Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Institut für Integrierte Systeme
Integrated Systems Laboratory

Department of
Information Technology and Electrical Engineering

Fall Semester 2025

Modality-Agnostic Framework for Biological Signal Analysis

Master Thesis

Federica Palmisano
fpalmisano@student.ethz.ch

01.05.2025

Advisors: Thorir Ingolfsson (ETHZ), thoriri@iis.ee.ethz.ch
Dr. Yawei Li (ETHZ), yawli@iis.ee.ethz.ch

Professor: Prof. Dr. Luca Benini (ETHZ), lbenini@iis.ee.ethz.ch

Abstract

Physiological signals, such as electroencephalography (EEG) and electrocardiography (ECG), are crucial for monitoring health, diagnosing conditions, and understanding complex physiological states. While deep learning has advanced their analysis, existing models often focus on single modalities or are tailored to specific tasks, hindering the development of unified, general-purpose frameworks capable of leveraging the inherent complementarity that multimodal biosignal data possesses. Furthermore, the scarcity of large-scale, multimodal labeled datasets remains a significant challenge in this domain.

This thesis introduces a novel multimodal foundation model designed for the comprehensive analysis of EEG and ECG signals. The proposed architecture integrates modality-specific encoders for both modalities, employing temporal and channel attention to learn intra-modal relationships, followed by a shared multimodal encoder that employs cross-attention to learn inter-modal interactions, and global self-attention for modality-agnostic multimodal fusion. To mitigate the reliance on multimodal labeled data, the model is pretrained using a multi-stage self-supervised learning strategy based on masked reconstruction, first training the unimodal encoders on large modality-specific datasets (TUEG for EEG, Icentia/PTB-XL for ECG), and subsequently training the shared multimodal encoder on a large multimodal dataset containing both EEG and ECG (SHHS). By effectively leveraging large amounts of unlabeled unimodal and multimodal data through this SSL approach, this work represents a step towards developing unified, general-purpose models for multimodal physiological data.

The efficacy and robustness of the pretrained foundation model were evaluated through finetuning on a diverse range of downstream tasks, including both multimodal (sleep stage classification, emotion recognition, transcranial electrical stimulation region classification) and unimodal EEG-only tasks (artifacts and slowing events classification, and abnormality detection). Experimental results demonstrate that the proposed multimodal architecture and multi-stage pretraining approach enable the model to learn powerful, transferable representations, achieving an increase of 3.2% in AUPRC for the TUAR dataset.

Acknowledgments

I would like to thank my supervisors, Thorir Mar Ingolfsson and Dr. Yawei Li, for their continuous support, for the helpful and insightful advice throughout this thesis.

I am grateful to Prof. Luca Benini, for having given me the opportunity and the resources to work on this thesis.

I would like to thank my collegues and friends, Nicolas, Berkay and Anna, for having made this experience both productive and fun.

Lastly, I would like to thank my dearest friend Chiara, my mother Simona, and my father, Antonio, for always having supported me, not just throughout this thesis, but during the whole duration of my studies.

Declaration of Originality

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every written paper or thesis authored during the course of studies. **In consultation with the supervisor**, one of the following two options must be selected:

□ I hereby declare that I authored the work in question independently, i.e. that no one helped me to author it. Suggestions from the supervisor regarding language and content are excepted. I used no generative artificial intelligence technologies[1].

☒ I hereby declare that I authored the work in question independently. In doing so I only used the authorised aids, which included suggestions from the supervisor regarding language and content and generative artificial intelligence technologies. The use of the latter and the respective source declarations proceeded in consultation with the supervisor.

Title of paper or thesis:

Modality-Agnostic Framework for Biological Signal Analysis

Authored by:
If the work was compiled in a group, the names of all authors are required.

Last name(s):	**First name(s):**
Palmisano	Federica

With my signature I confirm the following:

- I have adhered to the rules set out in the Citation Guidelines.
- I have documented all methods, data and processes truthfully and fully.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for originality.

Place, date	**Signature(s)**
Zürich, 01/05/2025	Federica Palmisano

If the work was compiled in a group, the names of all authors are required. Through their signatures they vouch jointly for the entire content of the written work.

[1] For further information please consult the ETH Zurich websites, e.g. https://ethz.ch/en/the-eth-zurich/education/ai-in-education.html and https://library.ethz.ch/en/researching-and-publishing/scientific-writing-at-eth-zurich.html (subject to change).

Contents

Chapter 1

Introduction

The increasing demand for objective and quantitative insights into human physiological and cognitive states underscores the critical role of biosignal analysis. Among the most valuable and widely studied biosignals are Electroencephalography (EEG), a non-invasive recording of the brain's electrical activity, and Electrocardiography (ECG), which captures the heart's electrical patterns. These modalities provide essential, and often complementary, information, providing invaluable insights for clinical diagnosis, patient monitoring, and neuroscience research. Recent years have seen a paradigm shift in automated biosignal analysis with the advent of deep learning. Notably, transformer architectures [1], leveraging powerful attention mechanisms, have proven exceptionally capable of modeling the complex temporal dependencies and multi-channel interactions inherent in both EEG and ECG data.

Despite the advances in deep learning and transformers for biosignal analysis, a significant gap remains. Current research often focuses on specialized models for individual modalities (EEG or ECG) or specific tasks and datasets. While the integration of multiple biosignals, particularly EEG and ECG, holds great potential for understanding complex physiological states, studies specifically focused on their joint analysis are relatively limited compared to unimodal research. Moreover, existing multimodal approaches within this domain are frequently tailored to a specific task, employing architectures and fusion strategies optimized solely for that purpose, which limits their broad applicability to other problems [2, 3, 4]. This fragmentation restricts knowledge transfer and hinders the creation of general-purpose AI tools for healthcare. Furthermore, the limited availability of large, high-quality, labeled multimodal biosignal datasets poses a major obstacle to training versatile supervised models. This context underscores the need for a unified, large-scale, general-purpose multimodal foundation model. Such a model should be capable of learning robust, shared representations from both unimodal and multimodal unlabeled data, enabling concurrent EEG and ECG analysis that generalizes effectively across diverse tasks.

This thesis proposes a novel framework to address this gap, introducing a multimodal foundation model specifically designed for the concurrent analysis of EEG and ECG signals. The proposed solution involves an architecture composed of modality-specific encoders for initial feature extraction, followed by a shared core fusion module. This core performs multimodal fusion through a combination of attention mechanisms: cross-attention, which explicitly models the relationships between EEG and ECG features, and self-attention, which processes the combined features in a modality-agnostic fashion, resulting in a shared representation. To mitigate the reliance on scarce labeled data and to yield more general representations, the model is trained through a multi-stage Self-Supervised Learning (SSL) strategy. This strategy leverages large-scale datasets, encompassing both unimodal EEG and ECG datasets in the initial pretraining stage, and multimodal EEG and ECG datasets for the pretraining of the shared core module. This approach aims to learn powerful representations of brain and heart activity that capture inter-modal relationships and can successfully generalize across different tasks.

The proposed multimodal foundation model is evaluated by assessing the generalizability and effectiveness of its learned representations across a variety of downstream tasks. This is achieved through finetuning the pretrained model on diverse labeled datasets. Evaluation includes multimodal tasks such as sleep stage classification and emotion recognition, and also unimodal EEG classification tasks, including artifact classification, slowing events classification and abnormality detection. Successful performance across these tasks demonstrates the model's ability to learn robust and clinically relevant representations from EEG and ECG signals.

This thesis is organized as follows: Chapter 2 provides the useful background on the fundamental concepts of physiological signals, specifically EEG and ECG, self-supervised learning, foundation models, and the datasets utilized. Chapter 3 reviews related works in foundation models for multimodal biosignals, further highlighting the research gap. Chapter 4 details the proposed method, including the full model architecture, the multi-stage self-supervised pretraining strategy, the data preprocessing pipeline, and the finetuning procedures. Chapter 5 presents and analyzes the experimental results. Last, Chapter 6 concludes the thesis by summarizing the key findings and discussing potential future developments.

Chapter 2

Background

This chapter will provide the relevant background knowledge necessary to understand the proposed methods and their applications. The chapter is divided into four areas, an introduction to biosignals, self-supervised learning for physiological signals, transformer models and attention, and lastly the datasets used in this thesis.

2.1. Signals Introduction

2.1.1. Biosignals

Biosignals are physiological signals that can be measured from living organisms, providing a window into the body's internal states and processes [5]. These signals, which can be electrical, chemical, mechanical, or optical in nature, carry valuable information about various biological functions, ranging from neural activity to cardiovascular dynamics. The analysis of biosignals has become increasingly crucial in diverse fields, including medicine, neuroscience, and biomedical engineering, offering objective and quantitative measures for diagnosis, monitoring, and understanding physiological mechanisms [6, 7]. Among the most widely studied and clinically relevant biosignals are electroencephalography (EEG) and electrocardiography (ECG), which will be discussed in detail below.

2.1.2. EEG

The human brain, the most complex organ in the body, is composed of billions of neurons that communicate with each other through electrical impulses. These impulses are generated by the flow of ions across the neuronal membrane, resulting in graded potentials. When a large number of neurons fire synchronously, the resulting electrical activity can be detected at the scalp level, leading to the formation of EEG signals. Electroencephalography (EEG) then stands as a critical non-invasive neurophysiological technique for capturing brain activity patterns. The signal is acquired by strategically placing electrodes on the scalp to detect and record the subtle electrical signals originating from the brain. These electrodes are typically arranged in a standardized manner, such as the 10-20 system [8], which ensures consistent electrode placement across different individuals and studies.

These multichannel EEG recordings are invaluable for monitoring sleep stages, analyzing cognitive and emotional states, but most importantly in clinical settings for the diagnosis and monitoring of various neurological conditions, including epilepsy, tumors, sleep disorders, and traumatic brain injuries [9, 10]. EEG signals are characterized by their complex and dynamic nature, reflecting the intricate interplay of neural activity across different brain regions. The signals are typically classified into different frequency bands, each associated with distinct brain activities and states of consciousness. The most commonly recognized frequency bands include delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–100 Hz) [11]. Each of these bands is linked to specific cognitive functions and mental states, providing valuable insights into the brain's functioning. For instance, delta waves are associated with deep sleep, while alpha waves are linked to relaxed wakefulness. The analysis of these frequency components allows clinicians and researchers to gain a deeper understanding of brain functioning and its relationship to various neurological conditions. EEG signals also present unique challenges due to their non-stationary behavior and non-linear dynamics, making it difficult to establish deterministic mathematical models that accurately capture the relationship between signal patterns and underlying brain conditions.

Consequently, machine learning and deep learning methodologies have become increasingly vital for modeling brain activity and uncovering relationships between EEG signal features and associated neurological states [12, 13]. Beyond diagnostics, EEG analysis extends to diverse applications, notably in Brain-Computer Interfaces (BCIs), where mental tasks like motor imagery are leveraged to control external devices [14, 15, 16, 17, 18], where advanced deep learning techniques are applied to decode emotional states from EEG patterns, sleep staging [17, 19], seizure detection [20, 21], artifact detection [22], and cognitive workload assessment.

2.1.3. ECG

Electrocardiography (ECG) stands as the primary non-invasive technique for assessing the heart's electrical function, a cornerstone in the diagnosis and management of cardiac conditions. By attaching electrodes to the limbs and chest, ECG captures the minute electrical currents generated by the heart muscle during each beat. These signals are displayed as characteristic waveforms, visually interpreted by cardiologists to identify patterns indicative of healthy or diseased cardiac function.

The ECG waveform is composed of recognizable segments – the P wave, QRS complex, and T wave – each reflecting a critical phase of the cardiac cycle, as outlined in established guidelines [23]. Specifically, the P wave represents atrial depolarization, the electrical signal that initiates contraction of the atria. Following the P wave, the atria contract, pumping blood into the ventricles. The QRS complex signifies ventricular depolarization, the much larger electrical event that triggers the contraction of the ventricles. Subsequently, the ventricles contract forcefully, ejecting blood into both the pulmonary and systemic circulations. Finally, the T wave corresponds to ventricular repolarization, the electrical recovery of the ventricles as they prepare for the next cycle. This repolarization process marks the electrical completion of a cardiac cycle, allowing the heart to relax and refill with blood before the next P wave initiates a new cycle.

These components provide a rich and fine-grained electrical representaion of the heart's activity, offering invaluable insights into overall cardiac health.

The clinical significance of ECG analysis is underscored by its crucial role in diagnosing cardiovascular diseases, the leading cause of mortality globally according to the World Health Organization. Within this broad category, cardiac arrhythmias stand out as particularly prevalent. Consequently, the accurate and reliable classification of these rhythm disturbances has become a central objective in both clinical practice and biomedical research, driven by the need for effective diagnosis and management strategies [24].

While ECG's non-invasive nature, ease of use, and cost-effectiveness [25] have solidified its position as a ubiquitous diagnostic tool, traditional visual analysis is resource-intensive, demanding expert review of lengthy recordings. This limitation has spurred significant research into machine learning applications for ECG processing [26, 27]. A fundamental step in automated ECG analysis is signal delineation, accurately segmenting and locating key features like the QRS complex in the signal [28]. Subsequently, machine learning models are employed to classify individual heartbeats as normal or abnormal, and to identify specific arrhythmia types such as Premature Atrial Contraction (PAC), Premature Ventricular Contractions (PVC) [29, 30, 31], or Atrial Fibrillation [32, 33]. Recent advancements highlight the growing role of transformer-based architectures in tackling diverse ECG analysis tasks, from event detection to beat and rhythm clssification and comprehensive arrhythmia classification .

2.1.4. EEG and ECG Comparison

Despite their different physiological origins, EEG and ECG share several properties as biosignals: both are time-series data, often multichannel, non-stationary, and susceptible to artifacts. However, they also differ significantly in terms of waveform characteristics, dominant frequency bands, and spatial dimensionality. These differences present challenges when attempting to develop unified models that can generalize across modalities.

- **Signal Origin and Generation**: EEG signals originate from the cerebral cortex, primarily reflecting the summed activity of post-synaptic potentials in large populations of neurons, while ECG signals are generated by the coordinated depolarization and repolarization of myocardial cells within the heart.

- **Frequency Content**: EEG signals exhibit a broader and generally higher frequency range, typically spanning from less than 0.5 Hz (delta waves) to over 100 Hz (gamma waves). ECG signals instead operate within a narrower and generally lower frequency range, predominantly below 100 Hz, with the most clinically significant information concentrated below 40-50 Hz. The primary focus in ECG is on the fundamental heart rate and the morphology of specific waves (P, QRS, T) rather than a broad spectrum of frequencies.

- **Amplitude**: EEG signals are characterized by relatively low amplitudes, typically measured in microvolts (µV). This low amplitude makes them highly susceptible to noise and requires significant amplification during recording. ECG signals, on the other hand, possess significantly higher amplitudes, typically in the millivolt (mV) range. This stronger signal makes ECG less prone to certain types of noise contamination compared to EEG.
- **Morphology and Complexity**: EEG waveforms are highly variable and complex, reflecting the intricate and dynamic nature of brain activity. While EEG rhythms exist, the overall morphology is less structured and repetitive compared to ECG. EEG patterns are often described in terms of bursts, spindles, sharp waves, and complex non-linear features. ECG waveforms, instead, exhibit a more structured and repetitive pattern with each heartbeat, characterized by well-defined waves that represent specific phases of the cardiac cycle.
- **Spatial Resolution**: scalp EEG has inherently poor spatial resolution due to volume conduction, where electrical signals spread widely through the conductive tissues of the scalp, skull, and brain. Electrodes pick up signals from a relatively broad area, making precise localization of neural sources challenging. ECG instead offers better spatial information relative to the heart, also due to the placement of the ECG electrodes, which captures different electrical vectors of the heart.
- **Temporal Resolution**: both EEG and ECG boast excellent temporal resolution, capable of capturing rapid changes in electrical activity on the order of milliseconds.
- **Susceptibility to Artifacts**: EEG recordings are notoriously susceptible to various artifacts, including muscle activity (EMG), eye blinks (EOG), movement artifacts, electrode artifacts, and power line interference. Robust artifact removal techniques are often essential for EEG data analysis. ECG is generally less artifact-prone than EEG, but it can still be affected by muscle artifacts, movement artifacts, baseline wander, and power line interference. However, ECG artifacts are often more readily identifiable and removed compared to the complex artifact landscape of EEG.

In summary, while both EEG and ECG are valuable tools for physiological monitoring, their distinct signal characteristics reflect their origins and the specific physiological processes they capture. EEG provides a complex and nuanced window into brain activity, while ECG offers a more structured and rhythmic representation of cardiac function. Understanding these differences is crucial for appropriately applying and interpreting these signals in both clinical and research settings, and for leveraging their complementarity in multimodal approaches, as explored in this thesis.

2.2. Self-supervised Learning for EEG and ECG

2.2.1. The Challenge of Labeled Data Scarcity

A significant impediment to advancing supervised machine learning for biosignal analysis is the scarcity of expertly labeled datasets. Unlike domains with readily available annotations, biosignal labeling, particularly for complex signals like EEG and ECG, demands specialized clinical knowledge and is inherently time-consuming. This annotation bottleneck restricts the scale and diversity of labeled biosignal data, limiting the applicability of purely supervised approaches and hindering the development of robust, generalizable models capable of addressing the full spectrum of clinical and research needs.

2.2.2. Leveraging unlabeled data

Self-Supervised Learning (SSL) offers a compelling paradigm shift to overcome the limitations imposed by labeled data scarcity in the biosignal domain. SSL techniques, such as masked signal modeling, capitalize on the inherent structure of unlabeled biosignals, enabling models to learn meaningful representations without relying on explicit annotations. Fortunately, vast quantities of unlabeled biosignal recordings, particularly EEG and ECG, are routinely collected in clinical and research settings. By harnessing this readily available unlabeled data, SSL unlocks the potential to pretrain robust models that can then be effectively adapted to downstream tasks, significantly reducing the dependence on expensive and limited labeled datasets and paving the way for more powerful biosignal AI.

