include generalizing learning and inference algorithms with more complicated structures to truly model the underlying relationships at different scales spanning from the single cell spike train up to macroscopic oscillations.

Chapter 5

Deep Latent Variable Joint Cognitive Modeling of Neural Signals and Human Behavior

5.1 Introduction

Current approaches to understanding brain function emphasize the search for statistical relationships between human behavior and individual physiological measures (EEG, fMRI, fNIRS, etc.; e.g. Itthipuripat et al., 2019). Behavioral measures, such as accuracy and speed of responses, reflect latent cognitive processes that underlie decision making that are not observed directly and must be inferred by cognitive models (Lee and Wagenmakers, 2014). An ongoing challenge in computational cognitive neuroscience research is formulating the link between brain activity and latent cognitive processes. Here, we present a novel approach that allows a theoretical account of the cognitive process of decision-making, and artificial neural networks to estimate a joint latent space to link cognitive parameters to both neural

signals and behavioral measures. This joint latent space model is a valuable new framework for computational cognitive neuroscience, allowing for new forms of inference and hypothesis generation.

Previous work has focused on neurocognitive relationships between human neural data and behavioral data in decision-making tasks (Nunez et al., 2015, 2017, 2019; Lui et al., 2021; Turner et al., 2013, 2016). The hierarchical Bayesian models used in these projects make strong predictions about the relationships between brain activity and the speed of decision-making. These models typically make use of the drift-diffusion model (DDM; Ratcliff and McKoon, 2008), a widely-used cognitive model in decision-making, as their generative model of choice and reaction time data. To integrate neural signals, these models require knowledge of previously discovered features of the neural data (e.g., known functional signals in the cognitive neuroscience literature) that are then linked by prescribed (usually linear) relationships to the latent cognitive variables in a Bayesian hierarchical model. The resulting *neurocognitive* models test the relationship between neural signals and cognitive variables, and enhance the accuracy of predictions of behavior directly from brain signals (Turner et al., 2016; Nunez et al., 2017). This can be thought of as one domain of the larger field of *model-based cognitive neuroscience* (Forstmann and Wagenmakers, 2015).

A limitation of this approach is that we must know in advance which brain signals are possibly linked to cognitive functions. However, advances in frameworks and tools for neuroscience allow for the discovery of previously unknown neural features that we could use to explain latent cognitive variables. Ideally, such frameworks operate across observations, experimental manipulations, and individual differences. Deterministic models that leverage deep learning have been proposed for learning feature representation of EEG data to analyze and decode brain activity (Roy et al., 2019). As a notable example, Sun et al. (2022) have proposed a SincNet-based neural network that made use of EEG signals to learn the latent cognitive variables of the DDM on individual decisions. This approach identifies time windows of

information processing and frequency bands that can be used to predict latent processes directly from EEG data as a trial-level association between neural features, choice, and response time.

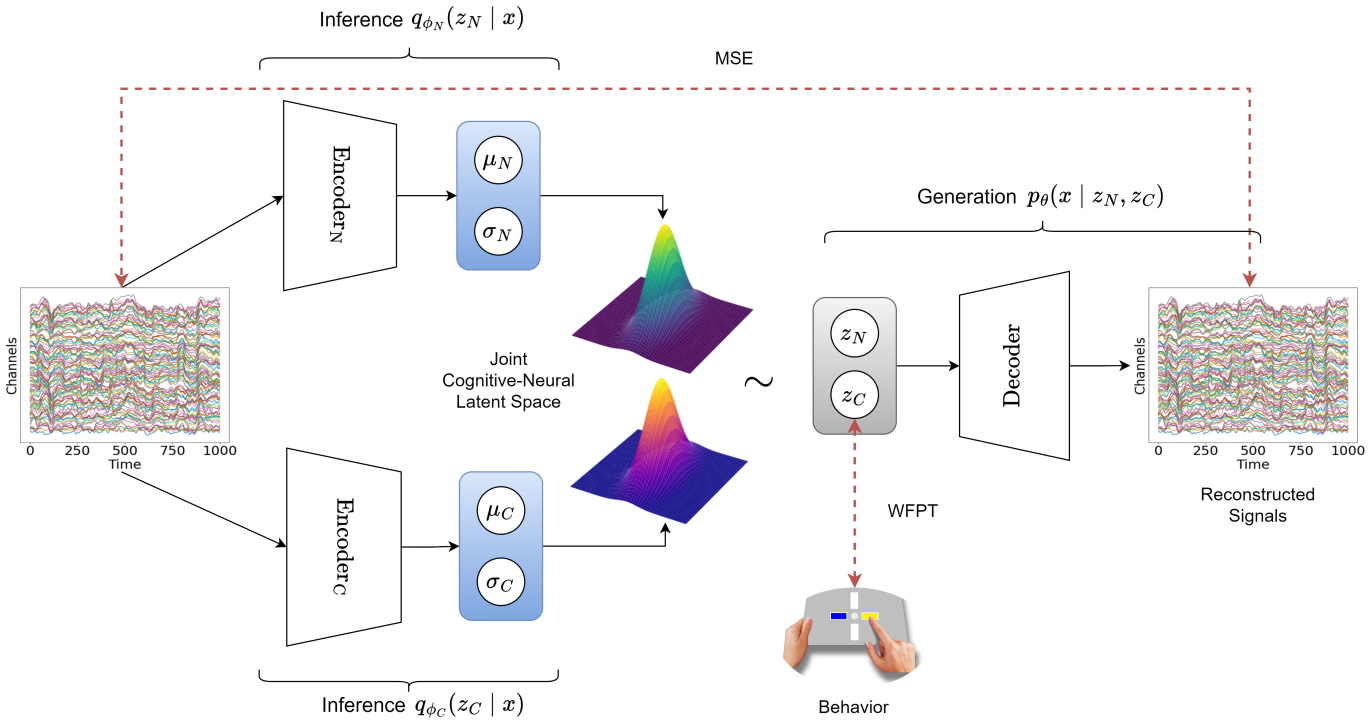
This work aims to develop a deep probabilistic method for linking neural data from EEG to the latent parameters of a cognitive model. The innovation of our work lies in the use of a theoretical account of the cognitive process. This theoretical account drives the analysis of neural and behavioral measures. The framework allows for one-step, joint inference on integrative neurocognitive models that map EEG and behavior into a joint latent space. Uniquely, this new approach has the potential to allow us to generate task-relevant EEG signals from behavioral data, and *predict modulation of EEG signals by cognitive model parameters*. By combining the exploratory potential of modern latent variable methods with the theoretical appeal of human-interpretable cognitive model parameters, the proposed technique can be used to make predictions of brain signals and cognitive parameters in future experiments to test neurocognitive theories.

5.2 Neurocognitive Variational Autoencoders

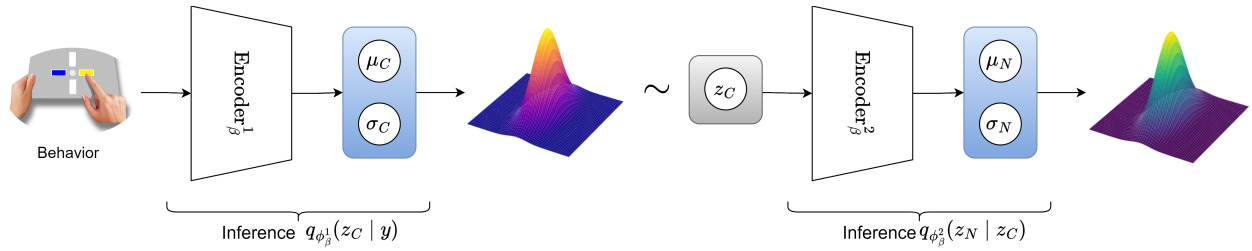
5.2.1 Generative EEG Modeling with VAEs

Consider first a data set $\mathcal{P} \stackrel{\text{def}}{=} \{\mathcal{D}_1, \ldots, \mathcal{D}_M\}$ containing M subjects, where each subject $\mathcal{D}_m \stackrel{\text{def}}{=} \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_I\}$ consists of I trials $\boldsymbol{x}_i \in \mathbb{R}^{C \times T}$ that are EEG signals of C channels by T time samples. Throughout the paper, the subscript m is omitted when we refer to only one subject or when it is clear from the context.

For each subject m, we aim to learn an EEG generative process with a latent-variable model comprising of a fixed Gaussian prior over latent variables $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z} \mid \boldsymbol{0}, \boldsymbol{I})$, where $\boldsymbol{I}$ is



(a) Generative Process



(b) Regularized Discriminative Process

Figure 5.1: The Neurocognitive VAE. After the generative process (a) learns the joint latent neurocognitive variables (Section 5.2.2), the regularized discriminative process (b) retrofits its hierarchical latent space to the joint latent space (Section 5.2.3). Inference networks q and Generation networks p contain neural network parameters θ and ϕ. Black arrows: flows of operations. Red arrows: loss functions. MSE and WFPT stand for Mean Squared Error and Wiener First Passage Time, respectively. The heatmaps represent the probability distributions in the latent spaces. Plasma color maps are for the drift-diffusion variables ($z_C \in \mathbb{R}^3$), while greenery color maps are for residual neural variables ($z_N \in \mathbb{R}^{32}$). Blue blocks contain μ and σ, which are the parameters of the multivariate Gaussian latent spaces. Gray blocks contain z sampled ($\sim$) from the distributions. The variables x and y represent EEG signals and choice-RTs, respectively. Each trapezoid represents a different convolutional neural network (see Table 5.2 for detailed architectures).

the identity covariance matrix, and a parametric non-linear Gaussian likelihood $p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$. The learning process finds θ such that the Kullback-Leibler (KL) divergence is minimized between the true data generating distribution $p_{\mathcal{D}}$ and the model p_θ:

$$\begin{aligned} &\arg\min_{\theta} KL\left(p_{\mathcal{D}}(\boldsymbol{x}) \| p_\theta(\boldsymbol{x})\right) \\ &= \arg\max_{\theta} \mathbb{E}_{p_{\mathcal{D}}(\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x})\right] \end{aligned} \tag{5.1}$$

where $p_\theta(\boldsymbol{x}) = \int_{\mathcal{Z}} p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z}$ is the likelihood of data point $\boldsymbol{x}$, approximated by averaging over the latent $\boldsymbol{z}$.