2.2.3. Self-supervised Learning Techniques

Within the realm of Self-Supervised Learning (SSL), a variety of techniques have emerged, each leveraging different strategies to learn from unlabeled data. Two particularly prominent and impactful approaches are masked reconstruction and contrastive learning.

In Masked Reconstruction [34], a portion of the input signal is intentionally masked or removed, and the model is then trained to predict or reconstruct the missing segments based on the available context. This process compels the model to learn the underlying structure and dependencies within the data, effectively capturing temporal patterns and signal characteristics necessary for successful reconstruction. By training on this pretext task, the model develops robust representations that can be subsequently leveraged for various downstream tasks, demonstrating the power of learning through signal completion. Masked reconstruction has shown notable success across diverse domains, including natural language processing, computer vision, and increasingly, in the analysis of physiological signals.

Contrastive Learning [35] focuses on learning representations by discriminating between similar and dissimilar data points. The core idea is to train a model to pull representations of "positive" pairs (e.g., different views or augmentations of the same original data sample) closer together in an embedding space, while simultaneously pushing representations of "negative" pairs (e.g., unrelated data samples) further apart. This comparative learning process encourages the model to capture features that are invariant to certain transformations (for positive pairs) and discriminative against irrelevant variations (for negative pairs). Contrastive learning has proven highly effective in scenarios where defining a precise reconstruction target is challenging, and where the goal is to learn robust and discriminative features for tasks like classification or clustering. Like masked reconstruction, contrastive learning has found success in various domains, including computer vision, natural language processing, and is also gaining traction in biosignal analysis for learning representations that capture essential signal characteristics and variations.

2.3. Foundation Models

Foundation models mark a shift in machine learning towards broadly applicable, general-purpose models, trained on massive datasets instead of task-specific designs [36]. Characterized by their large scale and adaptability, these models are pretrained using self-supervised learning on extensive data [34, 37]. This pretraining yields rich, generalizable representations, facilitating efficient transfer learning and fine-tuning for diverse applications, even with limited labeled task-specific data [38].

For biosignal analysis, particularly EEG and ECG, foundation models offer significant potential. Traditional approaches often necessitate task-specific models and extensive labeled data. In contrast, a foundation model for EEG and ECG aims to learn universal representations from large unlabeled datasets. This pretraining enables a single model to be effectively adapted to numerous downstream tasks like sleep staging or anomaly detection, reducing task-specific training needs and promoting more robust and generalizable AI in healthcare and neuroscience.

This thesis contributes to this emerging field by introducing a modality-agnostic foundation model specifically designed to process and integrate both EEG and ECG signals, leveraging their inherent complementarity and addressing the growing need for unified and versatile models in biosignal analysis.

2.3.1. Multimodal Transformers

Traditional transformer architectures are substantially augmented by multimodal transformers, which are engineered to seamlessly process and integrate data from diverse sources, including textual, visual, and auditory information [39]. These sophisticated models capitalize on the inherent capabilities of transformers to discern temporal patterns and extract intricate details from varied data streams. This enables the creation of systems capable of generating outputs that synthesize information from multiple modalities, leading to richer, more contextually aware applications. For instance, in healthcare, multimodal transformers demonstrate exceptional utility by combining clinical narratives with medical imagery or physiological signals, thereby significantly improving the efficacy of diagnostic and classification procedures [40, 41].

Multimodal transformers can achieve the integration of information from multiple modalities through various fusion strategies, which can be broadly categorized into three types [42]:

- **Early Fusion**: the different modalities are combined at the input level before being fed into the transformer model. This approach allows for simultaneous processing of all modalities, enabling the model to learn joint representations from the outset. However, it may limit the model's ability to learn modality-specific representations before fusion, as the modalities are mixed early in the processing pipeline.
- **Late Fusion**: each modality is separately processed through individual transformer branches and then the outputs are combined at a later stage. This approach allows for specialized processing of each modality, potentially leading to better performance in tasks where modalities have distinct characteristics. However, it may miss out on some inter-modal relationships that could be captured in early fusion.
- **Intermediate Fusion**: strikes a balance between early and late fusion by integrating information from different modalities at various stages within the transformer architecture. This approach allows for flexible and adaptive learning of inter-modal relationships while still maintaining the benefits of separate processing. This can be achieved using techniques like cross-attention mechanisms, where representations from one modality attend to representations from another, or by concatenating or combining intermediate feature maps from different modality encoders.

A multimodal approach, leveraging diverse data sources, has proven effective in improving the performance of signal-based tasks. Research demonstrates that attention mechanisms benefit from the inclusion of additional inputs like text, enabling better integration and understanding of signal data. For instance, biosignals see a significant boost in transformer model performance when combined with other biosignals, clinical notes, or visual modalities.

For example, a combined analysis of ECG, PPG, and PCG signals allows for a more complete and reliable assessment of cardiac output [43, 44], while the integration of EEG, EMG, and EOG signals was shown to be beneficial in emotion recognition and sleep staging [45, 46, 3].

This ability to integrate and process diverse data streams highlights multimodal transformers as a pivotal technology for creating more precise, dependable, and context-aware AI solutions across domains such as healthcare to human-computer interaction.

2.3.2. Attention Mechanism

Attention mechanisms, particularly self-attention and cross-attention, are pivotal components in the proposed modality-agnostic foundation model, enabling it to effectively process and integrate EEG and ECG signals. Self-attention, at its core, allows the model to weigh the importance of different parts of the input sequence when processing each time point. For time-series data like EEG and ECG, this is crucial as it enables the capture of long-range temporal dependencies, where events distant in time can significantly influence the current state.

Self-Attention

Self-attention is a core mechanism within Transformer architectures that enables a model to weigh the importance of different parts of the input sequence when processing each element [1]. Unlike recurrent neural networks that process sequences sequentially, self-attention allows for parallel processing and directly captures relationships between all positions in the input, regardless of their temporal distance. This is particularly crucial for time-series data like EEG and ECG, where long-range dependencies are often significant.

At its heart, self-attention operates by calculating attention weights based on three learned matrices derived from the input sequence: Queries (Q), Keys (K), and Values (V). For each position in the input, a Query vector is generated. This Query is then compared to Key vectors from all positions in the sequence. The similarity between the Query and each Key is computed, typically using a scaled dot product, to determine the attention weights. These weights, often normalized using a Softmax function, indicate the relevance of each position in the input sequence to the current position being processed. Finally, these attention weights are used to perform a weighted sum of the Value vectors. This weighted sum produces the self-attention output for the current position, effectively incorporating information from relevant parts of the input sequence.

Mathematically, the self-attention mechanism can be expressed by the following formula:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_\text{k}}}\right)V \tag{2.1}$$

where Q is the Query matrix, K is the Key matrix, V is the Value matrix, and d_k is the dimensionality of the Key vectors. The scaling factor $\sqrt{d_\text{k}}$ is used to prevent excessively large dot product values, which can lead to saturation in the softmax function.

Cross-Attention

Cross-attention extends the attention mechanism to enable interaction and information fusion between two distinct input sequences [47]. Let us consider two input sequences, X and Y. While self-attention allows a sequence to attend to itself, cross-attention allows sequence X to attend to sequence Y (and potentially vice versa in a bidirectional manner). This mechanism is fundamental in scenarios where information from one modality or sequence needs to be integrated with another, such as in multimodal models or encoder-decoder architectures. In cross-attention, the Queries (Q) are derived from one input sequence, for instance X. The Keys (K) and Values (V), however, are derived from the other input sequence, Y. The attention weights are then computed by assessing the similarity between Queries from X and Keys from Y, effectively determining which parts of Y are most relevant to each element in X. These weights are subsequently used to aggregate the Value vectors from Y, resulting in an output representation for X that is informed by and aligned with relevant aspects of Y.

Mathematically, cross-attention shares a similar formulaic structure with self-attention, but with the key distinction in the origin of Q, K, and V:

$$\text{CrossAttention}(Q_\text{X}, K_\text{Y}, V_\text{Y}) = \text{softmax}\left(\frac{Q_\text{X} {K_\text{Y}}^T}{\sqrt{d_\text{k}}}\right) V_\text{Y} \tag{2.2}$$

Where:

- Q_X represents the Query matrix, derived from input sequence X.
- K_Y represents the Key matrix, derived from input sequence Y.
- V_Y represents the Value matrix, derived from input sequence Y.
- Q_X, K_Y, and V_Y are obtained through linear transformations of their respective input sequences.
- The rest of the formula (matrix multiplication, scaling, Softmax, weighted sum) is analogous to self-attention.

This cross-attention mechanism provides a powerful and flexible means for models to perform inter-sequence attention, enabling them to learn alignments, relationships, and fuse information between different modalities or input sources.

Temporal and Channel Attention

For complex sequential data, especially those acquired through multi-channel sensors, attention mechanisms can be further refined to capture nuanced relationships along different data dimensions. Beyond capturing temporal dynamics within the sequence itself (temporal attention), it is often beneficial to also consider the varying importance of different input features or channels (feature or channel attention). Assume the input tensor is denoted as X with shape $[B, N, D]$, where B is batch size, N is the total number of tokens (representing flattened time and channel dimensions, $N = T * C$ with T being time steps and C being channels), and D is the feature dimension.

- **Channel Attention**: To apply channel attention, we first reshape the input X to group by batch and time, treating channels as the sequence dimension for attention:

$$X_\text{channel} = \text{X: (B, N, D)} \rightarrow \text{(B*T, C, D)} \tag{2.3}$$

Here, X_{channel} now has shape $(B * T, C, D)$. Then self-attention is applied along the channel dimension (dimension C), to learn the importance of each channel, and then the output is reshaped back to the original shape (B, N, D).

- **Temporal Attention**: For temporal attention, we reshape the input X to group by batch and channels, treating time steps as the sequence dimension for attention:

  $$X_{\text{temporal}} = \text{X: (B, N, D)} \rightarrow \text{(B*C, T, D)} \tag{2.4}$$

  Now, X_{temporal} has shape $(B * C, T, D)$. Then self-attention is applied along the temporal dimension (dimension T) to capture temporal dynamics, and then the output is reshaped back to the original shape (B, N, D).

This approach allows the model to learn the importance of each channel and time step independently, enhancing its ability to capture complex relationships in the data. Combining channel and temporal attention allows for a more comprehensive and flexible processing of complex sequential data. This approach is particularly beneficial for signals like EEG and ECG where both spatial (channel-related) and temporal characteristics are crucial for understanding the underlying physiological processes.

2.4. Datasets

Several large-scale datasets support the development of EEG and ECG models. For EEG, the Temple University EEG Corpus (TUEG) provides extensive unlabeled recordings, while datasets like Sleep-EDF, SHHS, and TUAB offer labeled data for sleep staging, pathology detection, and more. For ECG, PTB-XL and Icentia 11k offer diverse examples of cardiac activity with clinical annotations. These datasets vary in sampling frequency, channel configuration, and labeling schemes, requiring careful preprocessing and standardization to enable unified modeling.

2.4.1. Pretraining Datasets

Temple University EEG Corpus (TUEG)

TUEG [48] is the world's largest publically available corpus of clinical EEG data. It contains data acquired from 10,874 unique subjects, for a total of 16,986 sessions and over 21,000 hours of recordings. The dataset presents a balanced gender distribution, and the age of the subjects ranges from less than 1 to over 90 years old. The EEG signals were acquired by placing electrodes on the scalp according to the 10-20 system, and the sampling frequency of the recordings is variable, with the majority of signals sampled at 250 Hz.

PTB-XL

The PTB-XL ECG dataset [49, 50, 51], is a large-scale dataset contaning 21,799 clinical 12-lead ECG records from 18869 patients, each of 10 second length, for a total of 60 hours of recordings.

The raw signals were acquired with 16 bit precision at a resolution of 1 microV/LSB and a sampling frequency of 500Hz. The signals were acquired following the standard set of 12-leads (I, II, III, AVL, AVR, AVF, V1, V2, V3, V4, V5, V6) with reference electrodes on the right arm, and they were collected with devices from Schiller AG between October 1989 and June 1996. The patients in the dataset are aged between 0 and 95 years, with a median age of 62 years, and the gender distribution is relatively balanced, with 52 percent of the patients being male and 48 percent being female.

This raw waveforms were annotated with ECG statements (conform to the SCP-ECG standard [52]) for each record, and they cover diagnostic, form, and rhythm statements. The distribution of the diagnosis in the PTB-XL is described in Table 2.1. It should be noted that the sum of the statements exceeds the number of records because of potentially multiple labels per record, and in fact, the value of the dataset results from the collection of several co-occurring pathologies.

Table 2.1.: Summary of PTB-XL Diagnosis

Diagnosis	**Count**
Normal ECG	9514
Myocardial Infarction	5469
ST/T Change	5235
Conduction Disturbance	4898
Hypertrophy	2649

Icentia-11k

Icentia-11k [53, 54, 51] is a large-scale ECG dataset designed for arrhythmia classification.

The Icentia-11k dataset consists of raw ECG signals recorded from 11,000 patients, and it contains 541,794 segments of 70-min each, for a total of 632,092 hours of recordings with 2 billion labeled beats.

The raw signals were acquired with the CardioSTAT device, a single-lead heart monitor from Icentia [55], in a modified lead 1 position. The signals were recorded with a 16-bit resolution and sampled at 250 Hz. They were acquired from patients who wore the device for up to 2 weeks, although most patients wore the device for one week. Each patient's record was then segmented into segments of 70 minutes each, and 50 segments were randomly selected per patient. The average age of the patients is 62.2±17.4 years. Furthermore, as it is noted by the authors, the dataset does not represent a true random sample of the global population, since the CardioSTAT device is mostly used for third line exam, therefore the majority of records in the dataset exhibit arrhythmias. The annotations were performed by 20 technologists, who classified both beat and rhythm types.

The beats are classified into four categories: Normal, Premature Atrial Contraction, Premature or ectopic supraventricular beat, Premature Ventricular Contraction, and Undefined. The rhythm types are classified into three categories: Normal Sinusal Rhythm (NSR), Atrial Fibrillation (AFib), and Atrial Flutter (AFlutter). The dataset contains 2 billion labeled beats, with each beat being annotated by twenty technologists.

The distribution of the beats in the Icentia-11k dataset is described in Table 2.2, and the distribution of the rhythm types is described in Table 2.3.

Table 2.2.: Summary of Icentia-11k Beat Labels

Label Description	**Count**
Normal	2,061,141,216
Premature or Ectopic Supraventricular beat, Premature Atrial Contraction	19,346,728
Premature Ventricular Contraction, Premature Ventricular Contraction	17,203,041
Unclassifiable beat	676,364,002

Table 2.3.: Summary of Icentia-11k Rhythm Labels

Rhythm Label	**Count**
Normal Sinusal Rhythm	16,083,158
Atrial Fibrillation	848,564
Atrial Flutter	313,251

SHHS

The Sleep Heart Health Study [56, 57] is a multi-center cohort study implemented by the National Heart Lung and Blood Institute to determine the cardiovascular and other consequences of sleep-disordered breathing. It tests whether sleep-related breathing is associated with an increased risk of coronary heart disease, stroke, all cause mortality, and hypertension.

This SHHS dataset, which contains polysomnograms, is divided into two parts, SHHS1 and SHHS2.

- SHHS1: contains data collected from 6441 subjects between 1995 and 1998
- SHHS2: contains data collected from 3295 of the 6441 subjects between 2001 and 2003

Of the 6441 subjects available, only the data for 5804 patients was made publically available. The SHHS dataset contains 73,728 hours of recordings.

The polysomnograms were obtained in an unattended setting (most often participants' home), and the recording montage consisted of: two-channels EEG (C3/A2 and C4/A1), ECG from a bipolar lead (both EEG and ECG sampled at 125 Hz), and other signals, such as EOG, EMG, and airflow.

The participants of the SHHS were recruited from nine existing epidemiological studies which collected data on cardiovascular risk factors. The original cohorts include: The Framingham Offspring Cohort, The Hagerstown and Minneapolis/St. Paul sites of the Atherosclerosis Risk in Communities (ARIC) study, The Hagerstown, Sacramento and Pittsburgh sites of the Cardiovascular Health Study (CHS), The Strong Heart Study sites in South Dakota, Oklahoma, and Arizona and Studies of respiratory disease in Tucson and of hypertension in New York. In many of these cohorts snorers were purpusefuly over-sampled in order to increase the dataset prevalence of sleep-disordered breathing.

The sample of participants selected from this cohort was recruited if the subjects were aged 40 years or older, had no history of treatment of sleep apnea, no tracheostomy, and no current home oxygen therapy.

The dataset is also annotated with a hypnogram, EEG arousal events, respiratory events, and oxymeter artifact annotations.

2.4.2. Finetuning Datasets

DREAMER

The DREAMER [58] dataset is a multimodal emotional dataset. It contains data from 23 subjects (age range: 22-33, average age: 26.6), of which 14 are male and 9 are female. In the experiment, the subjects watched 18 audio-visual film clips (each clip duration: 65 to 393 s, average duration: 199 s), designed to solicit specific emotions, such as amusement, surprise or anger. Each session lasted approximately one hour, therefore the dataset amounts to a total of 23 hours of recordings. The dataset contains EEG and ECG recordings, and the participants' self-assessment of their affective state after each stimuli.

The recordings contain 14 channels of EEG signals, sampled at 128 Hz and using electrodes placed according to the 10-20 international system, and 2 channels of electrocardiogram (ECG) signals, sampled at 256 Hz.

For the self assessment, the participants had to assign a score from 1 to 5 to Valence, Arousal and Dominance.

For each subject, a 61-s pre-trial baseline was recorded before each clip, and they were asked to rate the valence, arousal, and dominance levels for each film clip from 1 to 5.