Nevertheless, estimating $p_\theta(\boldsymbol{x})$ is typically intractable. This issue can be mitigated by introducing a parametric inference model $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ to construct a variational evidence lower bound on the log-likelihood $\log p_\theta(\boldsymbol{x})$ as follows:

$$\begin{aligned} &\mathcal{L}(\boldsymbol{x}; \theta, \phi) \\ &\quad \stackrel{\text{def}}{=} \log p_\theta(\boldsymbol{x}) - KL\left(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \| p_\theta(\boldsymbol{z} \mid \boldsymbol{x})\right) \\ &\quad = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})\right] - KL\left(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \| p(\boldsymbol{z})\right) \end{aligned} \tag{5.2}$$

Taking the likelihood model $p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$ to be a decoder and the inference model $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ to be an encoder, a variational autoencoder (VAE; Kingma and Welling, 2013; Sohn et al., 2015) considers this objective from a deep probabilistic autoencoder perspective. Here, θ and ϕ are neural network parameters, and learning takes place via stochastic gradient ascent using unbiased estimates of $\nabla_{\theta,\phi} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(\boldsymbol{x}_i; \theta, \phi\right)$.

In the following sections, we extend the traditional VAE to create the Neurocognitive VAE (NCVA) (Figure 5.1). This model allows us to model a joint distribution of neural and behavioral data. Instead of a training technique that encourages disentanglement, as in β-VAE (Higgins et al., 2016), NCVA imposes restrictions on latent space by using a cognitive model that provides interpretability and controllable generation.

5.2.2 Disentangled Cognitive Latent Space of EEG

Now consider the data $\mathcal{D}_m \stackrel{\text{def}}{=} \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_I, \boldsymbol{y}_I)\}$, consisting, on the one hand, of N trials of the EEG data $\boldsymbol{x}_i$ and, on the other hand, of the corresponding choice response times (choice-RT) $\boldsymbol{y}_i$. Both $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ are associated with a context vector $\boldsymbol{c}_i$ (where the applicable context might be an experimental condition; say, noise conditions $\boldsymbol{c}_i$). For mathematical simplicity, the context vector $\boldsymbol{c}$ is not mentioned when we refer to one of the data modalities.

Crucially, we propose a generative model with two sources of variation: $\boldsymbol{z}_C$, which is cognitively specific, and $\boldsymbol{z}_N$, which captures any residual neural variations left in $\boldsymbol{x}$. We assume the approximate posterior $q_\phi(\boldsymbol{z}_N, \boldsymbol{z}_C \mid \boldsymbol{x})$ has the following fully factorized form:

$$
\begin{aligned}
q_\phi\left(\boldsymbol{z}_N, \boldsymbol{z}_C \mid \boldsymbol{x}\right) &= q_{\phi_N}\left(\boldsymbol{z}_N \mid \boldsymbol{x}\right) q_{\phi_C}\left(\boldsymbol{z}_C \mid \boldsymbol{x}\right) \\
q_{\phi_N}\left(\boldsymbol{z}_N \mid \boldsymbol{x}\right) &= \mathcal{N}\left(\boldsymbol{z}_N \mid \boldsymbol{\mu}_{\phi_N}(\boldsymbol{x}), \operatorname{diag}\left(\boldsymbol{\sigma}^2_{\phi_N}(\boldsymbol{x})\right)\right) \\
q_{\phi_C}\left(\boldsymbol{z}_C \mid \boldsymbol{x}\right) &= \mathcal{N}\left(\boldsymbol{z}_C \mid \boldsymbol{\mu}_{\phi_D}(\boldsymbol{x}), \operatorname{diag}\left(\boldsymbol{\sigma}^2_{\phi_D}(\boldsymbol{x})\right)\right)
\end{aligned}
\tag{5.3}
$$

A Gaussian prior over latent variables $p(\boldsymbol{z}_C)$ can be chosen for each subject. We use subject priors obtained from a Bayesian hierarchical fitting of a DDM using the Markov chain Monte Carlo (MCMC) (Nunez et al., 2019).

We learn the generative model by maximizing the lower bound on $\log p_\theta(\boldsymbol{x}, \boldsymbol{y})$ as:

$$
\begin{aligned}
&\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}; \theta, \phi_N, \phi_C) \\
&= \mathbb{E}_{q_\phi(\boldsymbol{z}_N, \boldsymbol{z}_C \mid \boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z}_N, \boldsymbol{z}_C) + \log p(\boldsymbol{y} \mid \boldsymbol{z}_C)\right] \\
&\quad - KL\left(q_{\phi_N}(\boldsymbol{z}_N \mid \boldsymbol{x}) \| p(\boldsymbol{z}_N)\right) \\
&\quad - KL\left(q_{\phi_C}(\boldsymbol{z}_C \mid \boldsymbol{x}) \| p(\boldsymbol{z_C})\right)
\end{aligned}
\tag{5.4}
$$

where $p_\theta(\boldsymbol{x} \mid \boldsymbol{z}_N, \boldsymbol{z}_C) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_\theta(\boldsymbol{z}_N, \boldsymbol{z}_C), \boldsymbol{I})$ and $p(\boldsymbol{y}|\boldsymbol{z}_C)$ can be any neurocognitive likelihood. This work applies the Wiener First Passage Time distribution (WFPT; Navarro and

Fuss, 2009) corresponding to the lower boundary:

$$
\begin{aligned}
& p(\boldsymbol{y}|\boldsymbol{z}_C) \\
&= \text{ Wiener } (\text{RT} \mid \alpha, \tau, \delta) \\
&= \frac{\pi}{\alpha^2} e^{-\frac{1}{2}\left(\alpha\delta + \delta^2(RT-\tau)\right)} \\
&\quad \times \sum_{k=1}^{+\infty} \left[k \sin\left(\frac{\pi k}{2}\right) e^{-\frac{k^2\pi^2}{2\alpha^2}(RT-\tau)} \right]
\end{aligned}
\tag{5.5}
$$

The probability at the upper boundary is obtained by setting $\delta' = -\delta$. $\boldsymbol{z}_C$ comprises of three parameters including drift rate δ, boundary α, non-decision time (ndt) τ. The bias towards correct or incorrect responses is fixed at 0.5, that is, the starting point is always unbiased.

The joint inference is performed using only EEG $\boldsymbol{x}$ to ensure that encoder θ_C would learn to extract neural features that are tailored to cognitive parameters, without relying on choice-RT $\boldsymbol{y}$. This has the advantage of providing more accurate trial-level parameter estimates that are associated with the EEG data.

Note that the dimension of the cognitive space is significantly lower than that of the residual neural space. This facilitates the representation of the variation in neural signals only through flexible $\boldsymbol{z}_N$. Maximizing the likelihood of observing neural signals does not guarantee decoder θ utilizing $\boldsymbol{z}_C$ to output $\boldsymbol{x}$. In the next section, we present an approach to capture the correlation between behavior and cognition, as well as the mapping of the variability of behavior and cognition to neural signals.

5.2.3 Structured EEG Modeling from Behavior

Here, we propose a discriminative model regularized by the generative model learned in the previous section. We aim to discriminatively learn the distribution of the cognitive parameters conditioned on behaviors, and the distribution of the neural latent variables

conditioned on cognitive parameters. The joint latent space inferred from the behavior can be factorized into the two-level latent space as follows:

$$q_{\phi_B}\left(\boldsymbol{z}_N, \boldsymbol{z}_C \mid \boldsymbol{y}_i\right) = q_{\phi_B^2}\left(\boldsymbol{z}_N \mid \boldsymbol{z}_C\right) q_{\phi_B^1}\left(\boldsymbol{z}_C \mid \boldsymbol{y}_i\right) \tag{5.6}$$

Inspired by Suzuki et al. (2016), we learn the following approximations, w.r.t parameter ϕ_B^1:

$$\mathbb{E}_{p_{\mathcal{D}}}\left[KL\left(q_{\phi_C}(\boldsymbol{z}_C \mid \boldsymbol{x}) \mid q_{\phi_B^1}(\boldsymbol{z}_C \mid \boldsymbol{y})\right)\right] \tag{5.7}$$

and w.r.t parameter ϕ_B^2:

$$\mathbb{E}_{p_{\mathcal{D}}}\left[KL\left(q_{\phi_N}(\boldsymbol{z}_N \mid \boldsymbol{x}) \mid q_{\phi_B^2}(\boldsymbol{z}_N \mid \boldsymbol{z}_C)\right)\right] \tag{5.8}$$

By decomposing the KL divergences as in Hoffman and Johnson (2016); Vedantam et al. (2017), we effectively minimize $KL\left(q_{\phi_C}^{\text{avg}}(\boldsymbol{z}_C \mid \boldsymbol{x}) \mid q_{\phi_B^1}(\boldsymbol{z}_C \mid \boldsymbol{y})\right)$ and $KL\left(q_{\phi_N}^{\text{avg}}(\boldsymbol{z}_N \mid \boldsymbol{x}) \mid q_{\phi_B^2}(\boldsymbol{z}_N \mid \boldsymbol{z}_C)\right)$, where $q_\phi^{\text{avg}}(\boldsymbol{z} \mid \boldsymbol{x}) = \mathbb{E}_{p(\boldsymbol{x}\mid\boldsymbol{y})}\left[q_\phi(\boldsymbol{z} \mid \boldsymbol{x})\right]$. As there is little posterior uncertainty once conditioned on an EEG signal $\boldsymbol{x}_i$, the approximations are close to the average posterior induced by each of the EEG $\boldsymbol{x}_i$ associated with similar $\boldsymbol{y}$.

Having fit both the generative and discriminative models, we can now explore the three-way relationship between behavior, brain activity, and cognitive processes.

5.3 Experiments

5.3.1 EEG and Behavioral Dataset

We used behavioral and EEG data collected while participants performed a two-alternative forced-choice task where they had to decide whether a Gabor patch presented with added dynamic noise is higher or lower spatial frequency (for details, see Experiment 2 by Nunez et al., 2019). Task difficulty was manipulated by adding spatial white noise to manipulate the quality of the perceptual evidence available to make the discrimination. The signal and the noise flickered at 30 and 40 Hz frequencies, respectively. 4 participants performed the task in blocks of trials at 3 added noise levels (low, medium, and high). Each subject performed approximately 3000 trials over 7 experimental sessions, while 128 channels of EEG and behavioral data were recorded. The independent component analysis (ICA)-based artifact rejection method was used on EEG data to remove eyeblinks, electrical noise, and muscle artifacts. A subset of 98 EEG channels were selected, excluding channels located in the outer ring. EEG data were bandpass filtered to 1 to 45 Hz in the frequency domain and then downsampled from 1000 Hz to 250 Hz in the time domain prior to data analysis. The data for each subject were divided into 80% for training and validation and the remaining 20% for testing.