A neutral clip was also shown to the participants before each film clip, in order to help the subjects return to a neutral emotional state.

CAP Sleep

The CAP Sleep Database [59, 51] contains 108 polysomnograms from 108 subjects, for a total of 1009 hours, and it was collected at the Sleep Disorders Center of the Ospedale Maggiore of Parma, Italy.

The recordings include EEG signals (variable channel number and montages across subjects), ECG (bipolar lead), EOG, EMG, airflow, respiratory effort and SaO2. The CAP Sleep is also annotated with sleep stages (W, S1, S2, S3, S4, REM) and with CAP events annotations according to the Rechtschaffen and Kales rules [60]. The Cyclic Alternating Pattern (CAP) is a periodic EEG activity occurring during NREM sleep, whose abnormal amounts are associated with a variety of sleep-related disorders.

Of the 108 subjects, 16 are healthy subjects and 92 are patients with sleep disorders, as it is shown in Table 2.4 below. In the participants cohort, there are 66 males and 42 females, with age ranging from 14 to 82 years old, with an average age of 39.8 years.

Table 2.4.: Summary of CAP Sleep Participants

Pathology	**Count**
Healthy subjects	16
Bruxism	2
Insomnia	9
Narcolepsy	5
NFLE	40
PLM	10
RBD	22
SDB	4

Dataset for TES

The Dataset for TES [61] provides a rich, multimodal collection of data from human participants undergoing High-Definition transcranial electrical stimulation (HD-tES). It combines high-density electroencephalography (EEG), physiological signals (ECG, electrooculography/EOG), and continuous behavioral measures (vigilance/alertness metrics from a Compensatory Tracking Task). The data was collected across multiple sessions where participants received nine different HD-tES stimulation montages, varying by cortical target region (frontal, motor, parietal) and stimulation waveform frequency

(DC, 5 Hz, 30 Hz). The dataset is designed to support research into the acute effects of tES on brain-body-behavior interactions.

The dataset was acquired from 20 neurologically typical individuals, aged 19-43 years (mean age: 29.10 ± 6.75 years), among which 13 are males and 7 are females. Of these 20 subjects, one was excluded from the study, due to inability to follow task instructions, resulting in a total of 19 subjects.

The dataset is divided into two parts, Experiment 1 and Experiment 2, and it includes 62 sessions, which lasted 70 minutes for Experiment 1 and 70.5 minutes for Experiment 2. This amounts to over 72 hours of continuous EEG, ECG, EOG, and behavioral recordings across all sessions.

The EEG data was acquired using a 32-channel EEG cap, with electrodes placed according to the 10/10 international placement system. These electrodes were interleaved with 9 HD-tES stimulation electrodes.

The ECG data was acquired using a Lead-I bipolar configuration, with two bipolar electrodes placed on the chest (approximately 5 cm below left and right clavicle bone) and a ground electrode on the left hip. Both EEG and ECG signals were sampled at 2 kHz.

The labels present in the dataset indicate the specific HD-tES condition applied during each stimulation trial, which depend on the cortical region and the frequency used, for a total of nine labels: F0, F5, F30, M0, M5, M30, P0, P5, P30.

TUAR (Artifact Corpus)

TUAR is a specialized dataset for artifact detection. It is annotated with common artifacts for EEG signals, such as eye blinks and muscle artifacts, for a total of six classes. This dataset contains data from 213 subjects, for a total of 213 hours of recordings.

TUAB (Abnormal EEG Corpus)

TUAB is a specialized dataset for abnormality detection in EEG signals. It is a subset of the Temple University EEG Corpus (TUEG) and is specifically annotated for abnormality detection, with recordings labeled as either 'normal' or 'abnormal'. The dataset contains a total of 2,329 subjects, for a total of 2,329 hours of recordings, and features a balanced class distribution.

TUSL (Slowing Corpus)

TUSL is a specialized dataset for EEG slowing event detection. It is annotated for EEG slowing events, with recordings labeled as 'slowing events', 'seizures', 'complex background' or 'normal', for a total for four classes. This dataset contains data from 38 subjects, for a total of 27.54 hours of recordings, and presents a highly imbalanced class distribution.

Table 2.5.: Summary of Pretraining Datasets

Dataset	Type	Task	Num. Subj.	Recording hrs.	Num. Electr.	Sampl. Frequency
TUEG	EEG	EEG monitoring	10,874	21,000+	22	250 Hz
PTB-XL	ECG	ECG Diagnosis	21,799	60	12	500 Hz
Icentia-11k	ECG	Arrhythmia Classification	11,000	632,092	1	250 Hz
SHHS	EEG	Sleep study dataset	5804	73,728	3	125

Table 2.6.: Summary of Finetuning Datasets

Dataset	Type	Task	Num. Subj.	Recording hrs.	Num. Electr.	Sampl. Frequency
DREAMER	EEG, ECG	Emotion recognition	23	23	16	128 (EEG), 256 (ECG)
CAP Sleep	EEG, ECG	Sleep staging	108	1009	>2	[128, 512]
TES	EEG, ECG	Transcranial electrical stimulation	19	72	32 (EEG), 1 (ECG)	2000
TUAR	EEG	Artifact detection	213	83.74	22	256
TUAB	EEG	Abnormality detection	2,329	1139.31	22	256
TUSL	EEG	Slowing event detection	38	27.54	22	256

Chapter 3

Related Work

In this section, previous works relevant to this study will be discussed, with a particular focus on multimodal and modality-agnostic models. Finally, we highlight the gaps in current research that motivate our proposed approach.

3.1. Multimodal and Modality-Agnostic Models

BIOT [62] is a flexible biosignal encoder architecture designed specifically to overcome the challenges posed by format mismatches present in different datasets, such as different number of channels, recording lengths, sampling rates, and missing values. BIOT aims to enable "cross-data learning" and joint pretraining and finetuning on diverse datasets, with different formats. To prove this, the authors pretrained and evaluated the model on tasks which are EEG-only, and ECG-only, for instance seizure detection, abnormality detection, event classification (EEG), and arrhythmia phenotype prediction (ECG). The experiments carried out in this paper also show that the model pretrained on multiple unimodal datasets perform better on that modality's downstram tasks compared to models that were pretrained from scratch on a single dataset. Even though this paper addresses the crucial problem of format mismatch, it does not feature multimodal pretraining, and does not include training the model on a dataset that combines both EEG and ECG inputs concurrently.

Abbaspourazad et al. (2024) [63] present a foundation model trained on large-scale wearable biosignal datasets containing ECG and PPG. In this work, the authors wanted to address the challenge of limited labeled medical data and therefore leveraged the use of SSL to train their models on a large-scale unlabeled dataset. Within the realm of SSL, they used CL with participant-level positive pair selection. In this study, the authors separately trained different models in ECG and PPG, without explicitly including any fusion of the two modalities. They also extensively evaluated the models' performances by using linear probing on the pretrained embeddings. Although this study leverages a massive unlabeled dataset, makes use of SSL frameworks, and trains on multimodal data, it does not explicitly address the problem of training on two modalities at the same time, nor does it include any fusion of the two modalities.

Recent efforts, such as the large-scale CLIMB benchmark [64] , have highlighted the critical need for unified datasets to advance multimodal clinical AI and have empirically validated the benefits of large-scale pretraining. Their evaluations, which covered modalities including EEG and ECG, confirmed that pretraining on diverse clinical data significantly improves model performance. Crucially, this pre-training fosters generalizability; unlike hyperspecialized, dataset-specific models that can achieve maximum performance on narrow tasks but often fail to adapt to new scenarios, as described in [64], models pretrained on broad benchmarks like CLIMB demonstrate better adaptability across diverse clinical modalities, tasks, and contexts. Furthermore, CLIMB demonstrated that strong unimodal representations learned during pre-training transfer successfully enhance multimodal task performance, when making use of appropriate fusion strategies. This aligns directly with and supports our approach of first developing powerful, dedicated EEG and ECG encoders via pre-training on large respective datasets, which are then integrated using our proposed cross-modal attention fusion mechanism.

In this work [65] , the authors propose a multimodal foundation model for physiological signals, specifically EEG, EMG, EOG and ECG from the PhysioNet 2018 sleep dataset. The key hypothesis tested in this study is that the quality of model representations can be improved by encouraging masked reconstruction of the modalities. In fact, a multimodal masked autoencoding (MultiMAE) objective is employed, where concatenated tokens from all modalities (this is an example of early fusion) are masked and they are all separately reconstructed by the modality-specific decoders. The authors also added modality dropout, showing that the performance of the introduced model surpasses that of the traditional MutiMAE in downstream tasks, such as sleep staging and arousal detection. Thus, Fang et al. effectively demonstrate that implicitly encouraging cross-modal learning through both the multimodal masked autoencoding and modality dropout improves representation quality. However, their work leaves a gap regarding the development of models specifically designed and trained for functional robustness when a modality is entirely unavailable during inference. Furthermore, the emphasis on complex decoders (here containing cross-attention) contrasts with the goal of building a powerful, generalizable encoder, subsequently paired with a simpler model head, which can benefit from pretraining on diverse, large-scale datasets beyond the single one utilized in their study. Lastly, their implicit cross-modal learning mechanism leaves unexplored explicit ones within the core fusion process.

Mostafaei et al. [66] propose a novel deep learning model, specifically designed for sleep stage classification. The model utilizes multiple physiological channels (EEG, EOG, ECG, EMG, respiratory) from the SHHS dataset, also employed in this work. In this work, the authors explored the use of raw physiological signals together with handcrafted features. In the model, features are initially extracted with different CNN-Attention blocks based on the characteristics of the different modalities, then two transformer encoders are employed to process features from the raw signals and the hand-crafted features, which are the two modalities, as is written in the paper. Then, cross-attention is used between the two modalities, meaning between the raw signal's features and the handcrafted features, for all the modalities together. Lastly, a transformer decoder is used to predict the sleep stages.

Lee et al. propose SynthSleepNet [67] a ViT model specifically designed for sleep analysis, using multimodal polysomnography data, EEG, EOG, EMG, and ECG. It proposes a hybrid SSL strategy that combines masked reconstruction and contrastive learning. This model contains modality-specific backbones, a multimodal ViT encoder, and separate modality-specific ViT decoders. The multimodal ViT is responsible for fusing the embeddings and uses masked reconstruction. The model also uses a contrastive loss to align the unimodal representations learned from modality-specific encoders to the multimodal average representation learned by the multimodal ViT encoder.

Gong et al. [3] propose a novel deep learning network for emotion recognition using multimodal physiological signals, specifically focusing on EEG and Peripheral Physiological Signals (PPS) like EOG, EMG, and ECG. Their key idea is to simultaneously consider inter-modality and intra-modality relationships. The way this is achieved is by using three parallel blocks, two intra-modal encoders with self-attention specific for EEG and PPS, and a pairwise cross-modal transformer to capture inter-modal correlations and complementarity. The model is trained with a joint loss that considers outputs from the three blocks. Modality-agnosticism or robustness to missing modalities is not explicitly addressed in this study.

Sun et al. [68] propose START, a transformer model designed specifically for sleep stage classification, which uses EEG and EOG signals. START uses two modality-specific transformer encoders with CLS tokens, and after this a cross-modal fusion block, which takes as input the CLS tokens produced by the two modality-specific encoders and trains in addition another CLS tokens, specific to learn cross-modal interactions. Then the final prediction is made using this CLS token, after passing it through a feedforward network.

sDREAMER [69] is a transformer-based model specific for automatic sleep staging that uses EEG and EMG signals from a mouse dataset. sDREAMER uses the MoME (Mixture-of-Modality-Experts) architecture with parallel paths and partially shared weights. Specifically, there are three paths; two are modality-specific for EEG and EMG, and the third is modality-agnostic, made with self-attention, for both signals together. To enhance cross-modal interaction, the authors use self-distillation training, in which the output of the shared path distills knowledge into the single-modality pathaways. This design is also inherently robust to missing modalities.

Twins-PainViT [2], is a modality-agnostic model specifically developed for automatic pain assessment that employs facial videos and fNIRS signals. The modality-agnosticism in this model is achieved by transforming all the inputs (which have different formats) into a standardized format, a 2D image, which can then be processed by a common ViT, independently of the original data type. Specifically, the authors used two different encoders, the first to extract embeddings from the data used, and the second to output pain assessment predictions.

Chang et al. propose VigilanceNet [70], a multimodal model for vigilance estimation that uses EEG and EOG. Their main contributions are the outer product embedding (to capture multiplicative relationships within EOG features), and the decoupling of intra- and inter-modality learning. Specifically, they enforce the intra-modality learning using separate feature extractors for the two signal modalities, and by making them independently predict the vigilance level. Inter-modality learning is allowed by a cross-modal fusion transformer that captures complementary information between EEG and EOG.

Götz et al. [4] propose a Modality-Agnostic Transformer-Based Self-Supervised Learning Specific for Emotion Recognition Using ECG and EDA from the WESAD dataset. The model uses modality-specific feature extractors (CNNs), followed by concatenation of the features, which are then processed together in a transformer encoder via self-attention. For the downstream task, the frozen CLS tokens are fed into a transformer decoder followed by a classifier that outputs the prediction. The model does not imply robustness to missing modalities during inference. While presenting a modality-agnostic approach, this model is not designed to be explicitly robust to missing modalities, employs a transformer decoder, and is emotion recognition specific.

3.2. Research Gap and Contribution

Despite significant progress in applying deep learning to multimodal physiological signals, several critical gaps remain, particularly in leveraging the synergistic potential of EEG and ECG.

First, there is a notable scarcity of studies explicitly focusing on fusing EEG and ECG signals, despite their combined importance in various clinical scenarios like sleep analysis and cardiovascular monitoring.

Second, many existing multimodal models are task-specific and often evaluated on limited or single datasets, hindering the assessment of their generalizability and robustness across diverse clinical conditions and data sources. This reliance on narrow evaluations contrasts sharply with the need for broadly applicable foundation models.

Third, architecturally, while various fusion methods exist, there is limited exploration of models that employ a dedicated intermediate fusion block utilizing explicit cross-attention mechanisms to directly model the interactions between separately processed EEG and ECG feature sequences, combined with joint self-attention.

Fourth, the crucial challenge of functional modality agnosticism, ensuring robust model performance even when one modality (EEG or ECG) is entirely absent during inference, has not been adequately addressed through targeted training strategies.

Finally, the potential benefits of specific multi-stage pretraining paradigms (large-scale unimodal pretraining followed by dedicated multimodal fusion training) and the integration of domain-specific handcrafted features within advanced transformer architectures for these biosignals remain largely untapped.

To address these gaps, this study introduces a novel foundation model framework specifically designed for robust and synergistic EEG-ECG analysis. Our main contributions are as follows.

- **EEG-ECG Focused Foundation Model**: We propose a novel foundation model architecture specifically targeting the integration of EEG and ECG signals, trained and evaluated on multiple large-scale and diverse datasets to promote generalizability.

- **Novel Intermediate Fusion Architecture**: Our model features a dedicated intermediate fusion block that processes outputs from separately pretrained EEG and ECG encoders. This block uses explicit cross-attention to model inter-modal dependencies between EEG and ECG and joint channel/temporal self-attention on the combined representation.

- **Targeted Training for Modality Agnosticism**: We explicitly train the fusion component using modality masking, directly optimizing the model for robust performance even when provided with only unimodal input during fine-tuning or inference.

- **Multi-Stage Training and Feature Integration**: We employ a distinct training approach that involves large-scale unimodal pretraining followed by multimodal fusion training and demonstrate the effective integration of both raw physiological signals and expert-derived handcrafted features within our transformer-based framework.

Through this approach, we aim to develop a generalizable, robust, and modality-agnostic foundation model for EEG-ECG analysis, improving performance and adaptability across various downstream tasks compared to task-specific or less robust multimodal models.

Chapter 4

Implementation

4.1. Model Architecture Structure

This section details the architecture of the proposed modality-agnostic foundation model, designed for joint processing of EEG and ECG signals. The architecture is based on the Transformer framework and employs specialized attention mechanisms for both unimodal and cross-modal feature extraction, enabling the learning of shared representations. The main components of the architecture are the modality-specific EEG and ECG encoders, the fusion encoder, and the decoders (for pretraining and finetuning), as it will be discussed in this chapter.

As it can be seen from Figure Fig. 4.1, the model processes segments of both EEG and ECG signals, which are first processed separately in their modality-specific encoders, and then combined together in the cross-modal fusion encoder. After the encoder, the embeddings are processed by the patch reconstruction decoder during pretraining, or by the classification head, during finetuning.

4.1.1. Input Processing and Tokenization

The tokenization separately processes the raw time series, EEG and ECG, transforming them into token sequences that will later be input in the transformer encoders.

Before the tokenization, EEG and ECG have shapes:

$$[B, C_{\text{exg}}, T] \tag{4.1}$$

where B is batch size, C_{exg} is the number of channels for EEG or ECG, and T is the sample time points, which in this model corresponds to a 5-second window of signals with sampling frequency of 256 Hz.