5.3.2 Results

To validate the neurocognitive modeling approach, we first examine the trial-by-trial variability of the parameters within each subject and the generalization of the model to unseen data. Figures 5.2a and 5.2c show the trial-by-trial correlations between estimated DDM posteriors and observed choice-RTs in the training data from neural signals and behavior, respectively. Spearman correlations between fitted drift rates (δ) and choice-RTs are nega-

Table 5.1: Comparison of the sum of Wiener negative log-likelihood ($-\sum \log$ Wiener ($\mathrm{RT}_i \mid \boldsymbol{\omega}_i$)) of four subjects on the test sets. $\bar{\boldsymbol{\omega}}$ represents the median fitted cognitive parameters from the training set.

Subjects	$\boldsymbol{\omega}_i^{\text{test}}$	$\bar{\boldsymbol{\omega}}^{\text{train}}$
s1	−0.018	0.212
s2	−0.244	0.159
s3	0.264	0.735
s4	0.031	0.230

tively strong. At the same time, there are strong positive correlations between boundaries (α) and choice-RTs, as well as between non-decision time and choice-RTs. The estimates in NCVA are regularized by the subject priors obtained from a Bayesian hierarchical fitting of a DDM using MCMC Nunez et al. (2019). The model was individually fitted for each subject using choice-RT and accuracy only and accounted for between-condition variability within subjects. Clear clusters of drift rates and non-decision-time estimates depending on the noise conditions can be seen, though boundary estimates are highly overlapped. It is worth noting that uncertainties in the estimates can be inspected from the figures through the posterior covariance. Understandably, the uncertainties in the estimations from choice-RTs are significantly higher than from EEG signals, which agree with the theoretical derivations in Section 5.2.3. Figures 5.2b and 5.2d also demonstrate a satisfactory generalization to unseen data. The drift rates positively correlate with choice-RTs, whereas the boundaries and non-decision time negatively correlate with choice-RTs. The model successfully learns to extract the neural features that account for the choice-RT variability at each trial. To evaluate whether obtaining trial estimates of cognitive parameters improved the model of choice and choice-RT data, Table 5.1 presents the Wiener likelihood test for the neurocognitive generalization ability to unseen data. The results show that the use of single-trial predictions of cognitive parameters $\boldsymbol{\omega}_i$ provides higher likelihood than the median estimates $\bar{\boldsymbol{\omega}}$ fitted from the training data. This implies that single-trial estimates better account for new data compared to median estimates.

Figure 5.4a shows the average of signals generated by the neurocognitive autoencoder when given a set of approximately 800 test choice-RTs compared to the average of actual signals associated with the same choice-RTs. At the selected electrodes, the window of interest is 100 ms pre-stimulus to 500 ms post-stimulus, which captures the N200 waveform. The generated and original signals appear visually similar in the timing and amplitudes of the peaks and troughs. Figures 5.4b, 5.4c, and 5.4d depict the trial-averaged frequency spectra and corresponding ERP waveforms of the reconstructed signals. Regarding the frequency spectra, the most important features are the 30 and 40 Hz peaks, which correspond to the flicker frequency of the signal (Gabor patch) and spatial white noise, respectively. Interestingly, the generative model learns to structure output the steady-state visually evoked potentials (SSVEPs) that occur in response to a visual stimulus flickering at different frequencies, even though it was never explicitly encoded in the model. Moreover, in the low noise condition (b), the 30 Hz peak is large and the 40 Hz is small, while in the high noise condition (d), the 30 Hz peak is reduced and the 40 Hz peak is enhanced. In terms of ERP waveforms, the model captures the relationships of the N200 peak latencies with respect to the additive noise conditions. Higher additive white noise in the stimulus effectively increases the latency and decreases the amplitude of the N200. We focus on the N200 signal because the original study (Nunez et al., 2019) found strong relationships between N200 latency and choice-RT, and thus the N200 is a good validation of our model. These prove the convergence of the model in optimizing the lower bound of the conditional likelihood mapping from behavioral data to EEG features, which effectively encodes differences in the stimuli presented to the subjects in the latent variable space.

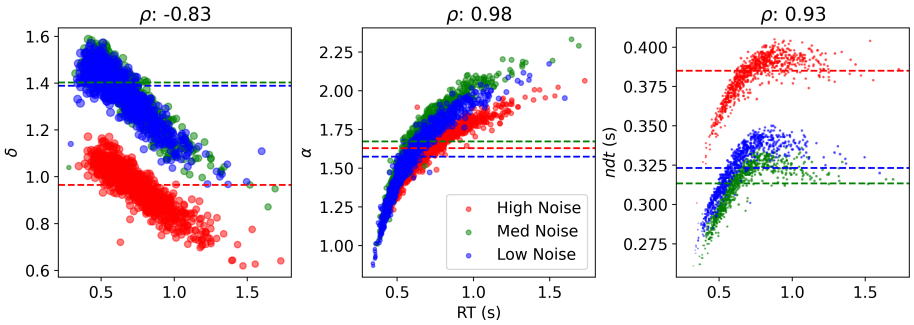
In addition to evaluating traditional ERP estimates (trial-averaged), we also assess the single-trial ERP estimate (channel-averaged). To increase the signal-to-noise ratio to better detect the N200, the first singular-value decomposition (SVD) component obtained from the ERP response is taken as a channel weighting function. More details of the SVD method can be seen at (Nunez et al., 2019). Figure 5.6 shows the performance of the model in learning the

N200 feature in each trial. As shown in Figure 5.6, the distributions of the single-trial N200 peak latencies, as well as the amplitudes calculated from the generated signals, closely match those of the original signals at three different noise levels. The peak amplitude distribution is somewhat broader than the original data's generated distribution. Importantly, the model can generate the variability of the N200 latency with the experimental manipulation of low, medium, and high noise, systematically increasing the N200 latency in the generated signals.

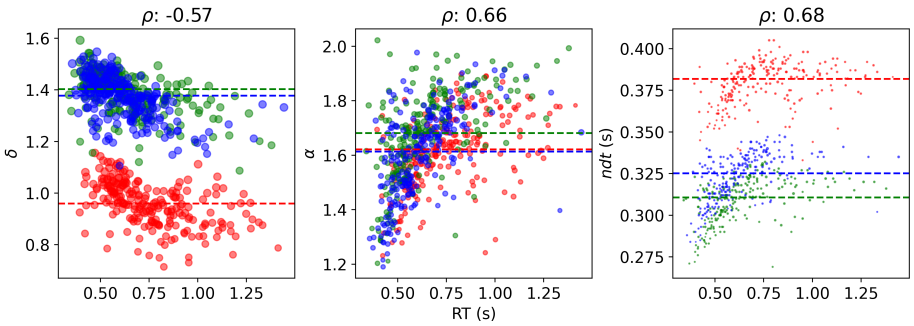
Figure 5.8 represents the sensitivity analysis of the choice-RT and drift-diffusion parameters regardless of the noise conditions. In the left column, we examine the sensitivity of the neural signals generated by the choice-RTs. We can see similar patterns across subjects where the increases in choice-RTs lead to significant declines in the 30 Hz and the rises of the N200 latencies. This confirms the minimization approach of the KL divergence between the latent spaces inferred from the behavioral data and the neural signals. Power at 40 Hz reflecting the neural response to the noise also changes according to the choice-RTs, though the pattern is not as strong as the subjects suppressed the noise signal in all conditions.

One of the powerful tools for exploring the relationship between cognitive processes is to examine the sensitivity of neural signals to cognitive parameters. The middle and right columns of Figure 5.8 depict the effect of *hypothetical* modulations of drift rates and non-decision time on the generated neural signals. The results show that our model reveals the intricate interactions between cognitive parameters and neural signals, which is consistent with prior discoveries in the cognitive modeling literature. As the non-decision time is faster, the N200 latencies are shorter, and the 30 Hz peaks are larger. Accordingly, the amplitudes of the N200 peaks are more prominent, though not shown in the figures for clarity. The same interactions are observed with the increase in drift rates, representing evidence accumulation. Again, the effects on 40 Hz peaks are weaker and depend on the subjects. We did not observe the effects of the boundary separation (caution) on the neural signals. The effect can be reversed with slower non-decision times and lower drift rates. The strongest effects can be

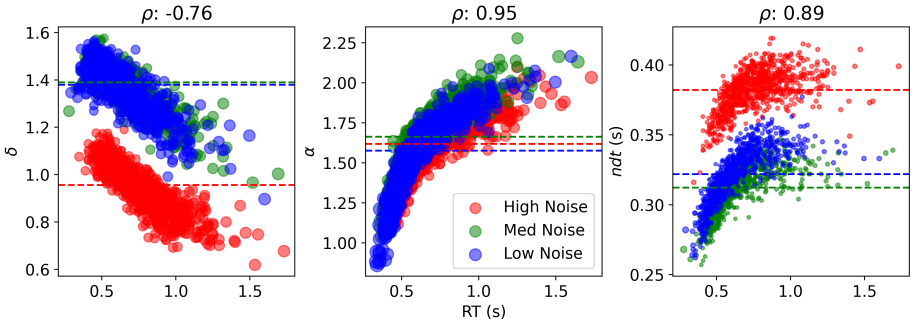
seen when both parameters influence neural signals. This demonstrates the effectiveness of the designs of the hierarchical latent variables inferred from choice-RTs and the disentangled latent space produced by the EEG data.



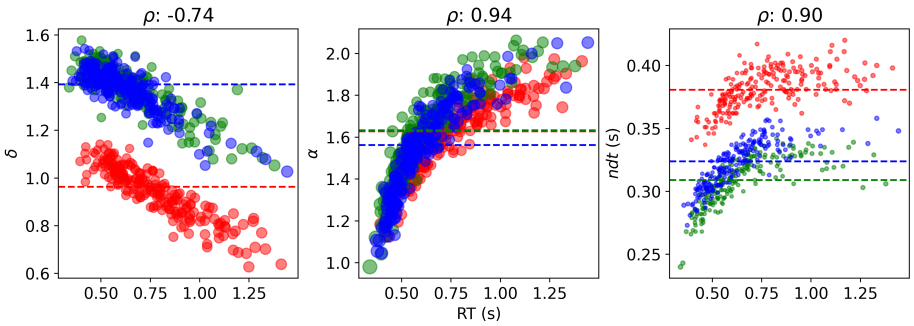
(a) Fitted from EEG (training data)



(b) Predicted from EEG (test data)