EEG and ECG are patched and tokenized separately, as a part of their modality-specific processing. The tokenization is performed using a 2D convolution with a kernel size of (1, patch_size) and a stride of (1, patch_size), which is applied independently for each

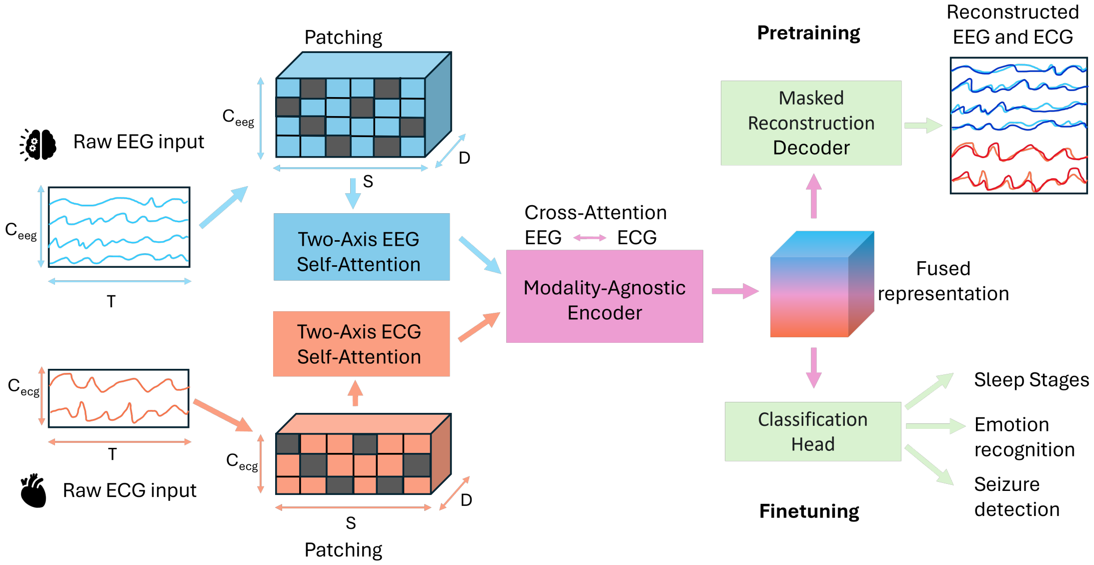

Figure 4.1.: Architecture of the Multimodal Foundation Model for EEG and ECG. EEG and ECG are first processed separately by the two modality-specific encoders (in blue for EEG and orange for ECG), and then combined together in the multimodal fusion block (in pink). The produced shared embedding is then input to one of the model heads (in green), specifically to the masked reconstruction decoder during pretraining, or to the classification head for the finetuning downstream tasks.

channel. This convolution effectively partitions the input signals into non-overlapping patches.

After the patching the shapes become:

$$[B, C_{\text{exg}}, T] \rightarrow [B, C_{\text{exg}}, S, P] \tag{4.2}$$

where S is the number of patches, and P is the patch size.

These patches are then mapped to an embedding dimension D:

$$[B, C_{\text{exg}}, S, P] \rightarrow [B, C_{\text{exg}}, S, D] \tag{4.3}$$

The patches are finally rearranged multiplying the number of channels C with the number of patches S:

$$[B, C_{\text{exg}}, S, D] \rightarrow [B, N_{\text{exg}}, D] \tag{4.4}$$

where $N_{\text{exg}} = C_{\text{exg}} * S$ is the number of tokens for the EEG modality.

4.1.2. Feature extraction

Before entering the core Transformer blocks, the tokenized sequence is augmented with additional handcrafted features extracted from the original raw signals. In the feature calculations, all channels are treated independently.

Both time and frequency features are calculated, specifically the features employed are: mean, standard deviation, zero-crossing rate, kurtosis, skewness, energy, entropy and a spectral feature to express frequency content.

This frequency feature is calculated using the Real Fast Fourier Transform (RFFT) applied along the time dimension. From this frequency representation, both the magnitude and the phase are extracted and concatenated, forming a frequency feature vector for all the patches. This frequency feature then represents the energy distributions across all frequencies in all the patches and channels.

It should be noted that the features used in this thesis are modality-agnostic, which means that they are not specific to either EEG or ECG.

Similar features were used in previous works, such as [66], where mean, standard deviation, kurtosis, skewness, zero-crossing rate and entropy are used. This is because these features are indicative of time-domain characteristics of the waveforms used, as pointed out in [71]. Differently from the cited works, in this thesis a frequency representation (agnostic to the modality type) is calculated, instead of specific features tailored specifically to EEG (e.g. band-specific frequencies) or to ECG (e.g. R peak features, R-R Interval).

The calculated features are then embedded using dedicated Multi Layer Perceptrons, which consists of three linear layers interleaved with a non-linear activation function (GELU), projecting the features into the tokens dimension, such that the features can be summed to the patched input $[B, N, D]$. This passage allows for the integration of the complementary information offered by the raw signal and its statistical and spectral features.

4.1.3. Positional and Channel Encoding

After the initial patching and feature extraction, learned positional and channel encodings are added to the tokens. These embeddings serve the purpose of capturing the original channel identity and the temporal position of each token. The channel and positional encodings are concatenated together, and are then added to the feature-augmented token embeddings, providing essential spatio-temporal context.

For the finetuning datasets, in which there is variability in the channel montages used (as will be discussed in Section 4.3.2), a channel index dictionary is employed to pair each channel with the correct channel encoding, to ensure compatibility between the channels of the finetuning data set and the encoding learned by the model on the pretraining dataset. This is done specifically for EEG, since ECG in the finetuning datasets contains lead I signals, and is then mapped correctly to the first position in the channel encoding.

4.1.4. Modality-Specific Encoders (EEG and ECG)

The modality-specific encoders, inspired by CEReBrO [72] take as input the tokenized sequence augmented by the handcrafted features, after this has been summed to the channel and positional encodings. There are two modality-specific encoders, one for EEG signals and one for ECG signals. Each of them internaly performs Multi-Head Two-Axis Self-Attention. The encoder initially reshapes the input as follows:

$$[B, N_{\text{modality}}, D] \rightarrow [B, C_{\text{modality}}, T, D] \tag{4.5}$$

and then applies *channel attention* along the channel dimension, C_{modality}, for each time step independently:

$$[B, C_{\text{modality}}, T, D] \rightarrow [B * T, \mathbf{C_{modality}}, D] \tag{4.6}$$

and *temporal attention* along the temporal dimension, T, for each channel independently:

$$[B, C_{\text{modality}}, T, D] \rightarrow [B * C_{\text{modality}}, \mathbf{T}, D] \tag{4.7}$$

where the sequence dimension in the attention mechanism is highlighted in bold.

The outputs of these two attention computations are then summed and linearly projected. This two-axis approach enables the unimodal encoders to capture complex spatio-temporal relationships within each modality.

The Two-Axis attention block (it should be noted that the same operations are also applied Hybrid Attention described below) receives an input the normalized x (Layer Normalization), then layer scaling is applied to its output, a residual connection is implemented to stabilize gradient flow through the network, and finally drop path is applied as a regularization technique.

The x processed as such is further normalized, then it is input into an standard MLP with GELU as the activation function, and finally layer scaling and drop path are applied again.

4.1.5. Shared Modality-Agnostic Encoder

The shared modality-agnostic encoder receives as input the concatenated outputs of the two modality-specific encoders, x_{eeg} and x_{ecg}, which result in the combined sequence x, with shape $[B, N, D]$, where:

$$N = N_{\text{eeg}} + N_{\text{ecg}} \tag{4.8}$$

where N_{eeg} and N_{ecg} are the number of tokens for EEG and ECG respectively.

This shared block implements a hybrid attention mechanism, implemented by sequentially applying Multi-Head Two-Axis Self-Attention on the full concatenated embedding x, and Multi-Head Cross-Attention between the EEG and the ECG components of the Two-Axis Self-Attention output.

Specifically, the Two-Axis Self-Attention performs self-attention along both the temporal and channel dimensions of the *concatenated token space*, as discussed in 2.3.2, refining the spatio-temporal relationships across the entire observation space. The output of this attention function has shape $[B, N, D]$.

This is followed by a Cross-Attention mechanism that explicitly facilitates interaction and information flow between the EEG and ECG modalities, enabling the model to learn how features in one signal relate to features in the other. This module takes the output of the Two-Axis Self-Attention step as its input. The Cross-Attention mechanism first splits the x into x_{eeg} and x_{ecg}, and then applies attention from EEG to ECG and from ECG to EEG.

In the first case, the attention is computed between Q_{eeg}, K_{ecg} and V_{ecg}, producing output of shape $[B, N_{\text{eeg}}, D]$:

$$x_{\text{eeg} \to \text{ecg}} = \text{Attention}(Q_{\text{eeg}}, K_{\text{ecg}}, V_{\text{ecg}}) \tag{4.9}$$

In the second case, the attention is computed between Q_{ecg}, K_{eeg} and V_{eeg}, producing output of shape $[B, N_{\text{ecg}}, D]$:

$$x_{\text{ecg} \to \text{eeg}} = \text{Attention}(Q_{\text{ecg}}, K_{\text{eeg}}, V_{\text{eeg}}) \tag{4.10}$$

These two outputs will then be concatenated in the tokens dimension, restoring the original shape $[B, N, D]$:

$$x = \text{Concat}(x_{\text{eeg} \to \text{ecg}}, x_{\text{ecg} \to \text{eeg}}) \tag{4.11}$$

where x has shape $[B, N, D]$.

To summarize, this sequential hybrid approach employed within every shared encoder block first refines the spatio-temporal context across the combined multimodal input space and then explicitly fuses information between the two modalities.

4.1.6. Modality-Agnosticism Analysis

While the model employs two modality-specific encoders, each designed to capture features unique to EEG or ECG, the subsequent shared encoder operates within a unified framework achieving modality-agnosticism.

The application of Two-Axis Self-Attention to the *full multimodal input* within the shared blocks is inherently *modality-agnostic*, as it processes all tokens in the combined space uniformly. This distinguishes this step from the Two-Axis attention computation in the modality-specific encoders, which is performed separately on EEG and ECG, in order to learn specific *intra-modal relationships*.

The following Cross-Attention mechanism necessarily relies on knowing the modalities' channels to be able to compute attention between the two sets of tokens, serving the purpose to enforce explicit *cross-modal relationships* learning.

Stacking these two attention mechanisms sequentially allows the model to first build robust, contextualized representations within the *combined multimodal space*, and then to leverage these refined representations for explicit *cross-modal fusion*.

The shared encoder outputs a single, integrated tensor which combines information from both EEG and ECG, allowing downstream task-specific modules to connect directly and function without needing to know the original modalities or channel configurations. This design provides flexibility, enabling the shared encoder to learn adaptable representations of inter-modal dynamics. These representations can then be used by task-specific modules trained on a variety of tasks, all leveraging this single, unified embedding.

4.1.7. Decoder

Two different decoders are used for this model, a patch reconstruction head for the pretraining, and a classification head for the finetuning.

The patch reconstruction head is used in the context of masked reconstruction, which is the SSL strategy employed in this project. This decoder takes as input the shared latent representation produced by the cross-modal block, and maps it back to the input signals. This module head consists of a simple linear layer that projects the embedding dimension D to the original patch size P.

The classification head is used for the downstream tasks, it also receives as input the latent representations output from the shared multimodal encoder, and it consists of a Multi Layer Perceptron. in the model head the input, with shape $[B, N, D]$, is first averaged on the tokens dimension, N, and then processed by the MLP, which produces classification logits, projecting the embedding dimension D to the number of classes.

4.2. Training Methodology

4.2.1. Overall Strategy

The overview of the model training will be detailed below, and it can be summarized as follows: first, a multi-stage pretraining is performed, utilizing the two modality-specific encoders in a first stage, and the shared modality encoder, in a second stage; then the pretrained model is finetuned on six different datasets and downstream tasks.

4.2.2. Multi-stage Pretraining

The proposed foundation model employs a multi-stage pretraining strategy designed to effectively learn from both unimodal and multimodal biosignal data. This motivations behind this approach are threefold.

First, EEG and ECG signals possess distinct characteristics, highlighting the need to learn modality-specific features.

Second, EEG and ECG often offer complementary information, when acquired simultaneously, highlighting the need to learn cross-modal interactions between the two signals.

Third, despire the abundance of unimodal datasets containing either EEG or ECG, even large-scale ones, such as TUEG or Icentia-11k, there is a scarcity of datasets, (especially large-scale) containing both modalities together.

These reasons motivate us to leverage the abundant unimodal data to learn intra-modal relationships in EEG and ECG in a first stage, and to leverage the available simultaneous data in a second stage, while refining inter-modal interactions in a modality-agnostic fashion.

Implementation

In the first stage, separate encoders are pretrained for EEG and ECG independently, utilizing modality-specific datasets, TUEG for EEG and an Icentia-11k subset combined with PTB-XL for ECG. This stage is crucial to allow each encoder to specialize in extracting features relevant to its unique signal properties, and to leverage the extensive data available for a single modality . EEG and ECG exhibit significant differences in frequency content, morphology, and physiological origin, necessitating distinct feature extraction strategies at the initial encoding layers.

In the second stage of the multi-stage pretraining strategy, the already pretrained EEG and ECG encoders (obtained from Stage 1) are loaded. These unimodal encoders' weights are initialized from their respective pretrained checkpoints. A newly initialized Shared Modality-Agnostic Encoder is then added on top of these two unimodal encoders, and initialized from scratch. Pretraining is continued on the entire model using the multimodal SHHS dataset, which contains both EEG and ECG. The objective remains masked reconstruction, now applied to the combined EEG and ECG inputs. This is the stage in which the Hybrid-Attention mechanism is trained, enabling the model to learn effective cross-modal information fusion and joint representations based on the previously learned unimodal features.

By learning both modality-specific and cross-modal representations, this strategy is expected to enhance the model's generalization capabilities and improve its performance on downstream tasks that leverage both EEG and ECG signals.

4.2.3. Pretraining Self-Supervised Learning

Masked reconstruction and MSE loss

The core self-supervised pretraining task for this model is Masked Reconstruction, which has been successfully employed in previous works, such as LaBram [73], CEReBrO [72] and [74]. Each input segment is first patched, and then a random subset of patches is masked, using a learnable mask embedding. The model then processes the input masked as such, and subsequently the decoder is tasked with the reconstruction of the signal, both in visible and in masked regions.

The model is trained using the L_2 or Mean Squared Error (MSE) loss function calculated in both the visible and masked region, as expressed below:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{masked}} + \alpha \cdot \mathcal{L}_{\text{visible}} \tag{4.12}$$

where $\mathcal{L}_{\text{visible}}$ is the $\mathcal{L}_2$ calculated for the visible patches, $\mathcal{L}_{\text{masked}}$ is the $\mathcal{L}_2$ calculated for the masked patches, and α is a hyperparameter set to 0.1.

The L_2 loss measures the squared differences between the predicted values for a set of patches and their true values, and is calculated as:

$$\mathcal{L}_2(X_{\text{pred}}, X_{\text{target}}) = (X_{\text{pred}} - X_{\text{target}})^2 \tag{4.13}$$

where X_{pred} is the prediction made by the model head, and $\mathrm{X}_{\text{target}}$ is the true, masked input.

Modality masking

During the pretraining process, specifically in Stage 2, a modality masking strategy is applied probabilistically to the input batches.

With a certain probability, the data for either EEG or ECG modality within a given batch (probability to mask EEG or ECG is set to 0.5), is randomly set to zero.

This forces the model to process batches where only a single modality is present, simulating scenarios where one signal might be unavailable or noisy. To ensure the model is not penalized for failing to reconstruct the modality that it did not receive, the masked reconstruction loss is calculated only based on the unmasked modalities in that specific batch.

This technique contributes to the modality-agnosticism of the learned representation by explicitly training the model to extract meaningful information and perform reconstruction when relying on input from just one modality, enhancing its ability to function even with partial multimodal inputs.

4.2.4. Task-Specific Supervised Finetuning

Finetuning Tasks

Following the multi-stage self-supervised pretraining, the learned representations within the model's encoders are utilized to solve specific downstream classification tasks.

The finetuning process involves leveraging the pretrained model as a feature extractor and finetuning its parameters on smaller, labeled datasets relevant to each target task. This is achieved by discarding the model head responsible for masked reconstruction, and attaching the classification head, described in 4.1.7.

The tasks used for evaluation in this project are:

- **Sleep stage classification**: Performed on the CAP Sleep dataset, multimodal EEG and ECG dataset.
- **Emotion recognition**: Evaluated on the DREAMER dataset multimodal EEG and ECG dataset.
- **TES region classification (frontal, motor, parietal)**: multimodal EEG and ECG dataset.
- **Artifact classification**: TUAR dataset, EEG only.

- **Slowing events detection**: TUSL dataset, EEG only.
- **Abnormality detection**: TUAB dataset, EEG only.

Layer-wise Learning Rate Decay

In this work, Layer-wise Learning Rate Decay is applied in both the second stage of self-supervised pretraining (Shared Encoder pretraining) and during task-specific supervised finetuning.

In the second pretraining stage, this strategy is used to continue pretraining the already pretrained EEG and ECG encoders with smaller learning rates while training the newly added Shared Encoder layers with larger rates.

During finetuning, Layer-wise Learning Rate Decay it is applied across the entire pretrained model, allowing the pretrained layers to retain their learned features while the new classification head adapts quickly to the downstream task. T

his approach helps stabilize training and improve performance when leveraging powerful pretrained representations.

Focal Loss

The loss function used for the finetuning tasks is the focal loss [75].

This choice was motivated by the severe class imbalances encountered in some of the finetuning datasets, in which the cross-entropy loss function would have rendered the model biased towards majority classes, resulting in poor performances on the minority classes.

Focal Loss modifies the standard Cross-Entropy loss to down-weight the contribution of easy-to-classify, abundant examples, focusing the training on hard, misclassified examples, which are often from the minority classes.

The focal loss formula is expressed as follows:

$$\mathcal{L}_{\text{focal}} = -\alpha \cdot (1 - p_{\text{t}})^{\gamma} \cdot \log(p_{\text{t}}) \tag{4.14}$$

where p_{t} is the model's estimated probability for each class, α is a balancing factor for the class, and γ is a focusing parameter that adjusts the rate at which easy examples are down-weighted.

The label distribution for the finetuning datasets is presented in . The FocalLoss implementation utilized takes two main parameters: α and γ.

The α parameter is a weighting factor assigned per class, which is set inversely proportional to the class-frequency to balance the importance of different classes.

The γ parameter is a modulating factor that adjusts the rate at which easy examples are down-weighted. By reducing the loss contribution from well-classified examples, Focal Loss effectively addresses the class imbalance issue by directing the model's learning capacity towards the hardest-to-classify samples from minority classes.

Evaluation Metrics

The performance of the finetuned model on each downstream task is evaluated using a set of metrics, which include:

- **Accuracy**: calculates the proportion of correctly classified examples.
- **Precision**: measures the proportion of true positive predictions among all positive predictions (relevant to false positives).
- **Recall**: measures the proportion of true positive predictions among all actual positive examples (relevant to false negatives).
- **AUROC (Area Under the Receiver Operating Characteristic Curve)**: quantifies the overall performance of a binary classifier by measuring its ability to discriminate between positive and negative instances across a range of classification thresholds.
- **AUPRC (Area Under the Precision-Recall Curve)**: informative for binary classification tasks with severe class imbalance, as it focuses on the performance on the positive class.
- **Cohen's Kappa**: measures the agreement between the model's predictions and the true labels, adjusted for the agreement expected by random chance.

The metrics were macro-averaged across all classes.

4.2.5. Data Handling for Varying Shapes

In this project a custom data loader was used in order to take care of the different number of channels present in the pretraining (TUEG, Icentia and PTB-XL) and in the finetuning datasets (CAP Sleep). Specifically, regarding the TUEG, this contains samples with 22-channels or 20-channels, for the ECG encoder the Icentia-11k contains 1-channel, and the PTB-XL contains 12-channels, and for the CAP Sleep this contains twelve different EEG channels montages.

This custom dataloader creates batches with samples having the same number of channels, which is required in order to stack the samples in a batch for further processing in the transformers. The dataloader reads files created with a fixed number of channels, such that samples (also potentially belonging to different files) with equal channel number will be stacked together. This dataloader has been extended to work also for the ECG pretraining block, in which the Icentia and the PTB-XL datasets were also processed in different formats, respectively LMDB and HDF5. For the CAP Sleep, during data preprocessing a combination index was assigned to each subject based on the channels number and types it contained after data cleaning, preprocessing and channel standardization. This combination index was then stored in the HDF5 files as attribute for each group. During data loading, the combination index is read, and batches are created with samples having the same specific montage. In the data loading process, one dataloader is created for all the samples sharing the same combination index. These dataloaders are then combined and iterated through sequentially.

4.3. Data Preprocessing

This section describes the diverse biosignal datasets used for both pretraining and finetuning, outlining the standardized preprocessing pipeline applied to ensure compatibility with the model architecture, and the necessary label handling for the finetuning datasets.

4.3.1. Overview of Data Sources

In this project diverse datasets were employed, containing both EEG and ECG, or just one modality, as described in Tables 2.5 and 2.6.

Specifically, these datasets are:

- **Unimodal Pretraining datasets:** TUEG, Icentia-11k, PTB-XL.
- **Multimodal Pretraining datasets:** SHHS.
- **Multimodal Finetuning datasets:** CAP Sleep, DREAMER, Dataset for tES.
- **Multimodal Finetuning datasets:** TUAB, TUAR, TUSL.

These datasets exhibit significant heterogeneity, in terms of sampling frequencies, ranging from 125 Hz to 2000 Hz, number of electrodes and montage used (especially for EEG signals), ranging from 1-channel to over 30, level of noise present in the data, labeling scheme, and overall dataset structure.

Furthermore, the total recording durations span a vast range, from tens of hours (e.g., TUAR, TUSL, DREAMER) to hundreds of thousands of hours (e.g., Icentia-11k).

This substantial initial diversity across datasets, subjects, signal types, recording durations, channel configurations, and sampling rates necessitates a consistent preprocessing and standardization pipeline to create a unified data format suitable for the training and finetuning of the proposed foundation model.

4.3.2. Core Standardization Pipeline

To prepare the heterogeneous biosignal data for the Transformer-based model, a consistent core preprocessing pipeline was applied across all datasets. This pipeline transforms raw signals into standardized, fixed-length segments with uniform characteristics:

- **Discard invalid values**: data was checked for invalid values (NaN/Inf). Segments with excessive invalid data were discarded, while limited NaN or Inf values were replaced with zero.

- **Bandpass Filtering**: signals were filtered to remove unwanted frequency components.

  Specifically, EEG signals was bandpass filtered with frequencies of 0.1 Hz and 75.0 Hz.

  ECG signals were bandpass filtered with frequencies of 0.5 Hz and 120 Hz.

  The sampling was modified for the CAP Sleep dataset, in accordance with the guidelines of the American Academy of Sleep medicine [76], which suggests to filter EEG with frequencies of 0.3 Hz and 35 Hz, and ECG with frequencies of 0.3 Hz and 70 Hz. Thus, a narrower frequency range was considered in order to obtain optimal conditions for sleep scoring.

- **Notch Filter**: in the various datasets, a notch filter of either 50 Hz or 60 Hz was applied in order to eliminate contaminations from powerline interferences.

- **Resampling to common frequency**: a crucial aspect of this preprocessing pipeline, is to take care of the different sampling frequencies different datasets use for the same modality, but most importantly to take care of the inherent differences between EEG and ECG signals, that, if as evident as a a different sampling frequency of different order of magnitudes for the amplitudes, could bias the model into trivially distinguishing between the two modalities.

  Hence, it was necessary to resample EEG and ECG in all the datasets to a common frequency, that was chosen to be 256 Hz. This effectively standardized the temporal resolution across all datasets and modalities.

- **Channel Standardization and Ordering**: the channel montage was standardized across all datasets, especially regarding the EEG signals, which were acquired following the 10-10, the 10-20 international system or unique layouts as seen in the CAP Sleep dataset. This standardization involved: converting original unipolar measurements to a bipolar format, and ordering the resulting channels according to a consistent layout, in this case the TCP montage was chosen. To accurately manage the variability in original channels and montages, a mapping dictionary was generated for each subject during preprocessing. This dictionary explicitly defined a correspondence between the position of a channel in the data and in the standardized TCP montage. This mapping was crucial to maintain consistency in the channel encodings, guaranteeing that the model associated its learned representations with the same anatomical locations *across all data*, which is vital for effective learning with channel-specific features.

  The case of the CAP Sleep dataset proved particularly challenging, due to strong internal montage variations, which amounted to twelve different channel montages across all subjects. It should also be noted that for the CAP dataset, out of the 95 subjects considered (the ramaining 13 were excluded due to corrupt files), 4 were excluded because they contained no standard EEG channels.

- **Windowing**: continuous recordings were segmented into fixed-length, non-overlapping 5-seconds windows, equivalent to 1280 timepoints at 256 Hz. These windows serve as the fundamental input samples for the model.
- **Label Extraction**: for the finetuning datasets preprocessed in this project (DREAMER, CAP Sleep and Dataset for tES), the labels were extracted from the annotation data. Details for each dataset are discussed below in 4.3.3.
- **Normalization**: min-Max normalization in the range on [-1, 1] was applied to all data, independently for each 5-second sample and for each channel.

  This was another crucial step in the preprocessing pipeline, together with the resampling, to avoid biasing the model, which would otherwise receive amplitudes in the range of mV for ECG and of μV for EEG signals. This also avoided potential training instability due the very different input amplitudes.

Filtering thresholds motivation

A standard practice while filtering EEG or ECG signals, involves applying bandpass filters to remove unwanted frequency components using linear digital filters with zero phase distortion, implemented here by applying a Butterworth filter forwards and backwards.

Specific frequency cutoffs are chosen to align with the physiologically relevant signal content and common noise sources for each modality, while also being constrained by the original data acquisition parameters (sampling frequency).

For both modalities, a high-pass filter is applied to remove slow baseline drifts, respiration artifacts, and DC offsets. A cutoff of 0.1 Hz is used for EEG, while a slightly higher cutoff of 0.5 Hz is applied for ECG. For ECG, common recommendations for the low-frequency cutoff in diagnostic recordings, which aim to minimize ST segment distortion and to preserve the overall waveform, range from 0.05 Hz for routine filters up to 0.67 Hz for linear digital filters with zero phase distortion, as discussed in the Recommendations for the Standardization and Interpretation of the Electrocardiogram from the American Heart Association Electrocardiography and Arrhythmias Committee [77]. Our 0.5 Hz high-pass filter for ECG falls within this clinically recommended range for zero-phase filters.

The upper cutoff frequency, however, differs more significantly between the modalities and is fundamentally limited by the original data sampling rates. For EEG, a low-pass filter at 75 Hz is applied, in accordance with the guidelineds of the American Clinical Neurophysiology Society [78], which recommend to use high-frequencies not below 30 Hz, and low frequencies not above 1 Hz. This cutoff is chosen to encompass the full range of established brain rhythms, including important gamma band activity (up to around 70-100 Hz), while effectively attenuating higher-frequency noise such as muscle artifacts (EMG). In contrast, for ECG, while many works in the literature use a high cutoff of 100 Hz [27], clinical guidelines emphasize that preserving even higher frequencies is fundamental for accurate measurement of rapid deflections and detailed QRS complex morphology, with recommendations often suggesting an upper-frequency cutoff of at least 150 Hz, or even 250 Hz for infants, for precise amplitude measurements [77]. However, achieving such bandwidth is fundamentally limited by the data's original sampling rate due to the Nyquist theorem. Our ECG datasets were originally sampled at varying rates, including 500 Hz, 250 Hz, and 125 Hz, corresponding to Nyquist frequencies of 250 Hz, 125 Hz, and 62.5 Hz. Consequently, a consistent upper cutoff of 150 Hz is not feasible across all data due to these Nyquist limits. We therefore selected a uniform low-pass filter at 120 Hz for all ECG data.

4.3.3. Handling Finetuning Datasets

For the downstream finetuning tasks, specific procedures were implemented to process the diverse original label formats into a consistent representation compatible with the model's classification head.

For CAP Sleep, the original labels were provided as hypnograms for 30-second epochs, and they were propagated to align with our standardized 5-second analysis windows. This involved assigning the 30-second epoch label to all 5-second windows falling entirely within that epoch. Missing labels in the original hypnogram were also propagated to the corresponding 5-second windows. Furthermore, original sleep stage labels (W, S1, S2, S3, S4, REM) were converted to the standard (W, N1, N2, N3, REM) nomenclature according to established guidelines [79].

For DREAMER, which provided self-assessment scores on a scale from 1 to 5 for valence, arousal, and dominance, these scores were further discretized into binary categories: scores of 1 to 3 were mapped to a 'Low' category, and scores of 4 and 5 were mapped to a 'High' category, as previously done in the literature. Additionally, a distinct label ('0') was created and assigned to data segments corresponding to the baseline, neutral video signals. This was done to explicitly teach the model to differentiate between physiological states elicited by stimulus videos and those during a neutral baseline period.

For the TES region classification task, labels were derived from the trigger annotations indicating occurring transcranial electrical stimulation (tES). The original triggers contain information on the specific stimulation montage, which includes information about the stimulated brain region (Frontal, Motor, or Parietal) and the stimulation current applied (DC, 5Hz, 30Hz). During preprocessing, data segments were labeled based on the brain region targeted by the tES, discarding the specific stimulation frequency information. The trigger sequences containing triggers for the start and end of the stimulation was checked for invalid trigger sequences, which were excluded from the finetuning data. Non-stimulation periods were implicitly treated as a separate category or assigned a baseline label, in this case 0.

4.3.4. Label Distribution

Below the label distribution for the finetuning datasets is shown. It can be seen that most datasets are imbalanced. This is in fact what motivated the use of the focal loss, as described in 4.2.4.

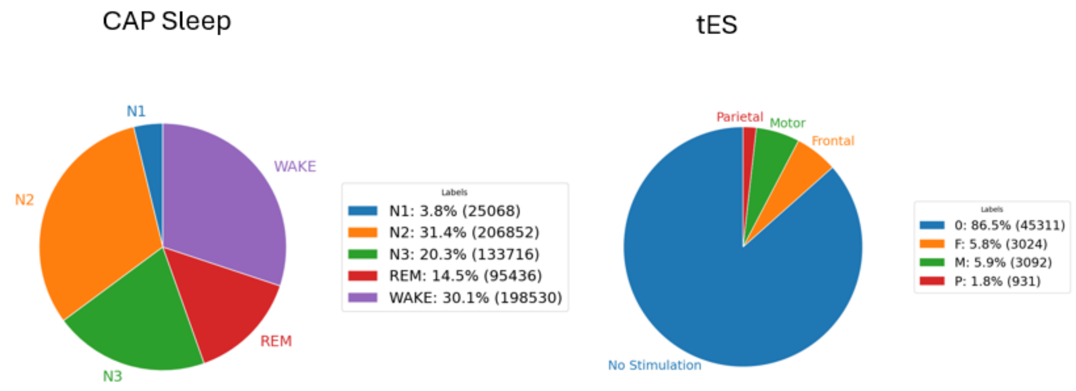

Figure 4.2.: Class distributions for the CAP Sleep and the tES datasets

As it can be seen in Figure 4.2, the datasets are imbalanced. For instance, the tES datasets, whose majority of the data corresponds to the absence of stimulation, has one class which only constitutes 1.8% of the data, while the majority makes up for 86.5% of the data. This is the original labels distribution, but in this project the TES classification task is actually a binary one, with the two classes being the frontal and the motor regions, which make up for a balanced dataset instead.

The CAP Sleep is also imbalanced, with the N1 class which is 3.8% of the data, and the N2 which corresponds to 31.4%.

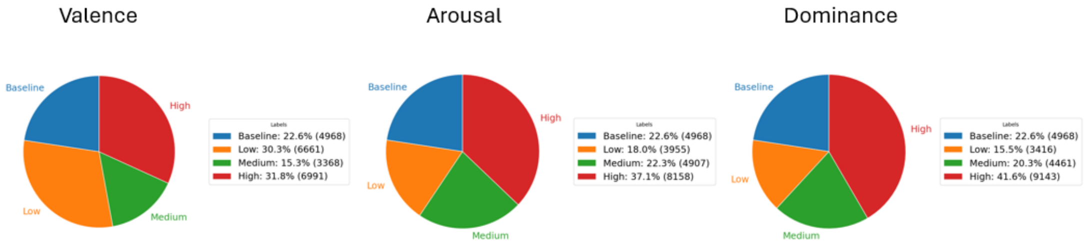

Figure 4.3.: Class distributions for the DREAMER dataset

The DREAMER dataset, while not perfectly balanced, is more evenly distributed, with the low and high classes being 40.4% and 41.2% respectively, while the neutral class is 18.4%.

As it can be seen in Figure 4.4, while the TUAB dataset is relatively balanced, the TUAR and most importantly the TUSL datasets are imbalanced. In the TUSL, one class constitues 97% of the total data, while the remaining three contribute to the total data for 1% each.

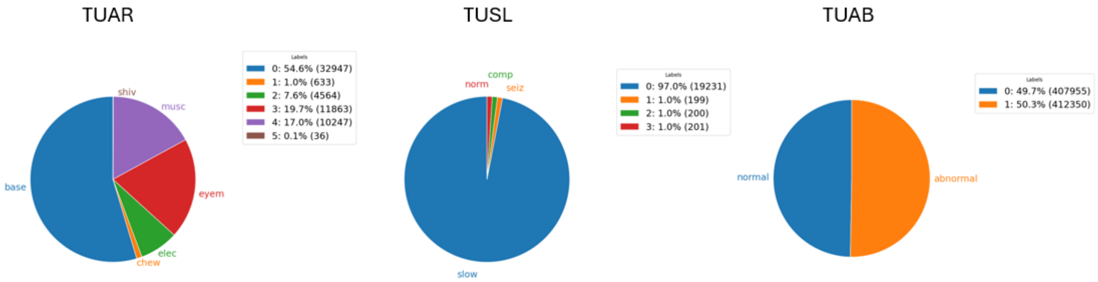

Figure 4.4.: Class distributions for the TUAR, TUSL and TUAB datasets

Chapter 5

Results

In this chapter, the results obtained from the evaluation of the proposed method are presented. The model is analyzed in the downstream task performance, and in the evaluation of the latent representations obtained from the pretrained model on the downstream task datasets.

5.1. Training Curves

The training curves, both for train and validation, are shown in Figure Fig. 5.1 and Figure Fig. 5.2.

These curves were obtained from the pretraining of the shared fusion model, starting from pretrained checkpoints obtained from the initial stage pretraining of the modality-specific encoders.

5.2. Masked Reconstructions

In this section the masked reconstructions produced by the pretraining of the fusion encoder are shown, with both EEG and ECG, in Figure and in Figure . The reconstructions, obtained during the validation stage, are presented here both for EEG and ECG, and they were obtained pretrained the shared encoder for 40 epochs.

These reconstructions are produced from the SHHS dataset, which contains two EEG channels, and one ECG channel.

5.3. Downstream datasets embedding visualizations

In this section, the t-SNE visualizations for some of the finetuning datasets are shown. These embedding visualizations were obtained using the pretrained encoder (in this case the three encoders, modality-specific and modality-agnostic) and using it to produce latent representations of the finetuning datasets. Figure Fig. 5.6 shows the embedding

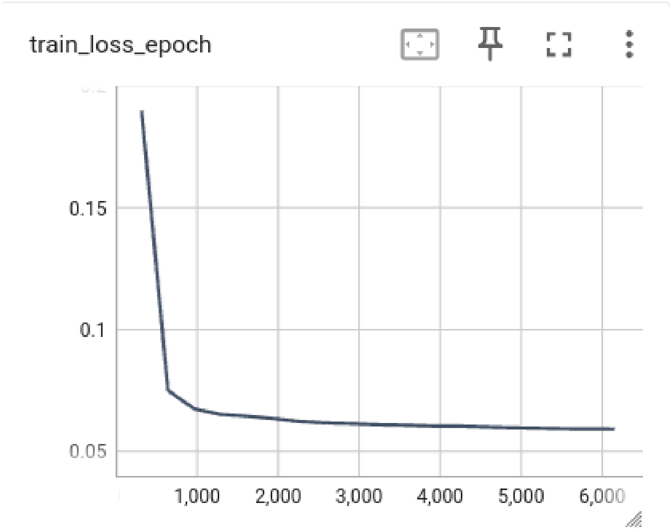

Figure 5.1.: Training curve for the shared fusion model.