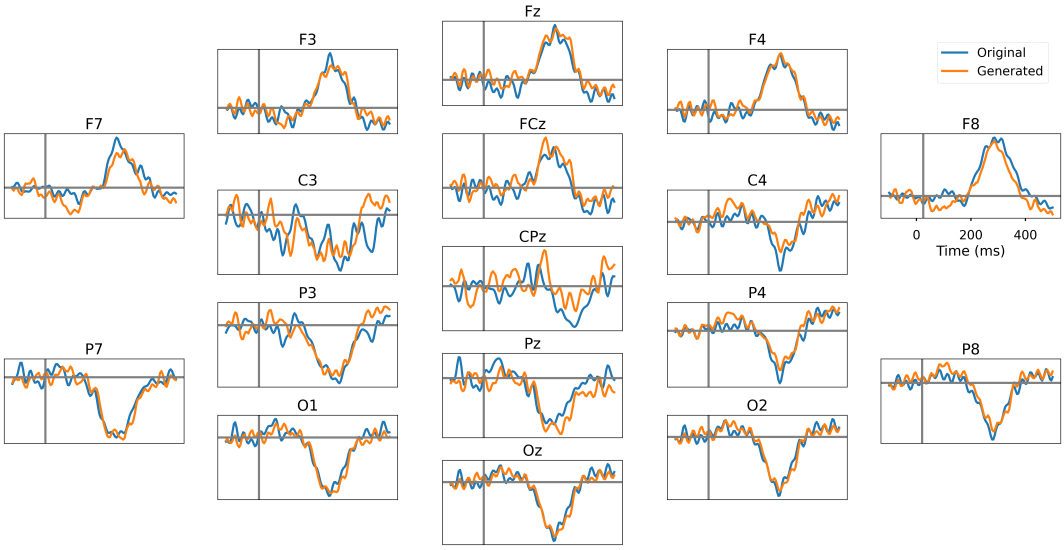


(c) Fitted from choice-RTs (training data)

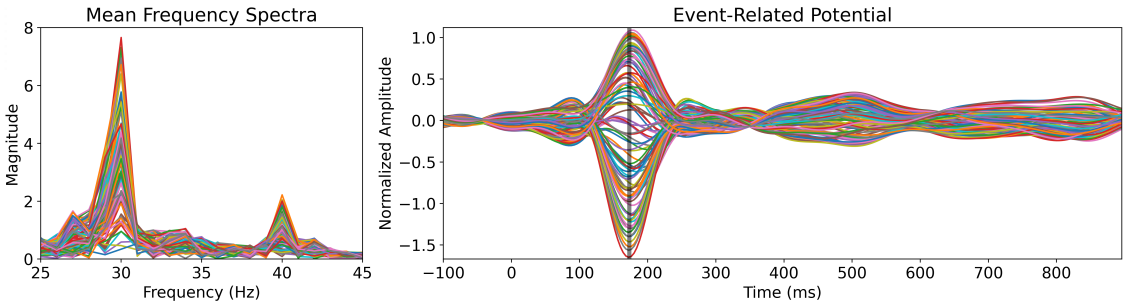


(d) Predicted from choice-RTs (test data)

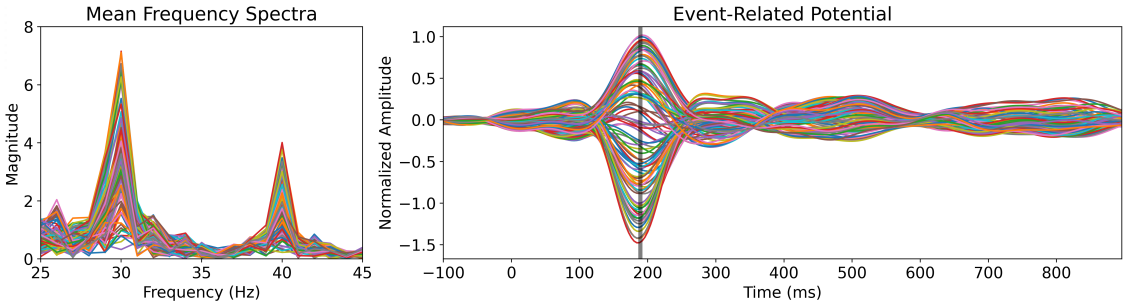
Figure 5.3: Drift-diffusion single-trial parameter estimations from correct responses of subject s1. The parameters are constrained by the subject priors resulting from a Bayesian MCMC modeling (without EEG data). Scatter plots illustrate the relationship between the parameters and the observed choice-RTs for each trial. The top two rows are posterior inferences from neural signals, while the bottom two are from behaviors. The left column shows the drift-rate (δ) estimates, the middle column shows boundary (α) estimates, and the right column presents non-decision time (ndt) estimates. The correlations between the choice-RTs and the inferred DDM parameters are consistent with what is expected. On top of each panel are the Spearman correlation coefficients (ρ). The covariances of the inferred parameters are indicated by circles, which correspond to contours having one standard deviation. For clarity, each circle is magnified 300 times.



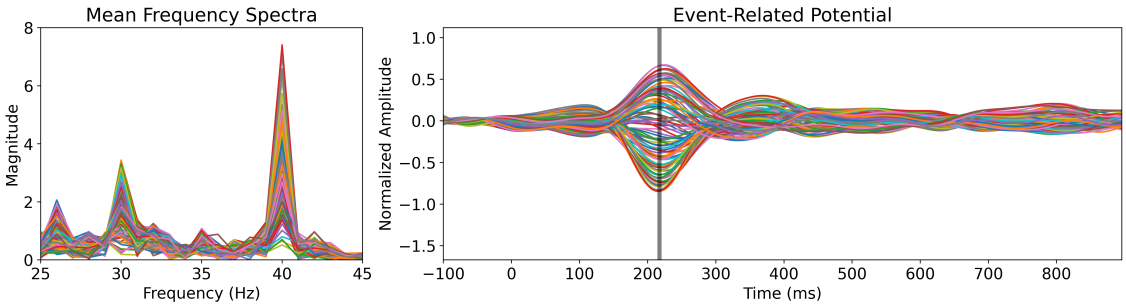
(a) EEG data at the selected electrodes



(b) Low noise condition

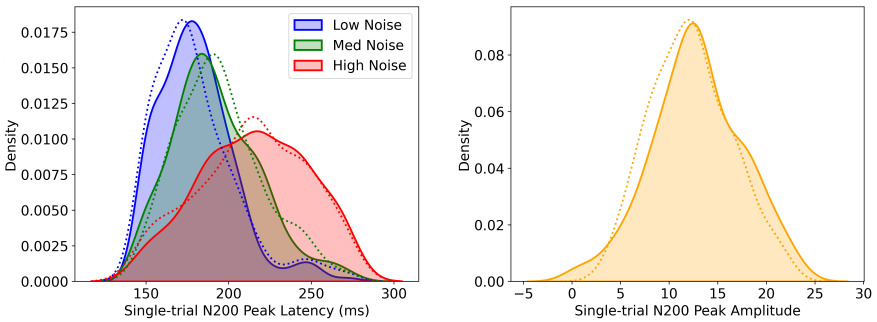


(c) Medium noise condition

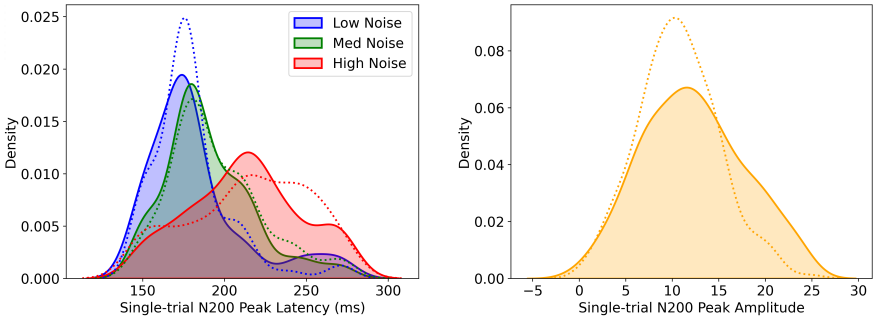


(d) High noise condition

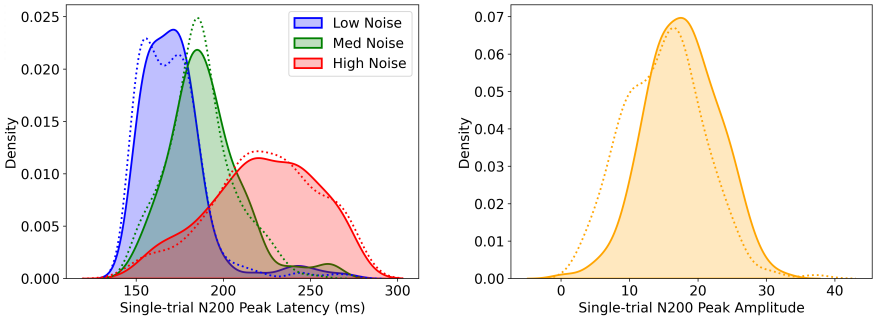
Figure 5.5: Performance of the model in reconstructing 98 EEG channels of subject s1 by averaging ≈ 800 predicted EEG trials from ≈ 800 choice-RTs in the test set. Time point zero denotes the time point of stimulus onset. The first row displays the original (blue) and generated (orange) trial-averaged EEG data at the pooled electrodes. The x-axis denotes the time in milliseconds from stimulus onset, and the y-axis denotes the signal amplitude. The second, third, and fourth rows are (left) frequency spectra and (right) EEG signals averaged over all test choice-RT trials ($\approx 800/3$ per condition). The signals on the right are low-pass filtered at 15 Hz for clarity of N200 peaks. Each colored line corresponds to one reconstructed EEG channel. In low-noise conditions, the spectra show a strong peak at the Gabor flicker frequency of 30 Hz, and the ERP waveform shows a shorter N200 latency and larger peak amplitude. Under high-noise conditions, the spectra show a strong peak at the noise flicker frequency of 40 Hz, and the ERP waveform shows a longer N200 latency and a smaller peak amplitude.