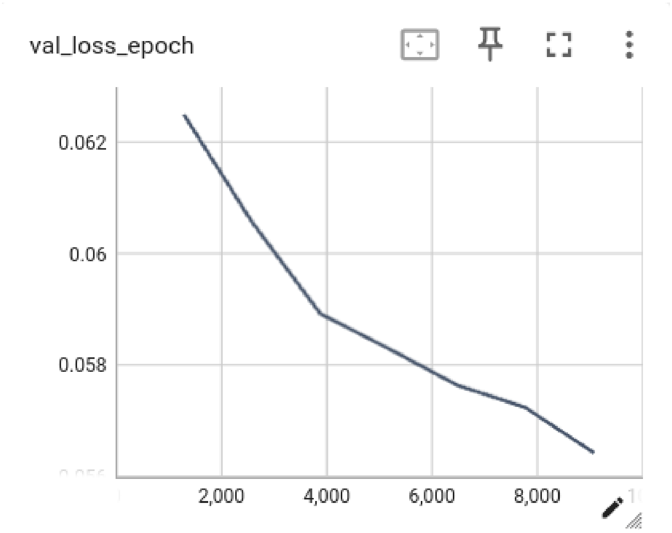

Figure 5.2.: Validation curve for the shared fusion model.

visualizations for the TES dataset. In Figure Fig. 5.5 the embeddings for the arousal prediction of the DREAMER dataset are shown. Lastly, in Figure Fig. 5.7 the embeddings produced using the TUAR dataset, EEG-only, are shown.

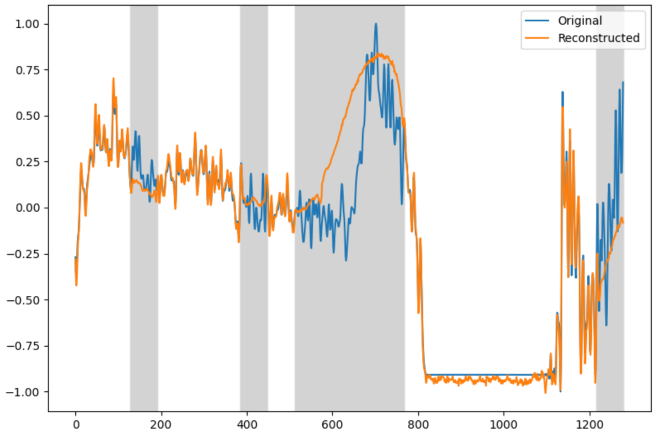

Figure 5.3.: EEG validation reconstruction produced by the shared encoder

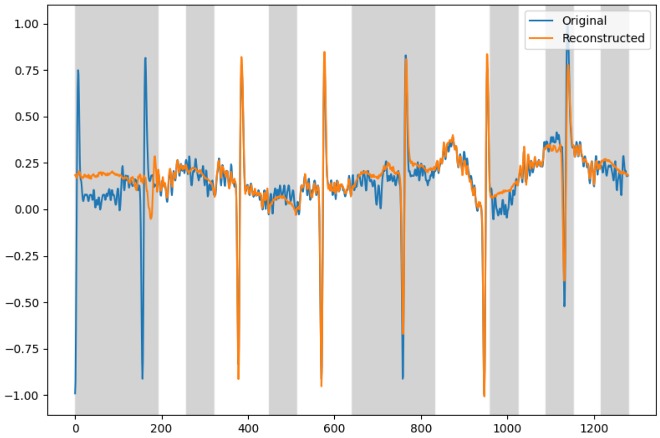

Figure 5.4.: ECG validation reconstruction produced by the shared encoder

5.4. Finetuning Results

In this section, the finetuning results obtained from the downstream tasks conducted on TES, DREAMER, CAP, TUAB, TUAR and TUSL are shown.

For the EEG only datasets, the finetuning results on TUAB, TUAR and TUSL are reported respectively in Table 5.1, Table 5.2 and Table 5.3.

The results obtained finetuning on the TES dataset, are instead shown in Table 5.4. The results obtained on the CAP Sleep dataset are shown in Table 5.5 .

The results obtained after finetuning on the DREAMER dataset are shown in Table 5.7.

In the finetuning results presented in Tables 5.5, 5.7, 5.4 we can observe the model performance on EEG and ECG datasets. Comparing the metrics for the full model and the baseline, we can observe that for the DREAMER dataset, the baseline model, which was only pretrained on the SHHS dataset, and not on the TUEG, the Icentia-11k, and

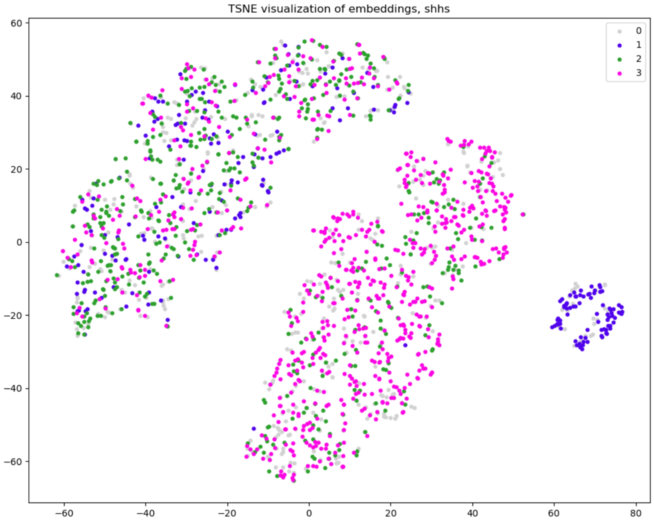

Figure 5.5.: Embedding visualization of Arousal prediction of the DREAMER dataset

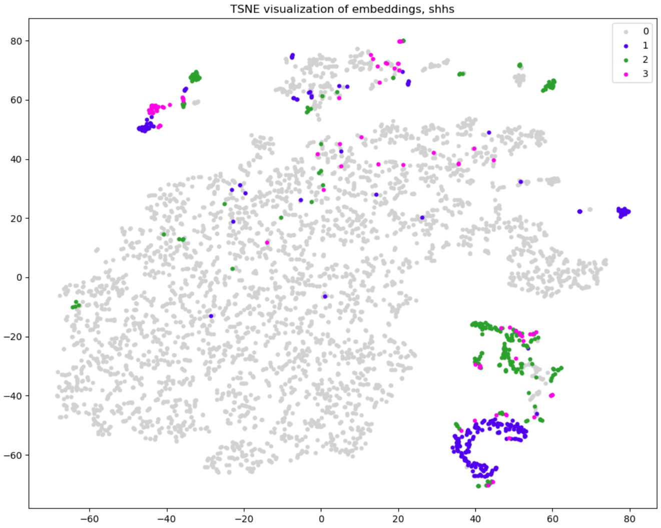

Figure 5.6.: Embedding visualization of the TES dataset

the PTB-XL, achieves slightly higher metric values than the full model. The general hypothesis is that the baseline model, being pretrained on a less datasets, and also without multi-stage pretraining, should perform worse than the full model. Although the slightly higher metric values seem to contradict this hypothesis, it should also be noted than these values are chance values for a binary classification, therefore the baseline does not actually perform better than the full model, given the meaning of these metrics for a binary classification task.

Looking at the finetuning results on the CAP dataset, a five classes classification task

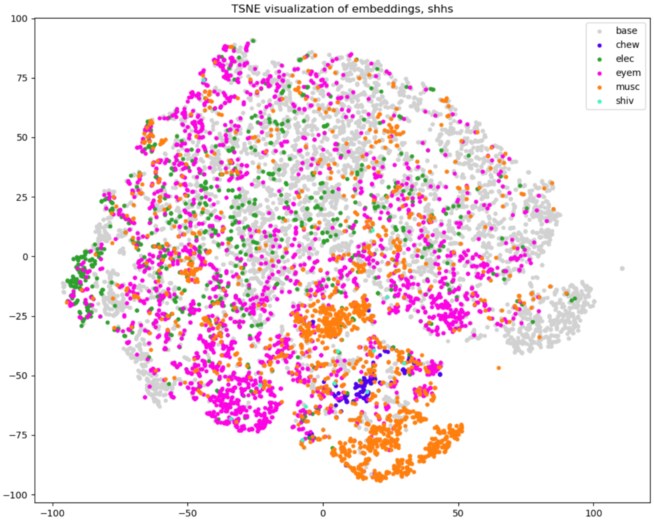

Figure 5.7.: Embedding visualization of the TUAR dataset

Table 5.1.: Results on TUAB

Models	AUROC	AUPRC	Accuracy
Full, no feat	0.8949 ± 0.0018	0.9036 ± 0.0019	**0.8229 ± 0.0037**
Full, feat	0.8936 ± 0.0020	0.9031 ± 0.0016	0.8227 ± 0.0045
Full, mod mask	**0.8968 ± 0.0029**	**0.9059 ± 0.0033**	0.8205 ± 0.0070
Baseline, no feat	0.8929 ± 0.0017	0.9025 ± 0.0012	0.8200 ± 0.0026
Baseline, feat	0.8883 ± 0.0004	0.8984 ± 0.0009	0.8165 ± 0.0012
Baseline, mod mask	0.8829 ± 0.0038	0.8914 ± 0.0042	0.8096 ± 0.0065

for sleep staging, we can see that the performance achieved my baseline and full model is comparable, the full model does not perform consistently better than the baseline. It should still be noted that for the CAP dataset, as for the DREAMER, the performance of the multimodal EEG-ECG model is suboptimal, and also inferior to the state-of-the-art results for the F1 score metric, the only one reported in the two previous works.

While performance on DREAMER and CAP is suboptimal, performance onthe TES dataset is more promising. The TES dataset confirms the hypothesis according which the full model should perform better than the baseline model, and the hypothesis according to which adding handcrafted features should improve model performance. We can in fact see that the model with the features, both full and baseline, performs consistently better than the corresponding model version without features. The modality masking experiment, intended primarily for EEG only datasets, was also tested here. We can see that for the full model, it achieved comparable results to the features-enriched model, while for the baseline model, it achieved much worse performance. This can confirm

Table 5.2.: Results on TUAR

Models	AUROC	AUPRC	Accuracy
Full, no feat	**0.9161** ± **0.0068**	**0.5909** ± **0.0238**	**0.4743** ± **0.0087**
Full, feat	0.8966 ± 0.0076	0.5048 ± 0.0108	0.4423 ± 0.0068
Full, mod mask	0.8925 ± 0.0159	0.5072 ± 0.0252	0.4511 ± 0.0203
Baseline, no feat	0.9051 ± 0.0175	0.5398 ± 0.0471	0.4565 ± 0.0263
Baseline, feat	0.9020 ± 0.0092	0.5185 ± 0.0169	0.4501 ± 0.0075
Baseline, mod mask	0.9021 ± 0.0100	0.5095 ± 0.0187	0.4408 ± 0.0104

Table 5.3.: Results on TUSL

Models	AUROC	AUPRC	Accuracy
Full, no feat	0.7058 ± 0.0971	0.2791 ± 0.0184	0.4166 ± 0.0530
Full, feat	0.6730 ± 0.0527	0.2720 ± 0.0064	0.3266 ± 0.0794
Full, mod mask	0.6995 ± 0.0367	**0.2895** ± **0.0172**	0.3650 ± 0.0806
Baseline, no feat	0.7053 ± 0.0409	0.2722 ± 0.0075	**0.4960** ± **0.0754**
Baseline, feat	**0.7347** ± **0.0451**	0.2816 ± 0.0132	0.3308 ± 0.1088
Baseline, mod mask	0.6902 ± 0.0330	0.2718 ± 0.0054	0.3290 ± 0.0496

the hypothesis according to which training a model on more data, as it was done for the full model, would improve the quality of the learned representations of the network. In fact, while the baseline model, trained on less data, was not robust to modality masking for what concerns the TES classification task, the full model was instead more robust, keeping comparable performance levels on the TES dataset despite being trained on less data.

For the EEG only datasets, we can observe that there are no significant differences between the different models for the TUAB dataset, while for the TUAR dataset, the models without the features seem to perform better than the versions that include them. For the TUSL instead, the baseline model (with features) is the one that achieves highest AUROC and Accuracy among them, appearing to be performing worse than the full model.

5.5. Comparison to related work

In this section, the proposed method is compared to state-of-the-art-works. From Table 5.8 we can observe than the proposed multimodal model achieves results comparable with the state of the art for the TUAB dataset, specifically slightly higher than Femba-

Table 5.4.: Results on TES

Models	AUROC	AUPRC	Accuracy	Recall
Full, no feat	0.8514 ± 0.1083	0.8892 ± 0.0659	0.7024 ± 0.2230	0.7506 ± 0.1544
Full, feat	**0.9746** ± 0.0308	**0.9878** ± 0.0147	0.9167 ± 0.0660	**0.9028** ± **0.0481**
Full, mod mask	0.9737 ± 0.0105	0.9870 ± 0.0057	**0.9259** ± **0.0150**	0.8982 ± 0.0212
Baseline, no feat	0.5077 ± 0.0652	0.6986 ± 0.0187	0.2949 ± 0.0084	0.4888 ± 0.0156
Baseline, feat	0.8012 ± 0.1133	0.8478 ± 0.0939	0.6789 ± 0.3117	0.7091 ± 0.2089
Baseline, mod mask	0.7417 ± 0.1210	0.7980 ± 0.1015	0.4437 ± 0.2521	0.5620 ± 0.1695

Huge, but still lower then LaBram-Huge and CBraMod.

From Table 5.9 we can observe the performance of the proposed model on TUAR and TUSL datasets. For TUSL, we can observe that the model achieves comparable performance to state of the art for the AUROC metric, but still lower performance than EEGFormer-Base. For TUAR instead we can observe that the model achieves slightly higer AUROC tham previous methods, including EEGFormer-Base and Femba-Large, and most importantly it achieves +3.2% AUPRC with respect to previous works. For th CAP dataset comparison in Table 5.11, we can see that the performance achieved by the model in terms of F1 score, is consistently lower than the one achieved by the present model.

Lastly, the TES dataset performance compasison is reported in Table 5.10, where we can see that the performance of the model is inferior to the one reported in the GIST paper.

5.6. Current limitations

This project presents some limitations, which are discussed below. First, the EEG-only self-supervised pretraining phase was conducted on the full TUEG dataset, including data from the TUAR and TUSL subsets later used for finetuning and evaluation. Although no labels from the evaluation subset were used during pretraining since the pretraining process remained entirely unsupervised, this overlap may have caused the model to have implicitly adapted to the distribution of the finetuning data, even without accessing labels. For future work, the EEG-only encoder should be pretrained on the TUEG subset that excludes subject in TUAR and TUSL, allowing a more rigorous assessment of the model's ability to generalize to entirely unseen data.

Second, due to time and resource constraints, the full model (shared encoder pre-

Table 5.5.: Results on CAP

Models	AUROC	AUPRC	Accuracy	F1 Score
Full, no feat	**0.6753** ± **0.0132**	**0.3293** ± **0.0117**	**0.3075** ± **0.0109**	**0.2773** ± **0.0110**
Full, feat	0.6196 ± 0.0631	0.2812 ± 0.0493	0.2512 ± 0.0475	0.1866 ± 0.0840
Full, mod mask	0.6554 ± 0.0074	0.3059 ± 0.0078	0.2778 ± 0.0161	0.2369 ± 0.0159
Baseline, no feat	0.6663 ± 0.0064	0.3174 ± 0.0033	0.2950 ± 0.0084	0.2637 ± 0.0083
Baseline, feat	0.6715 ± 0.0108	0.3208 ± 0.0118	0.2912 ± 0.0122	0.2498 ± 0.0146
Baseline, mod mask	0.6583 ± 0.0160	0.3085 ± 0.0135	0.2837 ± 0.0093	0.2456 ± 0.0101

Table 5.6.: Results on DREAMER

Models	AUROC	AUPRC	Accuracy
Full, no feat	0.4217 ± 0.0372	0.3227 ± 0.0264	0.3583 ± 0.0085
Full, feat	0.4464 ± 0.0540	0.3288 ± 0.0318	**0.3593** ± **0.0000**
Full, mod mask	0.4184 ± 0.0899	0.3219 ± 0.0595	**0.3593** ± **0.0000**
Baseline, no feat	0.4324 ± 0.0283	0.3274 ± 0.0170	0.3510 ± 0.0138
Baseline, feat	**0.4868** ± **0.0352**	**0.3874** ± **0.0147**	**0.3593** ± **0.0000**
Baseline, mod mask	0.4519 ± 0.0299	0.3538 ± 0.0280	**0.3593** ± **0.0000**

training) could not be pretrained for a higher number of epochs, which renders the reconstructions and the finetuning results suboptimal.

Third, the model was designed with the aim to learn EEG and ECG, focusing on intra-model knowledge and inter-model knowledge, which motivates the use of the modality-specific encoders, and the cross-attention in the shared encoder. These steps are not inherently modality-agnostic. Despite this, with the explicit objective of fusing the knowledge of the two modalities, a global self-attention was also introduced in the shared encoder, with the objective of learning a shared latent representation in a modality-agnostic fashion.

Last, the limited availability of multimodal models tailored to EEG and ECG, and of multimodal datasets, especially labeled, with EEG and ECG, limits the comparisons that can be made with the present project.

Table 5.7.: Performance Comparison on the DREAMER Dataset

Metric	Model		
	Proposed Model	Emotion Recognition [3]	Bi-ANN [16]
Valence			
Accuracy	60.69	99.47	92.68
F1	0.0	99.41	90.91
AUROC	58.06	–	–
Arousal			
Accuracy	51.55	99.47	92.95
F1	7.72	99.38	91.61
AUROC	63.52	–	–
Dominance			
Accuracy	51.41	–	–
F1	62.61	–	–
AUROC	65.21	–	–

Note: Accuracy, F1-Score (macro), e AUROC (macro) are given for each prediction (Valence, Arousal, Dominance). The values '–' indicate the absence of the data in the studies here reported.

Table 5.8.: TUAB Literature comparison

Models	TUAB AUROC	TUAB AUPRC	TUAB Accuracy
Multimodal model (no feat)	0.8949	0.9036	0.8229
Multimodal model (features)	0.8936	0.9031	0.8227
Multimodal model (mod. mask)	0.8968	0.9059	0.8205
CBraMod [80]	0.9156	**0.9221**	0.8249
LaBraM-Huge [73]	**0.9162**	0.9204	**0.8258**
Femba-Huge [74]	0.8921	0.9005	0.8182

Table 5.9.: TUAR and TUSL literature comparison

Models	TUAR AUROC	TUAR AUPRC	TUSL AUROC	TUSL AUPRC
Multimodal model (no feat)	**0.9161**	**0.5909**	0.7058	0.2791
Multimodal model (features)	0.8966	0.5048	0.673	0.272
Multimodal model (mod. mask)	0.8925	0.5072	0.6995	0.2895
Femba-Large [74]	0.915	0.521	0.714	0.282
Femba-Base [74]	0.900	0.559	**0.731**	0.289
EEGFormer-Base [22]	0.847	0.483	0.713	**0.393**

5.7. Pretraining and Finetuning setup

The hyperparameters used in pretraining and finetuning are summarized respectively in Table 5.12, for the pretraining setup, in Table 5.13 for the EEG-only finetuning datasets and in Table 5.14 for the EEG and ECG finetuning datasets.

Table 5.10.: TES literature comparison

Models	**TES Accuracy**	**TES Recall**
Multimodal model (no feat)	0.7024	0.7506
Multimodal model (features)	0.9167	0.9028
Multimodal model (mod. mask)	0.9259	0.8982
GIST [81]	**0.994**	**0.988**

Table 5.11.: CAP literature comparison

Models	**F1 Score**
Multimodal model (no feat)	0.2773 ± 0.0110
Multimodal model (features)	0.1866 ± 0.0840
Multimodal model (mod. mask)	0.2369 ± 0.0159
U-Time [82]	0.68
CNN [83]	0.68

Table 5.12.: Hyperparameters for EEG and ECG pretraining

Hyperparameters	**EEG encoder**	**ECG Encoder**	**Shared Encoder**
Non-masked regions loss weight	0.1		
Patch size	64		
Number of layers	10		
Embedding dimension	768		
Attention type	Two Axis	Two Axis	Hybrid
Attention head number	12		
Image size	1280		
Model drop path	0.2		
Layerwise learning rate decay	None	None	0.9
Batch size	350	8000 (Icentia-11k), 512 (PTB-XL)	1024
Gradient accumulation	8		
Learning rate scheduler	CosineAnnealingLR		
Base learning rate	10e-4		
Warmup epochs	5		
Warmup initial learning rate	10e-6		
Minimum learning rate	10e-6		
Optimizer	AdamW		
Optimizer β	[0.9, 0.999]		
Weight decay	0.05		
Total epochs	195	223	52
Loss type	L2		
Mask ratio	0.5		
Precision	bf-16mixed		

Table 5.13.: Hyperparameters for EEG finetuning

Hyperparameters	**TUAB**	**TUAR**	**TUSL**
Model drop path	0.2	0.1	0.2
Layerwise learning rate decay	0.5	0.7	0.5
Batch size	128	128	128
Gradient accumulation	8		
Learning rate scheduler	CosineAnnealingLR		
Base learning rate	10e-4		
Warmup epochs	5		
Warmup initial learning rate	10e-5		
Minimum learning rate	10e-5		
Optimizer	AdamW		
Optimizer β	[0.9, 0.999]		
Weight decay	0.05		
Max epochs	20	20	20
Loss type	Focal	Focal	Focal
Focal loss weights	[1.42, 1.41]	[1.36, 9.99, 3.61, 2.27, 2.38, 45.30]	[0.01, 0.9, 1.1, 0.85]
Precision	bf-16mixed		

Table 5.14.: Hyperparameters for EEG and ECG finetuning

Hyperparameters	**TES**	**CAP**	**DREAMER**
Model drop path	0.2	0.2	0.2
Layerwise learning rate decay	0.8	0.7	0.75
Batch size	128	256	128
Gradient accumulation	8		
Learning rate scheduler	CosineAnnealingLR		
Base learning rate	10e-4		
Warmup epochs	5		
Warmup initial learning rate	10e-5		
Minimum learning rate	10e-5		
Optimizer	AdamW		
Optimizer β	[0.9, 0.999]		
Weight decay	0.05		
Max epochs	50	50	50
Loss type	Focal	Cross-entropy	Focal
Focal loss weights	[1.42, 1.41]	None	[0.78, 1.40]
Label smoothing	None	0.1	None
Precision	bf-16mixed		

Chapter 6

Conclusion and Future Work

This thesis addressed the challenge of developing generalizable AI models for diverse biosignals by proposing a novel multi-modal foundation model for simultaneous EEG and ECG analysis.

We introduced a Transformer-based architecture with initial modality-specific encoders and a shared multimodal encoder, trained through a multi-stage self-supervised learning strategy on large-scale unimodal (TUEG, Icentia-11k, PTB-XL) and multimodal (SHHS) datasets using masked reconstruction. This approach was able to learn rich, shared multimodal representations capable of integrating information from both signal types. Experimental evaluations across unimodal and multimodal finetuning tasks (including abnormality detection, artifact classification, and transcranial electrical stimulation region classification), validated the model's flexibility and the generalizability of its learned representations.

The key contributions of this project include the proposed multimodal architecture and multi-stage SSL pretraining strategy tailored for EEG and ECG, which employed four large-scale unimodal and multimodal datasets for pretraining, specifically TUEG (EEG), Icentia-11k and PTB-XL (ECG), and SHHS (EEG and ECG), and six unimodal and multimodal datasets for finetuning, respectively, TUAB, TUAR, TUSL, and TES, CAP and DREAMER.

Future works should primarily focus on evaluating the proposed approach on unimodal ECG datasets, and on leveraging explainability or interpretability methods to understand what the model "focuses on", for instance, in the attention mechanism. This could contribute to making clearer the role of the multimodal interaction, possibly highlighting correlations between EEG and ECG that the model is making use of for its predictions.

In conclusion, this work shows the potential of the proposed multimodal EEG and ECG Foundation Model in both multimodal and unimodal-EEG finetuning tasks, together with the robustness shown while using either one or both modalities at the same time, as shown by the increase of 3.2% obtained in AUPRC for the artifact classification task in the TUAR unimodal EEG dataset with respect to state-of-the-art works.

Appendix A

Task Description

Master Project at the Department of
Information Technology and Electrical Engineering

Autumn Semester 2024

Federica Palmisano

Modality-Agnostic Framework for Biological Signal Analysis

October 2, 2024

Advisors:	Thorir Mar Ingolfsson, thoriri@iis.ee.ethz.ch Dr. Yawei Li, yawei.li@vision.ee.ethz.ch
Supervisor:	Prof. Dr. Luca Benini, lbenini@iis.ee.ethz.ch

Handout:	October XX, 2024
Due:	April XX, 2025

The final report will be submitted in electronic format. All copies remain property of the Integrated Systems Laboratory.

Electroencephalography (EEG) is a non-invasive method used to record electrical activity in the brain, playing a critical role in both neurological research and clinical diagnostics [1]. By offering a window into the brain's activity, EEG helps diagnose and treat various neurological disorders. However, the analysis of EEG signals presents substantial challenges due to the complexity of these signals and the fine distinctions required between normal and abnormal brain activity [2].

Recent advancements in artificial intelligence and deep learning have opened up new possibilities for analyzing EEG data. Foundation models—large, pre-trained neural networks that can be fine-tuned for specific tasks—have reshaped fields such as natural language processing and computer vision [3]. Their potential to improve EEG analysis is promising but still in its early stages.

Progress has been made with models like LaBraM, which segments EEG signals into channel patches to enable cross-dataset learning and fine-tuning [4]. Additionally, Brant-X offers a unified physiological signal alignment framework by leveraging EEG alongside other physiological signals, improving performance in EEG classification tasks like sleep stage and emotion recognition [5]. Another model, EEGFormer, takes a self-supervised learning approach with vector quantization to pre-train on large-scale EEG datasets, producing state-of-the-art results in tasks like seizure detection [6].

Despite these advancements, significant challenges in EEG analysis remain:

1. **Limited Multi-Modality Integration**: Due to the scarcity of datasets containing multiple modalities recorded simultaneously, existing models often cannot leverage cross-modal information effectively.

2. **Data Augmentation**: The application of advanced data augmentation techniques in EEG analysis is underexplored, limiting the diversity and robustness of training data.

This project seeks to address these issues by developing a foundation model capable of learning from different modalities independently (EEG, iEEG, PPG, ECG) and leveraging shared representations to improve performance on downstream tasks, even when only one modality is available.

1 Project Description

The project aims to enhance EEG signal analysis through three primary objectives:

Objective 1: Develop a modality-agnostic foundation model capable of transferring knowledge across different physiological signal modalities (e.g., EEG, iEEG, PPG, ECG). The model should be designed to be pre-trained on one or more modalities and effectively fine-tuned on datasets consisting of only another modality, leveraging shared representations to enhance performance on downstream tasks.

Objective 2: Utilize selected data augmentation techniques to enrich the training data and improve model robustness, focusing on practical and feasible methods within the project timeline.

The primary tasks for this project are as follows:

- **Task I: Literature Review**
  Conduct a comprehensive review of recent advancements related to:
  - Focus on modality-agnostic models and transfer learning techniques.
  - Review practical data augmentation methods applicable to physiological signals.
- **Task II: Dataset Collection and Preprocessing**
  Gather and preprocess datasets to ensure diversity in subjects and modalities:
  - **Data Acquisition**: Collect publicly available datasets for EEG and one to two other modalities (such as iEEG, PPG, and ECG), even if they are not recorded simultaneously.
  - **Dataset Alignment**: Preprocess the data, ensuring consistency across modalities.
- **Task III: Modality-Agnostic Model Development**
  Develop a foundation model that can transfer knowledge across different modalities, allowing pre-training on certain modalities and fine-tuning on others:
  - **Design a Modality-Agnostic Architecture**:
    - Create a unified model with a shared encoder capable of processing various modalities.
    - Implement modality-specific adaptation layers for fine-tuning on new modalities.
  - **Pre-Training on Available Modalities**:
    - Train the model using available datasets (e.g., EEG, iEEG) to learn robust, generalizable features.
    - Focus on capturing shared patterns across physiological signals.
  - **Fine-Tuning on New Modalities**:
    - Adapt the pre-trained model to new modalities (e.g., PPG-only datasets) through fine-tuning.
    - Utilize the shared representations to improve learning efficiency and performance.

  - **Implement Cross-Modal Learning Techniques**:
    - Use multi-task learning, contrastive learning, or meta-learning to align representations across modalities.
    - Encourage the model to learn modality-invariant features.
  - **Evaluate Transfer Learning Effectiveness**:
    - Compare performance on new modalities against models trained from scratch.
    - Assess the benefits of shared representations and pre-training.

- **Task IV: Simplified Data Augmentation Techniques**
  Enhance training data using data augmentation methods:
  - **Implement Basic Augmentation Methods**:
    - Apply straightforward augmentation techniques like noise injection, scaling, and time-shifting.
  - **Evaluation of Augmented Data**:
    - Assess the quality and utility of synthetic data in improving model performance for each modality.
    - Compare models trained with augmented data to those trained on original data.

- **Task V: Evaluation and Performance Analysis**
  Conduct comprehensive evaluations to assess the model's effectiveness:
  - **Benchmarking**:
    - Compare the proposed model with existing state-of-the-art models on standard tasks for each modality.
    - Evaluate performance using metrics such as accuracy, F1-score, AUC, and computational efficiency.
  - **Ablation Studies**:
    - Analyze the impact of shared representations on downstream task performance.
    - Investigate the contribution of data augmentation techniques.
  - **Modality Transfer Evaluation**:
    - Test the model's ability to leverage knowledge from one modality to improve performance on another.

    - ∗ Assess the practical benefits of the modality-agnostic approach.

- **Task VIII: Report and Presentation Work**
  Prepare documentation and presentations to communicate the project's findings. Highlight key insights, challenges, and future directions for self-supervised pre-training in EEG research.

Expected Outcomes

- **A Modality-Agnostic Foundation Model**: A model capable of learning from different physiological signals independently, utilizing shared representations to improve downstream task performance even when only one modality is available.

- **Enhanced Training Data**: Augmented datasets for each modality demonstrating the effectiveness of data augmentation techniques.

- **Academic Contributions**: Potential publications and presentations contributing to the fields of EEG analysis and multi-modal deep learning.

2 Project Realization

2.1 Meetings

Weekly meetings and reports must be held. The exact time and location of these meetings will be determined within the first week of the project in order to fit the student's and the assistant's schedule. These meetings will be used to evaluate the status and progress of the project. Besides these regular meetings, additional meetings can be organized to address urgent issues as well.

2.2 Report

Documentation is an important and often overlooked aspect of engineering. One final report has to be completed within this project. The common language of engineering is de facto English. Therefore, the final report of the work is preferred to be written in English. Any form of word processing software is allowed for writing the reports, nevertheless, the use of LaTeX with Tgif[1] or any other vector drawing software (for block diagrams) is strongly encouraged by the IIS staff.

[1] See: `http://bourbon.usc.edu:8001/tgif/index.html` and `http://www.dz.ee.ethz.ch/en/information/how-to/drawing-schematics.html`.

Final Report The final report has to be presented at the end of the project and a digital copy need to be handed in. Note that this task description is part of your report and has to be attached to your final report.

2.3 Presentation

There will be a presentation (15 min for the semester thesis, and 20 min for the MS thesis presentation followed by 5 min Q&A) at the end of this project in order to present your results to a wider audience. The exact date will be determined towards the end of the work.

References

[1] S. Weisdorf, S. W. Gangstad, J. Duun-Henriksen, K. S. S. Mosholt, and T. W. Kjær, "High similarity between eeg from subcutaneous and proximate scalp electrodes in patients with temporal lobe epilepsy," *Journal of Neurophysiology*, vol. 120, no. 3, pp. 1451–1460, 2018.

[2] T. M. Ingolfsson, S. Benatti, X. Wang, A. Bernini, P. Ducouret, P. Ryvlin, S. Beniczky, L. Benini, and A. Cossettini, "Minimizing artifact-induced false-alarms for seizure detection in wearable eeg devices with gradient-boosted tree classifiers," *Scientific Reports*, vol. 14, no. 1, p. 2980, 2024.

[3] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[4] W.-B. Jiang, L.-M. Zhao, and B.-L. Lu, "Large brain model for learning generic representations with tremendous eeg data in bci," *arXiv preprint arXiv:2405.18765*, 2024.

[5] D. Zhang, Z. Yuan, J. Chen, K. Chen, and Y. Yang, "Brant-x: A unified physiological signal alignment framework," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 4155–4166.

[6] Y. Chen, K. Ren, K. Song, Y. Wang, Y. Wang, D. Li, and L. Qiu, "Eegformer: Towards transferable and interpretable large-scale eeg foundation model," *arXiv preprint arXiv:2401.10278*, 2024.

Zurich, October 2, 2024 Prof. Dr. Luca Benini

List of Figures

List of Tables

Bibliography

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[2] S. Gkikas and M. Tsiknakis, "Twins-PainViT: Towards a Modality-Agnostic Vision Transformer Framework for Multimodal Automatic Pain Assessment using Facial Videos and fNIRS," 2024. [Online]. Available: https://arxiv.org/abs/2407.19809

[3] L. Gong, W. Chen, M. Li, and T. Zhang, "Emotion recognition from multiple physiological signals using intra- and inter-modality attention fusion network," *Digital Signal Processing*, vol. 144, p. 104278, Jan. 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1051200423003731

[4] T. Goetz, P. Arora, F. X. Erick, N. Holzer, and S. Sawant, "Self-supervised representation learning using multimodal Transformer for emotion recognition," in *Proceedings of the 8th international Workshop on Sensor-Based Activity Recognition and Artificial Intelligence*. Lübeck Germany: ACM, Sep. 2023, pp. 1–8. [Online]. Available: https://dl.acm.org/doi/10.1145/3615834.3615837

[5] E. N. Bruce, *Biomedical Signal Processing and Signal Modeling*. New York: John Wiley & Sons, 2001.

[6] L. Clifton, D. A. Clifton, M. A. F. Pimentel, P. J. Watkinson, and L. Tarassenko, "Gaussian process regression in vital-sign early warning systems," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. San Diego, CA: IEEE, Aug. 2012, pp. 6161–6164. [Online]. Available: http://ieeexplore.ieee.org/document/6347400/

[7] P. L. Nunez and R. Srinivasan, *Electric Fields of the Brain: The Neurophysics of EEG*, 2nd ed. New York: Oxford University Press, 2006.