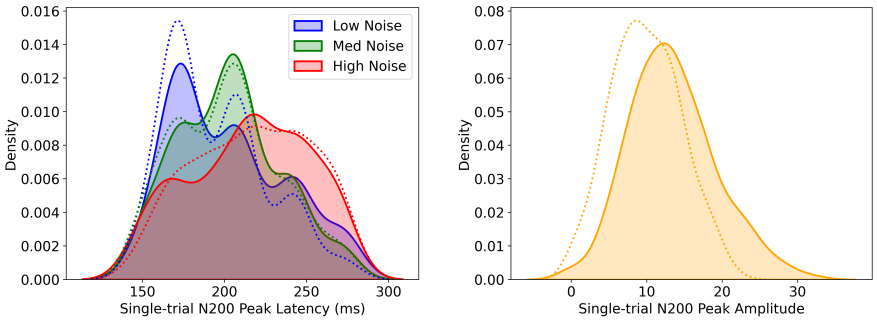
(a) Subject s1



(b) Subject s2

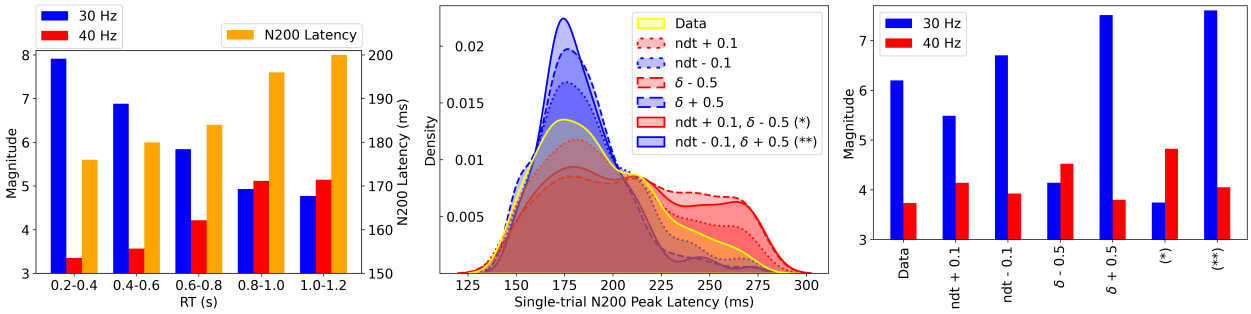


(c) Subject s3

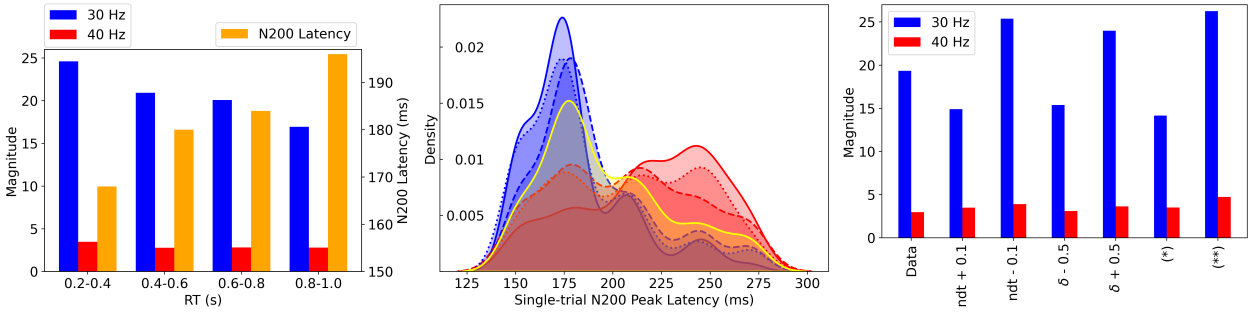


(d) Subject s4

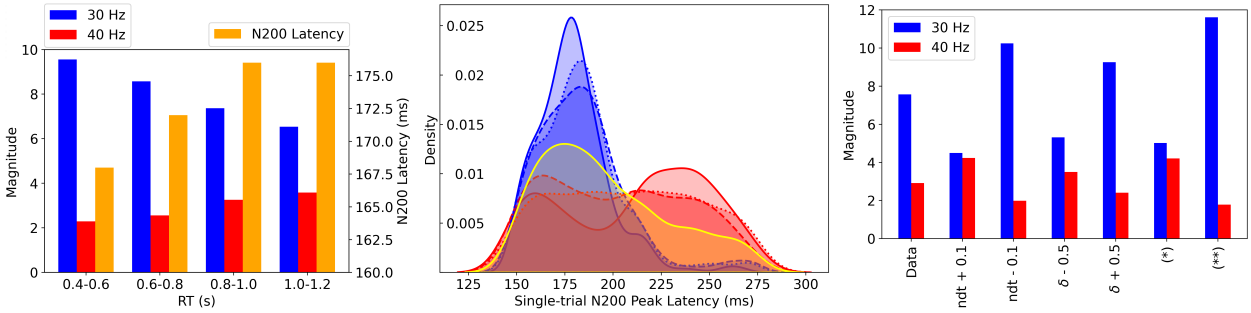
Figure 5.6: Performance of the model in reconstructing single-trial N200 peaks from choice-RTs in four subjects. The dotted lines are references to the original data. The distributions of (left) single-trial N200 peak latencies across three noise conditions and (right) the N200 peak amplitude statistics are shown. Single-trial observations of the peak latency of N200 are found using the SVD method (Nunez et al., 2019) for each subject and noise condition.



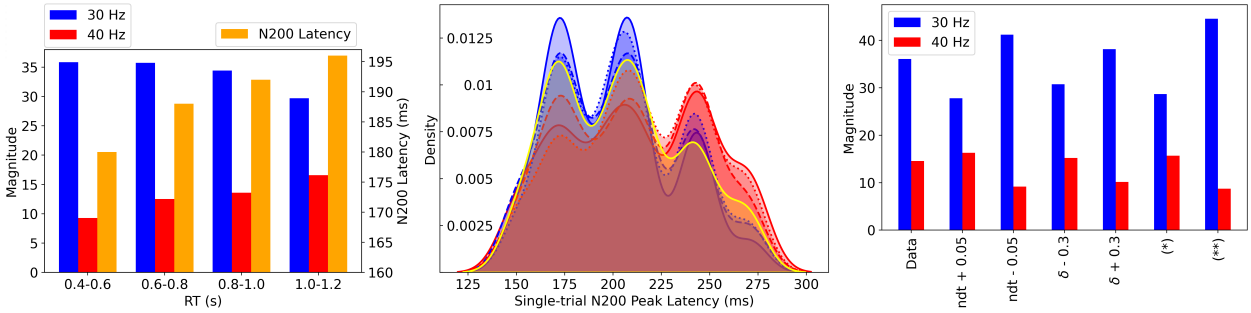
(a) Subject s1



(b) Subject s2



(c) Subject s3



(d) Subject s4

Figure 5.8: Sensitivity analysis of choice-RTs and latent drift-diffusion parameters on EEG signal generation in four subjects. The left column presents the effects of choice-RTs on the output neural signals. The blue bars represent the power at 30 Hz, while the red bars represent the power at 40 Hz. The orange bars show the N200 latencies. The middle column shows the changes in the single-trial N200 distribution w.r.t to *hypothetical* changes in the cognitive parameters. The yellow distribution represents the reference data, while the blue and red ones correspond to modified parameter settings that decrease or increase the N200 latencies, respectively. The modification in subject s4 (ndt $\pm$ 0.05, $\delta\pm$ 0.3) is different from other subjects. The right column characterizes the changes in 30 Hz and 40 Hz peaks w.r.t to the changes in the same cognitive parameters.

5.4 Conclusion

In this work, we proposed a joint behavioral and EEG modeling approach driven by a cognitive model of decision making. The experimental results demonstrate the effectiveness of our Neurocognitive VAE in simultaneously modeling high-dimensional EEG signals and low-dimensional behavioral data. Remarkably, the model learns essential task-relevant neural features, e.g. N200 peaks and SSVEP, without explicit specification in the optimization objective. Furthermore, the model captures how these features modulate behavior, specifically discovering relationships between brain activity and behavior consistent with other models based on prior knowledge. This suggests that the Neurocognitive VAE helps uncover neural signals linked to behavioral data by mapping to a structured latent space. Compared to the aforementioned published joint models (Nunez et al., 2015, 2017, 2019; Lui et al., 2021; Turner et al., 2013, 2016), our end-to-end model is capable of inferring task-relevant EEG features from behavior without prior knowledge of which features to optimize. The structured latent space allows the learning of behavioral variability to drive the EEG data generation process, leading to the prediction of the structure of EEG features in relation to the stimuli used in the experiments (N200 and SSVEP) and the behavioral performance (choice-RT). In addition, the model allows us to directly map the variability of cognitive parameters to neural signals, allowing for theoretical predictions that guide future experimental

studies. It should be noted that our framework does not serve to refine the functional form of process-oriented computational models. Instead, it presumes a set of fixed assumptions; in the DDM, a constant drift rate and boundary separation within trials. Importantly, our framework can be generalized to encompass any other neural measures combined with any cognitive model to explain behavior, provided that the cognitive model expresses a closed-form likelihood of behavioral data. Importantly, by parameterizing the likelihood by a deep neural network receiving neural data as input, trial-level parameter inferences are made possible. In this research, we assume a DDM posterior with a diagonal covariance matrix. This could lead to an overestimation of the variance of the marginal posteriors if the true posterior has dependencies. It would be beneficial to investigate the use of a full covariance matrix as an alternative. It is important to mention that our validation process focused on correct responses. Due to the low number of incorrect responses compared to correct ones, we lack confidence in interpreting the results in this study for the incorrect trials, although the direction of the trial-level parameter fits was consistent with the results for correct trials. We anticipate future research to explore strategies to address the class imbalance problem in deep learning models (Johnson and Khoshgoftaar, 2019). Further work with a larger dataset is needed to demonstrate that we can extend the model to new individuals. In principle, this would potentially allow us to predict brain activity in clinical populations with known behavioral differences.

Data and Code Availability Statement

The dataset analyzed during the current study is available on `https://zenodo.org/record/8381751`, and the implementation of the model is in the following repository `https://github.com/khuongav/neurocognitive_vae`.

5.5 Supplementary Materials

5.5.1 Neural Network Architectures and Training Hyperparameters

The inferential and generative processes are parameterized by deep neural networks, as shown by the flows in Figure 5.1. Table 5.2 details the architectures of the five networks. The input EEG signals are of size 98 x 250 (1 second of data of 98 channels at 250 Hz). The feature extraction layers in the EEG and cognitive encoders are similar to Vo et al. (2022). All the feature maps have 128 channels. Leaky ReLU (lReLU) activation functions are applied to all layers, with a slope of 0.1 to stimulate easier gradient flow. Batch normalizations (BN) (Ioffe and Szegedy, 2015) are used in each convolutional layer of the encoders and decoders. Self-attention layers (Zhang et al., 2019) are applied in the encoders and decoders to better account for long-range relationships in time series. $\boldsymbol{c}$ are noise condition embeddings as one-hot vectors (size 3). The size of $\boldsymbol{z}_N$ is set at 32 as increasing the dimension did not lead to any improvement in performance on a validation set.

In Equation 5.4, the term $\log p(\boldsymbol{y} \mid \boldsymbol{z}_C)$ is weighted by $\lambda = 2$ to scale up the likelihood of low-dimensional behavior. The KL terms are weighted by $\beta = 20$. The KL terms are normalized to balance the KL divergence loss and the reconstruction loss. Please refer to Sections 4.2 and A6 of (Higgins et al., 2016) for further information. The optimization of $q_{\phi_C}(\boldsymbol{z}_C \mid \boldsymbol{x})$ is divided into two stages. We first optimize the network w.r.t drift rate δ and boundary α, while non-decision time τ is set to $0.93 \cdot RT_{min}$ for each subject, approximating the results of the Bayesian MCMC modeling Nunez et al. (2019). Having trained ϕ_C for δ and α, we can proceed to train only the last fully connected layer that predicts τ. This procedure is to circumvent the difficulty of simultaneously optimizing the network for the boundary and the non-decision time on the experimental data. We used Adam (Kingma and Ba, 2014) for optimizations, with a learning rate of 5e-4 and exponential decay rates $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

Table 5.2: Neural network parametrization

Encoder$_{\mathbf{N}}$ – $q_{\phi_N}(\boldsymbol{z}_N \mid \boldsymbol{x})$ **maps EEG signals to neural latents**		**Encoder$_{\mathbf{C}}$ – $q_{\phi_C}(\boldsymbol{z}_C \mid \boldsymbol{x})$** **maps EEG signals to cognitive latents**	**Decoder - $p_\theta(\boldsymbol{x} \mid \boldsymbol{z}_N, \boldsymbol{z}_C)$** **reconstructs EEG signals**
Dropout(0.3) Conv 1, lReLU, 128 x 250		Conv 1, lReLU, 128 x 250 Conv 6, BN, lReLU, Dropout(0.7) Conv 6, Stride 2, BN, lReLU, Dropout(0.7)	Get $\boldsymbol{z}_C$ Linear 128, lReLU Linear 32, lReLU Concat $\boldsymbol{z}_N$, $\boldsymbol{c}$ Conv Transp 8, Stride 4, 512 Channels, BN, lReLU Conv Transp 8, Stride 4, 256 Channels, BN, lReLU Self Attention Conv Transp 6, Stride 3, 128 Channels, BN, lReLU Conv Transp 6, Stride 3, 128 Channels, BN, lReLU Self Attention Conv Transp 10, Stride 2, 98 Channels
Conv 6, BN, lReLU Conv 6, Stride 2, BN, lReLU	X 2		
Self Attention		Self Attention	
Conv 6, BN, lReLU Conv 6, Stride 2, BN, lReLU	X 2	Conv 6, BN, lReLU, Dropout(0.7) Conv 6, Stride 2, BN, lReLU, Dropout(0.7)	
Reshape 2048, Concat $\boldsymbol{c}$ Linear 32 (mean $\boldsymbol{z}_N$) Linear 32 (logvar $\boldsymbol{z}_N$)		Reshape 2048, Concat $\boldsymbol{c}$ Linear 1 (mean δ), Linear 1 (logvar δ) Linear 1, Softplus (mean α) Linear 1 (logvar α) Linear 1, Softplus (mean ndt) Linear 1 (logvar ndt)	

Encoder$^2_\beta$ – $q^2_\beta(\boldsymbol{z}_N \mid \boldsymbol{z}_C)$ **maps cognitive latents to neural latents**		**Encoder$^1_\beta$ – $q^1_\beta(\boldsymbol{z}_C \mid \boldsymbol{y}_i)$** **maps behaviors to cognitive latents**
Linear 128, lReLU Linear 128, lReLU Concat $\boldsymbol{c}$ Linear 64		Linear 128, lReLU Linear 128, lReLU Concat $\boldsymbol{c}$ Linear 6

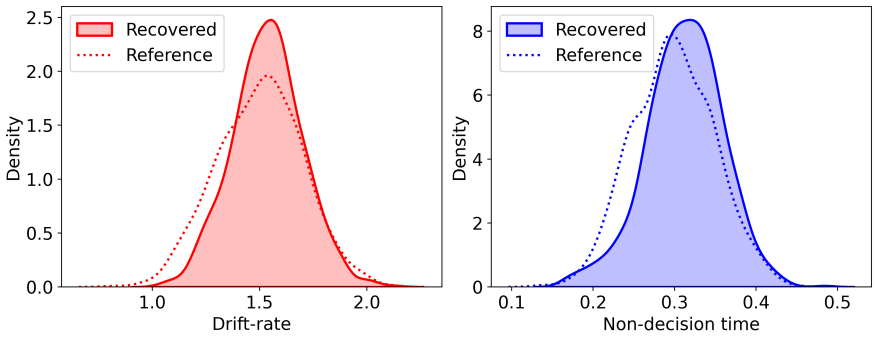
5.5.2 Simulation Studies

We assessed our ability to recover true non-decision time (NDT) and drift rate by simulating response time data and EEG signals. Response time data were simulated from a drift-diffusion model with trial-to-trial variability in NDT and evidence accumulation rate (i.e., drift rate). To simulate EEG signals with a known relationship with DDM parameters, we specifically focused on N200 due to the significant associations between N200 latency and NDT reported by Nunez et al. (2019). In our new experiments, we additionally observed a substantial relationship between drift rate and N200 latency, which we included in the simulation. Boundary separation was not included in the simulation, as we did not find any neural correlates of variability in boundary separation, and those are usually only found in tasks with trial-level accuracy feedback (Cavanagh and Frank, 2014; Nunez et al., 2024).

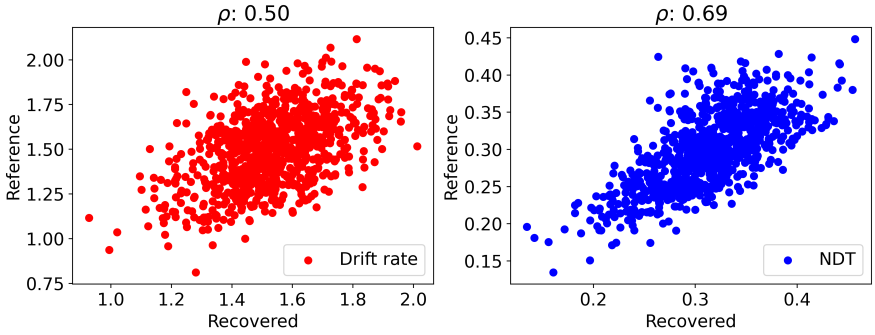
To simulate single-trial EEG signals, we shifted the true averaged ERP waveform based on each sample of trial-level NDT, using a linear regression slope of 1, as in Nunez et al. (2019). EEG noise was obtained from the original data, using independently sampled segments that did not include responses to stimuli. The resulting ERP and EEG waveforms were then combined to generate artificial EEG signals for each trial that carried the N200 latency

information and was associated with choice and response time.

It is evident from the results in Figure 5.9 that the model can accurately recover the original distributions of trial-specific parameters. In particular, the generating and recovered distributions strongly overlap, and the correlation plots indicate that our single-trial estimates of cognitive parameters exhibit good correlations with the reference parameters.



(a) Parameter distributions



(b) Parameter correlations

Figure 5.9: Drift-diffusion parameter estimates from neural signals in a simulation of trial-level choice RTs and EEG signals. The top panels show the overlap between the recovered and the original distributions of trial-specific drift-rate and NDT. The reference values for the drift rate and NDT are drawn from the normal distributions $\mathcal{N}(1.5, 0.2)$ and $\mathcal{N}(0.3, 0.05)$, respectively. The bottom scatter plots illustrate the relationship between the recovered parameters and the original parameters each trial. ρ are the Spearman correlation coefficients.

5.5.3 Experimental Tasks

Nunez et al. (2019) incorporated data from two experiments to test the hypothesis that N200 peak-latencies track Visual Encoding Time (VET). Both experiments required participants

to determine whether a Gabor stimulus had high or low spatial frequency content. The tasks took place in a dark room with participants fixating on a small spot while responding to the stimuli presented on a 61 cm LED monitor.

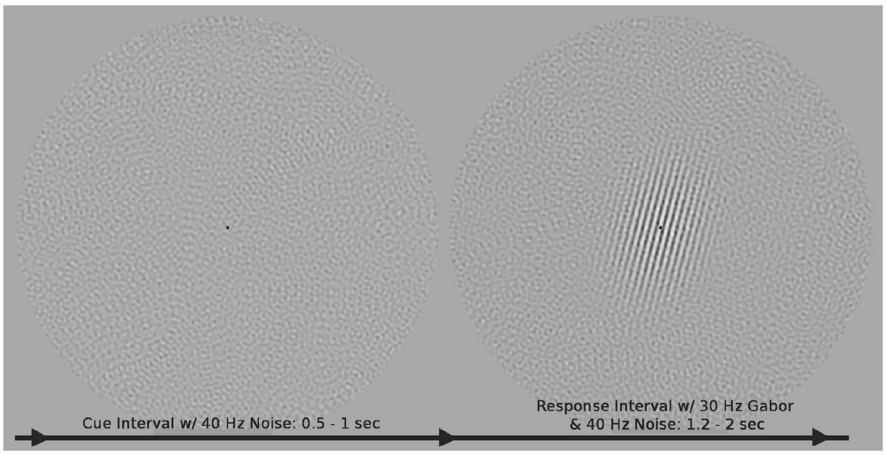


Figure 5.10: Example stimuli of the cue and response intervals of medium noise conditions (Nunez et al., 2019). In the response phase, participants identified the spatial-frequency target represented by each Gabor, using their left hand to press a button for a target with a low spatial frequency (2.4 cpd) and their right hand for a target with a high spatial frequency (2.6 cpd). N200 waveforms were calculated time-locked to the onset of the Gabor stimulus during the response intervals. The visual noise altered at a frequency of 40 Hz, while the Gabor signal modulated at 30 Hz, inducing 40 Hz and 30 Hz electrocortical responses that monitor attention to both noise and signal.

In these experiments, Gabors were sinusoidal grating patterns with a Gaussian falloff of contrast. The high and low spatial frequencies of the target Gabors were 2.4 and 2.6 cycles per degree visual angle (cpd) respectively. The experiments involved three conditions of visual noise contrast: high, medium, and low. Visual noise was displayed both before and concurrently with the Gabor targets at regular intervals. Example stimuli are given in Figure 5.10. Participants used a button box to respond, pressing with the left hand for low spatial frequency targets and the right hand for high spatial frequency targets. They maintained fixation on a central spot while identifying the spatial frequency of the Gabor stimuli embedded in noise.

EEG data was recorded using a 128-channel Geodesic sensor net. The visual noise changed at 40 Hz, and the Gabor signal flickered at 30 Hz, evoking specific electrocortical responses.

The primary objective was to assess whether N200 peak-latencies recorded by EEG reflected VET across varying visual noise conditions, thereby shedding light on the timing involved in perceptual decision-making processes.

5.5.4 Decision-Making Models - The Drift-Diffusion Model (DDMs)

The Drift-Diffusion Model (DDM) is a sequential sampling model of decision making. The model assumes that decision-making is the result of the accumulation of evidence in favor of one option or another. The evidence is represented by a random walk process, where the evidence accumulates over time, and the decision is made when the accumulated evidence exceeds a threshold.

Mathematically, the DDM is described by a set of equations that govern the accumulation of evidence and the decision-making process. The basic equation for the DDM following a Wiener process (Figure 5.11) is

$$dx = \delta dt + \varsigma dW \tag{5.9}$$

where dx is the evidence step, dt is the time step and ςdW denotes a Gaussian noise with a scale ς. The drift rate δ and the diffusion coefficient ς are parameters that describe the average rise in change over a unit interval during evidence accumulation and the instantaneous variation in the rate of change, respectively. The variance can be fixed at 1 for mathematical explicitness and simplicity. The distance between two options is described as the boundary separation, or α. The beginning position of evidence accumulation, which shows a bias toward one of the two options, is encoded by the parameter β. When β is 0.5, the beginning point is halfway between the two borders, and the evidence-building process can begin unbiasedly between the two options. Visual encoding time prior to evidence accumulation and motor execution time following evidence accumulation could be written as τ_e and τ_m,

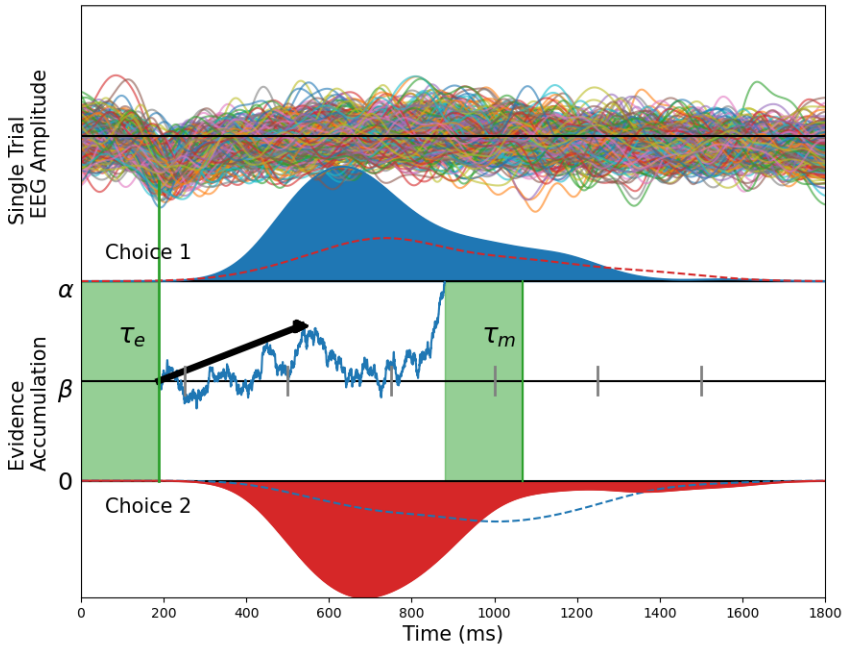


Figure 5.11: The DDM is illustrated in action during a two-choice task, with non-decision time shown in green. Following the visual encoding period, the decision variable (DV) begins evidence accumulation and reaches either the upper or lower limit for each trial. The black vector depicts the average rate of evidence accumulation. The blue curve depicts the distribution of response times when choice 1 is successfully picked, while the red curve depicts the distribution of reaction times when choice 2 is correctly selected. When the DV drifts towards the incorrect boundary owing to random noise, incorrect decisions are made. The distribution of reaction times for incorrect trials is depicted by the dotted curve. EEG data for each trial, processed using singular value decomposition to highlight N200, is shown on top that track the start of evidence accumulation.

respectively. Only the total of the two processes, non-decision time τ, can be observed with behavior alone.

The probability density function (pdf) of the Wiener diffusion model is bivariate (with one dimension for the latency (t) and one for the binary choice (c)); its analytical form can be

approximated as

$$\begin{cases} \text{Wiener } (t, c = 0; \alpha, \beta, \tau, \delta) \\ = \dfrac{\pi}{\alpha^2} e^{-\frac{1}{2}\left(2\alpha\beta\delta + \delta^2(t-\tau)\right)} \times \displaystyle\sum_{k=1}^{+\infty} \left[k \sin(\pi k \beta) e^{-\frac{1}{2}\frac{k^2\pi^2}{\alpha^2}(t-\tau)} \right] \\ \text{Wiener } (t, c = 1; \alpha, \beta, \tau, \delta) \\ = \text{ Wiener } (t, c = 0 \mid \alpha, 1-\beta, \tau, -\delta) \end{cases} \tag{5.10}$$

Efficient methods for the computation of the Wiener diffusion model density and distribution functions exist (Navarro and Fuss, 2009), making it a highly tractable model. In this work, β is set at 0.5, so that the starting point is always unbiased at $z = \beta\alpha$.

Chapter 6

Conclusions

6.1 Contributions

The growing accessibility of medical time series data is propelling the advancement of mathematical models and techniques that can analyze it broadly and efficiently. This dissertation is a move towards this goal, motivated by the recent achievements in probabilistic modeling and deep learning.

In this dissertation, we have constructed integrated frameworks that combine concepts from latent variable models, state-space models, and deep learning to model multidimensional dependencies in physiological signals. These frameworks are capable of modeling complex data distributions across various applications. This is accomplished through the development of probabilistic models that utilize deep neural networks to parameterize the underlying conditional distributions. Deep learning architectures serve as powerful function approximators, enabling the model to automatically extract features essential for wide applicability. Recent developments in deep learning can be seamlessly integrated into this framework.

The approaches proposed in this dissertation offer versatile frameworks for representing and learning from diverse types of physiological measures. By efficiently processing unlabeled datasets, these models enable the discovery of hidden structures and patterns, facilitating

data-driven hypothesis generation:

1. *Deep State-Space Model for Heart Electrical Waveforms* (Vo et al., 2023). This application has significant potential for clinical diagnoses, especially since it allows for heart disease assessment through wearable devices. The use of optically obtained signals as inputs adds to the innovation, potentially making diagnosis simpler and more accessible.

2. *Brain Signal Modeling with Probabilistic Graphical Models and Deep Adversarial Learning* (Vo et al., 2022). Combining these two approaches is promising for encoding neural oscillations' complexity while maintaining interpretability. Moreover, applying these techniques to epilepsy seizure detection as an unsupervised learning problem could lead to earlier and more accurate diagnoses, improving patient outcomes.

3. *Joint Modeling of Physiological Measures and Behavior* (Vo et al., 2024). This approach shows potential in tackling the modern challenge in amalgamation of diverse medical data sources. By analyzing the relationship between physiological measures and behavior, our method could uncover new insights into brain function and potentially revolutionize our understanding of neurocognitive processes.

6.2 Future Work

Every progression in a field brings forth a set of unanswered queries, usually more complex than the ones preceding them. Regarding the deep latent variable models (DLVMs) presented in this thesis, several open questions exist, the resolution of which could enhance our comprehension of their working principles as well as to better exploit their modeling power.

- Ensuring patient safety when implementing medical machine learning methods for clinical applications necessitates robust models. A key characteristic of robust models is their resilience to out-of-distribution (OOD) data, meaning they can still provide

accurate predictions when encountering data that differ from the training set. Such OOD data might include samples from varied patient demographics, different medical equipment and laboratory methods. Future research should focus on assessing the adaptability of the DLVMs to OOD data and enhancing model resilience against such data by leveraging established strategies in the field (Zhou et al., 2022; Wang et al., 2022).

- Probabilistic graphical models offer a principled approach to incorporating prior knowledge and structured frameworks into the model, utilizing current message-passing algorithms for approximate inference. A crucial yet difficult task for broader adoption of DLVMs is developing a message-passing library that seamlessly works with existing deep learning libraries (Johnson et al., 2016b; Bendekgey et al., 2024).

- Choosing the optimal model parameterization for a specific application can be challenging. This category of models derives all the complexities associated with defining the precise network architecture, such as the number of layers, units, and activation functions, from its deep learning components. Therefore, identifying a systematic method for hyperparameter optimization that is effective across different applications is essential (Yu and Zhu, 2020; He et al., 2021).

- Recent progress in physics-informed deep learning (Raissi et al., 2019; Nabian and Meidani, 2020) integrates the advantages of deep learning methods with physical principles to improve both model efficacy and generalization. In this approach, deep learning models are enhanced with a regularization term that serves as prior knowledge, reflecting the fundamental laws and penalizing deviations from these governing equations. Investigating DLVM approaches that adhere to any specified law of electrophysics, as characterized by stochastic differential equations, would be advantageous.

Bibliography

Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiol Meas.*, 28(3):1–39.

Aschbacher, K., Yilmaz, D., Kerem, Y., Crawford, S., Benaron, D., Liu, J., Eaton, M., Tison, G. H., Olgin, J. E., Li, Y., et al. (2020). Atrial fibrillation detection from raw photoplethysmography waveforms: A deep learning application. *Heart rhythm O2*, 1(1):3–9.

Aznan, N. K. N., Atapour-Abarghouei, A., Bonner, S., Connolly, J. D., Al Moubayed, N., and Breckon, T. P. (2019). 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.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Banerjee, R., Sinha, A., Choudhury, A. D., and Visvanathan, A. (2014). Photoecg: Photoplethysmographyto estimate ecg parameters. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4404–4408. IEEE.

Bau, D., Zhu, J.-Y., Wulff, J., Peebles, W., Strobelt, H., Zhou, B., and Torralba, A. (2019). Seeing what a gan cannot generate. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4502–4511.

Begoli, E., Bhattacharya, T., and Kusnezov, D. (2019). The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23.

Bendekgey, H. C., Hope, G., and Sudderth, E. (2024). Unbiased learning of deep generative models with structured discrete representations. *Advances in Neural Information Processing Systems*, 36.

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. (2015). Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45.

Cavanagh, J. F. and Frank, M. J. (2014). Frontal theta as a mechanism for cognitive control. *Trends in cognitive sciences*, 18(8):414–421.

Chiu, H.-Y., Shuai, H.-H., and Chao, P. C.-P. (2020). Reconstructing qrs complex from ppg by transformed attentional neural networks. *IEEE Sensors Journal*, 20(20):12374–12383.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

Donahue, J., Krähenbühl, P., and Darrell, T. (2016). Adversarial feature learning. *arXiv preprint arXiv:1605.09782*.

Dumoulin, V., Belghazi, I., Poole, B., Mastropietro, O., Lamb, A., Arjovsky, M., and Courville, A. (2016). Adversarially learned inference. *arXiv preprint arXiv:1606.00704*.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., et al. (2021). Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8.

Forstmann, B. U. and Wagenmakers, E.-J., editors (2015). *An Introduction to Model-Based Cognitive Neuroscience*. Springer New York, New York, NY.

Glomb, K., Cabral, J., Cattani, A., Mazzoni, A., Raj, A., and Franceschiello, B. (2020). Computational models in electroencephalography. *arXiv preprint arXiv:2009.08385*.

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

Hartmann, K. G., Schirrmeister, R. T., and Ball, T. (2018). Eeg-gan: Generative adversarial networks for electroencephalograhic (eeg) brain signals. *arXiv preprint arXiv:1806.01875*.

He, X., Zhao, K., and Chu, X. (2021). Automl: A survey of the state-of-the-art. *Knowledge-based systems*, 212:106622.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2016). beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*.

Hoffman, M. D. and Johnson, M. J. (2016). Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, volume 1.

Hong, S., Xiao, C., Ma, T., Li, H., and Sun, J. (2019). Mina: multilevel knowledge-guided attention for modeling electrocardiography signals. *arXiv preprint arXiv:1905.11333*.

Hong, S., Zhou, Y., Shang, J., Xiao, C., and Sun, J. (2020). Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. *Computers in biology and medicine*, 122:103801.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr.

Itthipuripat, S., Sprague, T. C., and Serences, J. T. (2019). Functional MRI and EEG index complementary attentional modulations. *Journal of Neuroscience*, 39(31):6162–6179.

Jabbar, A., Li, X., and Omar, B. (2020). A survey on generative adversarial networks: Variants, applications, and training. *arXiv preprint arXiv:2006.05132*.

Jang, E., Gu, S., and Poole, B. (2016). Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Jansen, B. H. and Rit, V. G. (1995). Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns. *Biological cybernetics*, 73(4):357–366.

Jansen, B. H., Zouridakis, G., and Brandt, M. E. (1993). A neurophysiologically-based mathematical model of flash visual evoked potentials. *Biological cybernetics*, 68(3):275–283.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016a). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Johnson, J. M. and Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54.

Johnson, M. J., Duvenaud, D. K., Wiltschko, A., Adams, R. P., and Datta, S. R. (2016b). Composing graphical models with neural networks for structured representations and fast inference. *Advances in neural information processing systems*, 29.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Krishnan, R., Shalit, U., and Sontag, D. (2017). Structured inference networks for nonlinear state space models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Krishnan, R. G., Shalit, U., and Sontag, D. (2015). Deep kalman filters. *arXiv preprint arXiv:1511.05121*.

Lee, M. D. and Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge university press.

Lee, T.-Y., Vo, K., Baek, W., Khine, M., and Dutt, N. (2020). Stint: selective transmission for low-energy physiological monitoring. In *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*, pages 115–120.

Li, C., Welling, M., Zhu, J., and Zhang, B. (2018). Graphical generative adversarial networks. *Advances in neural information processing systems*, 31.

Lu, G., Yang, F., Taylor, J., and Stein, J. F. (2009). A comparison of photoplethysmography and ecg recording to analyse heart rate variability in healthy subjects. *Journal of medical engineering & technology*, 33(8):634–641.

Lui, K. K., Nunez, M. D., Cassidy, J. M., Vandekerckhove, J., Cramer, S. C., and Srinivasan, R. (2021). Timing of readiness potentials reflect a decision-making process in the human brain. *Computational Brain & Behavior*, 4(3):264–283.

McSharry, P. E., Clifford, G. D., Tarassenko, L., and Smith, L. A. (2003). A dynamical model for generating synthetic electrocardiogram signals. *IEEE transactions on biomedical engineering*, 50(3):289–294.

Minka, T. P. (2013). Expectation propagation for approximate bayesian inference. *arXiv preprint arXiv:1301.2294*.

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

Mohamed, S. and Lakshminarayanan, B. (2016). Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*.

Moody, B., Moody, G., Villarroel, M., Clifford, G., and Silva, I. (2020). Mimic-iii waveform database matched subset. *MIMIC-III Waveform Database Matched Subset v1. 0*.

Nabian, M. A. and Meidani, H. (2020). Physics-driven regularization of deep neural networks for enhanced engineering design and analysis. *Journal of Computing and Information Science in Engineering*, 20(1).

Navarro, D. J. and Fuss, I. G. (2009). Fast and accurate calculations for first-passage times in wiener diffusion models. *Journal of mathematical psychology*, 53(4):222–230.

Niederer, S. A., Lumens, J., and Trayanova, N. A. (2019). Computational models in cardiology. *Nature Reviews Cardiology*, 16(2):100–111.

Nunez, M. D., Fernandez, K., Srinivasan, R., and Vandekerckhove, J. (2024). A tutorial on fitting joint models of m/eeg and behavior to understand cognition. *PsyArXiv*.

Nunez, M. D., Gosai, A., Vandekerckhove, J., and Srinivasan, R. (2019). The latency of a visual evoked potential tracks the onset of decision making. *Neuroimage*, 197:93–108.

Nunez, M. D., Srinivasan, R., and Vandekerckhove, J. (2015). Individual differences in attention influence perceptual decision making. *Frontiers in psychology*, 8:18.

Nunez, M. D., Vandekerckhove, J., and Srinivasan, R. (2017). How attention influences perceptual decision making: Single-trial EEG correlates of drift-diffusion model parameters. *Journal of mathematical psychology*, 76:117–130.

Nunez, P. L. and Srinivasan, R. (2006). *Electric fields of the brain: the neurophysics of EEG*. Oxford University Press, USA.

Olier, I., Ortega-Martorell, S., Pieroni, M., and Lip, G. Y. (2021). How machine learning is impacting research in atrial fibrillation: implications for risk prediction and future management. *Cardiovascular Research*, 117(7):1700–1717.

Pascual, D., Aminifar, A., Atienza, D., Ryvlin, P., and Wattenhofer, R. (2019). Synthetic epileptic brain activities using gans. *Machine Learning for Health (ML4H) at NeurIPS*.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707.

Ratcliff, R. and McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, 20(4):873–922.

Reisner, A., Shaltis, P. A., McCombie, D., and Asada, H. H. (2008). Utility of the photoplethysmogram in circulatory monitoring. *Anesthesiology: The Journal of the American Society of Anesthesiologists*, 108(5):950–958.

Rosiek, A. and Leksowski, K. (2016). The risk factors and prevention of cardiovascular disease: the importance of electrocardiogram in the diagnosis and treatment of acute coronary syndrome. *Therapeutics and clinical risk management*, 12:1223.

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., and Faubert, J. (2019). Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, 16(5):051001.

Sarkar, P. and Etemad, A. (2021). Cardiogan: Attentive generative adversarial network with dual discriminators for synthesis of ecg from ppg. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 488–496.

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

Siddiqui, M. K., Morales-Menendez, R., Huang, X., and Hussain, N. (2020). A review of epileptic seizure detection using machine learning classifiers. *Brain informatics*, 7:1–18.

Sohn, K., Lee, H., and Yan, X. (2015). Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28.

sudden (Retrieved on September 2020). Sudden death in young people: Heart problems often blamed. `https://www.mayoclinic.org/diseases-conditions/sudden-cardiac-arrest/in-depth/sudden-death/art-20047571`.

Sun, Q. J., Vo, K., Lui, K., Nunez, M., Vandekerckhove, J., and Srinivasan, R. (2022). Decision sincnet: Neurocognitive models of decision making that predict cognitive processes from neural signals. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE.

Suzuki, M., Nakayama, K., and Matsuo, Y. (2016). Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*.

Tang, Q., Chen, Z., Guo, Y., Liang, Y., Ward, R., Menon, C., and Elgendi, M. (2022). Robust reconstruction of electrocardiogram using photoplethysmography: A subject-based model. *Frontiers in Physiology*, page 645.

Tian, X., Zhu, Q., Li, Y., and Wu, M. (2020). Cross-domain joint dictionary learning for ecg reconstruction from ppg. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 936–940. IEEE.

Tran, B.-H., Shahbaba, B., Mandt, S., and Filippone, M. (2023). Fully bayesian autoencoders with latent sparse gaussian processes. In *International Conference on Machine Learning*, pages 34409–34430. PMLR.

Turner, B. M., Forstmann, B. U., Wagenmakers, E.-J., Brown, S. D., Sederberg, P. B., and Steyvers, M. (2013). A bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage*, 72:193–206.

Turner, B. M., Rodriguez, C. A., Norcia, T. M., McClure, S. M., and Steyvers, M. (2016). Why more is better: Simultaneous modeling of EEG, fMRI, and behavioral data. *NeuroImage*, 128:96–115.

van Gent, P., Farah, H., van Nes, N., and van Arem, B. (2019). Analysing noisy driver physiology real-time using off-the-shelf sensors: Heart rate analysis software from the taking the fast lane project. *Journal of Open Research Software*, 7(1).

Van Gent, P., Farah, H., Van Nes, N., and Van Arem, B. (2019). Heartpy: A novel heart rate algorithm for the analysis of noisy signals. *Transportation research part F: traffic psychology and behaviour*, 66:368–378.

Vedantam, R., Fischer, I., Huang, J., and Murphy, K. (2017). Generative models of visually grounded imagination. *arXiv preprint arXiv:1705.10762*.

Vo, K., El-Khamy, M., and Choi, Y. (2023). Ppg to ecg signal translation for continuous atrial fibrillation detection via attention-based deep state-space modeling. *arXiv preprint arXiv:2309.15375*.

Vo, K., Naeini, E. K., Naderi, A., Jilani, D., Rahmani, A. M., Dutt, N., and Cao, H. (2021). P2e-wgan: Ecg waveform synthesis from ppg with conditional wasserstein generative adversarial networks. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, pages 1030–1036.

Vo, K., Sun, Q. J., Nunez, M. D., Vandekerckhove, J., and Srinivasan, R. (2024). Deep latent variable joint cognitive modeling of neural signals and human behavior. *NeuroImage*, page 120559.

Vo, K., Vishwanath, M., Srinivasan, R., Dutt, N., and Cao, H. (2022). Composing graphical models with generative adversarial networks for EEG signal modeling. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1231–1235. IEEE.

Wainwright, M. J., Jordan, M. I., et al. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., and Philip, S. Y. (2022). Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8):8052–8072.

Wu, W., Nagarajan, S., and Chen, Z. (2015). Bayesian machine learning: Eeg\/meg signal processing measurements. *IEEE Signal Processing Magazine*, 33(1):14–36.

Yu, T. and Zhu, H. (2020). Hyper-parameter optimization: A review of algorithms and applications. *arXiv preprint arXiv:2003.05689*.

Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. (2019). Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., and Loy, C. C. (2022). Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415.

Zhu, Q., Tian, X., Wong, C.-W., and Wu, M. (2019). Ecg reconstruction via ppg: A pilot study. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 1–4. IEEE.