[8] J. Xu, S. Mitra, C. Van Hoof, R. F. Yazicioglu, and K. A. A. Makinwa, "Active Electrodes for Wearable EEG Acquisition: Review and Electronics Design Methodology," *IEEE Reviews in Biomedical Engineering*, vol. 10, pp. 187–198, 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7828037/

[9] Y. Potter, G. Zerveas, C. Eickhoff, and D. Duncan, "Unsupervised Multivariate Time-Series Transformers for Seizure Identification on EEG," in *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. Nassau, Bahamas: IEEE, Dec. 2022, pp. 1304–1311. [Online]. Available: https://ieeexplore.ieee.org/document/10069282/

[10] Y. Ma, C. Liu, M. S. Ma, Y. Yang, N. D. Truong, K. Kothur, A. Nikpour, and O. Kavehei, "TSD: Transformers for Seizure Detection," Jan. 2023. [Online]. Available: http://biorxiv.org/lookup/doi/10.1101/2023.01.24.525308

[11] J. A. Urigüen and B. Garcia-Zapirain, "EEG artifact removal—state-of-the-art and guidelines," *Journal of Neural Engineering*, vol. 12, no. 3, p. 031001, Jun. 2015. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2560/12/3/031001

[12] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, "BENDR: Using Transformers and a Contrastive Self-Supervised Learning Task to Learn From Massive Amounts of EEG Data," *Frontiers in Human Neuroscience*, vol. 15, p. 653659, 2021, 143 citations (Semantic Scholar/DOI) [2024-09-24] 96 citations (Crossref) [2024-09-24].

[13] J. Xie, J. Zhang, J. Sun, Z. Ma, L. Qin, G. Li, H. Zhou, and Y. Zhan, "A Transformer-Based Approach Combining Deep Learning Network and Spatial-Temporal Information for Raw EEG Classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 2126–2136, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9845479/

[14] P. Deny and K. W. Choi, "Hierarchical Transformer for Brain Computer Interface," in *2023 11th International Winter Conference on Brain-Computer Interface (BCI)*. Gangwon, Korea, Republic of: IEEE, Feb. 2023, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/10078473/

[15] J. Liu, H. Wu, L. Zhang, and Y. Zhao, "Spatial-temporal Transformers for EEG Emotion Recognition," in *2022 The 6th International Conference on Advances in Artificial Intelligence*. Birmingham United Kingdom: ACM, Oct. 2022, pp. 116–120. [Online]. Available: https://dl.acm.org/doi/10.1145/3571560.3571577

[16] X. Zhong, Y. Gu, Y. Luo, X. Zeng, and G. Liu, "Bi-hemisphere asymmetric attention network: recognizing emotion from EEG signals based on the transformer," *Applied Intelligence*, vol. 53, no. 12, pp. 15 278–15 294, Jun. 2023. [Online]. Available: https://link.springer.com/10.1007/s10489-022-04228-2

[17] D. Zhang, Z. Yuan, J. Chen, K. Chen, and Y. Yang, "Brant-X: A Unified Physiological Signal Alignment Framework," Aug. 2024, 0 citations (Semantic Scholar/arXiv) [2024-09-24] arXiv:2409.00122 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2409.00122

[18] Y. Song, Q. Zheng, B. Liu, and X. Gao, "EEG Conformer: Convolutional Transformer for EEG Decoding and Visualization," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 710–719, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/9991178/

[19] H. Phan, K. P. Lorenzen, E. Heremans, O. Y. Chén, M. C. Tran, P. Koch, A. Mertins, M. Baumert, K. B. Mikkelsen, and M. De Vos, "L-SeqSleepNet: Whole-cycle Long Sequence Modeling for Automatic Sleep Staging," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 10, pp. 4748–4757, Oct. 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10210638/

[20] Z. Yuan, F. Shen, M. Li, Y. Yu, C. Tan, and Y. Yang, "BrainWave: A Brain Signal Foundation Model for Clinical Applications," Sep. 2024, 2 citations (Semantic Scholar/arXiv) [2024-09-24] arXiv:2402.10251 [cs, eess, q-bio]. [Online]. Available: http://arxiv.org/abs/2402.10251

[21] L. Benfenati, T. M. Ingolfsson, A. Cossettini, D. J. Pagliari, A. Burrello, and L. Benini, "BISeizuRe: BERT-Inspired Seizure Data Representation to Improve Epilepsy Monitoring," Jun. 2024, arXiv:2406.19189 [cs]. [Online]. Available: http://arxiv.org/abs/2406.19189

[22] Y. Chen, K. Ren, K. Song, Y. Wang, Y. Wang, D. Li, and L. Qiu, "EEGFormer: Towards Transferable and Interpretable Large-Scale EEG Foundation Model," Mar. 2024. [Online]. Available: https://openreview.net/forum?id=MXRy6bYBfB

[23] B. Surawicz, R. Childers, B. J. Deal, and L. S. Gettes, "AHA/ACCF/HRS Recommendations for the Standardization and Interpretation of the Electrocardiogram," *Journal of the American College of Cardiology*, vol. 53, no. 11, pp. 976–981, Mar. 2009. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0735109708041351

[24] R. Ghorbani Afkhami, G. Azarnia, and M. A. Tinati, "Cardiac arrhythmia classification using statistical and mixture modeling features of ECG signals," *Pattern Recognition Letters*, vol. 70, pp. 45–51, Jan. 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0167865515004043

[25] M. Zehender, T. Meinertz, J. Keul, and H. Just, "ECG variants and cardiac arrhythmias in athletes: Clinical relevance and prognostic importance," *American Heart Journal*, vol. 119, no. 6, pp. 1378–1391, Jun. 1990. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0002870305801899

[26] N. Rafie, A. H. Kashou, and P. A. Noseworthy, "ECG Interpretation: Clinical Relevance, Challenges, and Advances," *Hearts*, vol. 2, no. 4, pp. 505–513, Nov. 2021. [Online]. Available: https://www.mdpi.com/2673-3846/2/4/39

[27] S. Kaplan Berkaya, A. K. Uysal, E. Sora Gunal, S. Ergin, S. Gunal, and M. B. Gulmezoglu, "A survey on ECG analysis," *Biomedical Signal*

Processing and Control, vol. 43, pp. 216–235, May 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1746809418300636

[28] T. Liu, Y. Si, W. Yang, J. Huang, Y. Yu, G. Zhang, and R. Zhou, "Inter-Patient Congestive Heart Failure Detection Using ECG-Convolution-Vision Transformer Network," *Sensors*, vol. 22, no. 9, p. 3283, Apr. 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/9/3283

[29] J. Guan, W. Wang, P. Feng, X. Wang, and W. Wang, "Low-Dimensional Denoising Embedding Transformer for ECG Classification," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, ON, Canada: IEEE, Jun. 2021, pp. 1285–1289. [Online]. Available: https://ieeexplore.ieee.org/document/9413766/

[30] A. Natarajan, G. Boverman, Y. Chang, C. Antonescu, and J. Rubin, "Convolution-Free Waveform Transformers for Multi-Lead ECG Classification," in *2021 Computing in Cardiology (CinC)*. Brno, Czech Republic: IEEE, Sep. 2021, pp. 1–4. [Online]. Available: https://ieeexplore.ieee.org/document/9662697/

[31] L. Meng, W. Tan, J. Ma, R. Wang, X. Yin, and Y. Zhang, "Enhancing dynamic ECG heartbeat classification with lightweight transformer model," *Artificial Intelligence in Medicine*, vol. 124, p. 102236, Feb. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S093336572200001X

[32] R. Alcaraz, F. Sandberg, L. Sörnmo, and J. J. Rieta, "Classification of Paroxysmal and Persistent Atrial Fibrillation in Ambulatory ECG Recordings," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 5, pp. 1441–1449, May 2011. [Online]. Available: http://ieeexplore.ieee.org/document/5710968/

[33] F. Murat, F. Sadak, O. Yildirim, M. Talo, E. Murat, M. Karabatak, Y. Demir, R.-S. Tan, and U. R. Acharya, "Review of Deep Learning-Based Atrial Fibrillation Detection Studies," *International Journal of Environmental Research and Public Health*, vol. 18, no. 21, p. 11302, Oct. 2021. [Online]. Available: https://www.mdpi.com/1660-4601/18/21/11302

[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[35] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," 2018. [Online]. Available: https://arxiv.org/abs/1807.03748

[36] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha,

T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang, "On the Opportunities and Risks of Foundation Models," 2021. [Online]. Available: https://arxiv.org/abs/2108.07258

[37] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, Jul. 2021, pp. 8748–8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[38] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," 2018. [Online]. Available: https://arxiv.org/abs/1801.06146

[39] A. Anwar, Y. Khalifa, J. L. Coyle, and E. Sejdic, "Transformers in biosignal analysis: A review," *Information Fusion*, vol. 114, p. 102697, Feb. 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253524004755

[40] A. Samanta, M. Karlov, M. Ravikumar, C. M. Clarke, J. Rajadas, and K. Hassani, "MVMTnet: A Multi-variate Multi-modal Transformer for Multi-class Classification of Cardiac Irregularities Using ECG Waveforms and Clinical Notes," 2023. [Online]. Available: https://arxiv.org/abs/2302.11021

[41] Z. Li, G. Zhang, S. Okada, L. Wang, B. Zhao, and J. Dang, "MBCFNet: A Multimodal Brain–Computer Fusion Network for human intention recognition," *Knowledge-Based Systems*, vol. 296, p. 111826, Jul. 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S095070512400460X

[42] S. Latif, A. Zaidi, H. Cuayahuitl, F. Shamshad, M. Shoukat, and J. Qadir, "Transformers in Speech Processing: A Survey," 2023. [Online]. Available: https://arxiv.org/abs/2303.11607

[43] H. Han, M. Xiang, C. Lian, D. Liu, and Z. Zeng, "A Multimodal Deep Neural Network for ECG and PCG Classification With Multimodal Fusion," in *2023 13th International Conference on Information Science and Technology (ICIST)*. Cairo, Egypt: IEEE, Dec. 2023, pp. 124–128. [Online]. Available: https://ieeexplore.ieee.org/document/10367180/

[44] S. Wang, Y. Li, T. Zhang, and K. Li, “TAGformer: A Multimodal Physiological Signals Fusion Network for Pilot Stress Recognition,” *IEEE Sensors Journal*, vol. 24, no. 13, pp. 20 842–20 854, Jul. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10530519/

[45] O. Ali, M. Saif-ur Rehman, T. Glasmachers, I. Iossifidis, and C. Klaes, “ConTraNet: A hybrid network for improving the classification of EEG and EMG signals with limited training data,” *Computers in Biology and Medicine*, vol. 168, p. 107649, Jan. 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0010482523011149

[46] Z. Zhang, B.-S. Lin, C.-W. Peng, and B.-S. Lin, “Multi-Modal Sleep Stage Classification With Two-Stream Encoder-Decoder,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 32, pp. 2096–2105, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10552163/

[47] M.-T. Luong, H. Pham, and C. D. Manning, “Effective Approaches to Attention-based Neural Machine Translation,” 2015. [Online]. Available: https://arxiv.org/abs/1508.04025

[48] I. Obeid and J. Picone, “The Temple University Hospital EEG Data Corpus,” *Frontiers in Neuroscience*, vol. 10, May 2016. [Online]. Available: http://journal.frontiersin.org/Article/10.3389/fnins.2016.00196/abstract

[49] P. Wagner, N. Strodthoff, R. Bousseljot, W. Samek, and T. Schaeffter, “Ptb-xl, a large publicly available electrocardiography dataset (version 1.0.3),” https://physionet.org/content/ptb-xl/1.0.3/, 2022, physioNet. Accessed: 2025-04-25.

[50] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter, “PTB-XL, a large publicly available electrocardiography dataset,” *Scientific Data*, vol. 7, no. 1, p. 154, May 2020. [Online]. Available: https://www.nature.com/articles/s41597-020-0495-6

[51] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, “PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals,” *Circulation*, vol. 101, no. 23, Jun. 2000. [Online]. Available: https://www.ahajournals.org/doi/10.1161/01.CIR.101.23.e215

[52] R. Bousseljot, D. Kreiseler, and A. Schnabel, “Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet,” *Biomedizinische Technik/Biomedical Engineering*, pp. 317–318, Jul. 2009. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/bmte.1995.40.s1.317/html

[53] J. Couderc, S. Kyal, Q. Li, B. Hall, and K. Meisel, “Icentia11k single lead continuous raw electrocardiogram dataset,” https://physionet.org/content/icentia11k/1.0.0/, 2020, accessed: 2025-04-25.

[54] S. Tan, G. Androz, A. Chamseddine, P. Fecteau, A. Courville, Y. Bengio, and J. P. Cohen, “Icentia11K: An Unsupervised Representation Learning Dataset for Arrhythmia Subtype Discovery,” 2019. [Online]. Available: https://arxiv.org/abs/1910.09570

[55] Icentia Inc., “Icentia – ecg monitoring analysis solutions,” 2024, accessed: 2025-04-25. [Online]. Available: https://www.icentia.com/

[56] G.-Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley, and S. Redline, “The National Sleep Research Resource: towards a sleep data commons,” *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1351–1358, Oct. 2018. [Online]. Available: https://academic.oup.com/jamia/article/25/10/1351/5026200

[57] “The Sleep Heart Health Study: Design, Rationale, and Methods,” *Sleep*, Dec. 1997. [Online]. Available: https://academic.oup.com/sleep/article/20/12/1077/2749934/The-Sleep-Heart-Health-Study-Design-Rationale-and

[58] S. Katsigiannis and N. Ramzan, “DREAMER: A Database for Emotion Recognition Through EEG and ECG Signals From Wireless Low-cost Off-the-Shelf Devices,” *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 98–107, Jan. 2018. [Online]. Available: http://ieeexplore.ieee.org/document/7887697/

[59] M. G. Terzano, L. Parrino, A. Smerieri, R. Chervin, S. Chokroverty, C. Guilleminault, M. Hirshkowitz, M. Mahowald, H. Moldofsky, A. Rosa, R. Thomas, and A. Walters, “Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep,” *Sleep Medicine*, vol. 3, no. 2, pp. 187–199, Mar. 2002. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1389945702000035

[60] R. Poryazova, E. Werth, L. Parrino, M. G. Terzano, and C. L. Bassetti, “Cyclic alternating pattern in narcolepsy patients and healthy controls after partial and total sleep deprivation,” *Clinical Neurophysiology*, vol. 122, no. 9, pp. 1788–1793, Sep. 2011. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1388245711001751

[61] N. Gebodh, Z. Esmaeilpour, A. Datta, and M. Bikson, “Dataset of Concurrent EEG, ECG, and Behavior with Multiple Doses of transcranial Electrical Stimulation,” May 2020. [Online]. Available: https://zenodo.org/record/3837212

[62] C. Yang, M. Westover, and J. Sun, “BIOT: Biosignal Transformer for Cross-data Learning in the Wild,” in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 78 240–78 260. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/f6b30f3e2dd9cb53bbf2024402d02295-Paper-Conference.pdf

[63] S. Abbaspourazad, O. Elachqar, A. C. Miller, S. Emrani, U. Nallasamy, and I. Shapiro, “Large-scale Training of Foundation Models for Wearable Biosignals,” Mar. 2024, 9 citations (Semantic Scholar/arXiv) [2024-09-24] 9 citations (Semantic Scholar/DOI) [2024-09-24] arXiv:2312.05409 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2312.05409

[64] W. Dai, P. Chen, M. Lu, D. Li, H. Wei, H. Cui, and P. P. Liang, “CLIMB: Data Foundations for Large Scale Multimodal Clinical Foundation Models,” 2025. [Online]. Available: https://arxiv.org/abs/2503.07667

[65] C. Fang, C. Sandino, B. Mahasseni, J. Minxha, H. Pouransari, E. Azemi, A. Moin, and E. Zippi, “Promoting cross-modal representations to improve multimodal foundation models for physiological signals,” 2024. [Online]. Available: https://arxiv.org/abs/2410.16424

[66] S. H. Mostafaei, J. Tanha, and A. Sharafkhaneh, “A novel deep learning model based on transformer and cross modality attention for classification of sleep stages,” *Journal of Biomedical Informatics*, vol. 157, p. 104689, Sep. 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1532046424001072

[67] C.-H. Lee, H. Kim, B. C. Yoon, and D.-J. Kim, “Toward Foundational Model for Sleep Analysis Using a Multimodal Hybrid Self-Supervised Learning Framework,” 2025. [Online]. Available: https://arxiv.org/abs/2502.17481

[68] J. Sun, R. Wang, G. Zhao, C. Chen, Y. Qu, J. Li, X. Hu, and Y. Yu, “START: Automatic Sleep Staging with Attention-based Cross-modal Learning Transformer,” in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Istanbul, Turkiye: IEEE, Dec. 2023, pp. 1421–1428. [Online]. Available: https://ieeexplore.ieee.org/document/10385393/

[69] J. Chen, Y. Yao, M. Anderson, N. Hauglund, C. Kjaerby, V. Untiet, M. Nedergaard, and J. Luo, “sDREAMER: Self-distilled Mixture-of-Modality-Experts Transformer for Automatic Sleep Staging,” in *2023 IEEE International Conference on Digital Health (ICDH)*. Chicago, IL, USA: IEEE, Jul. 2023, pp. 131–142. [Online]. Available: https://ieeexplore.ieee.org/document/10224751/

[70] X. Cheng, W. Wei, C. Du, S. Qiu, S. Tian, X. Ma, and H. He, “VigilanceNet: Decouple Intra- and Inter-Modality Learning for Multimodal Vigilance Estimation in RSVP-Based BCI,” in *Proceedings of the 30th ACM International Conference on Multimedia*. Lisboa Portugal: ACM, Oct. 2022, pp. 209–217. [Online]. Available: https://dl.acm.org/doi/10.1145/3503161.3548367

[71] J. Yin, M. Wu, Y. Yang, P. Li, F. Li, W. Liang, and Z. Lv, “Research on Multimodal Emotion Recognition Based on Fusion of Electroencephalogram and Electrooculography,” *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–12, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10445754/

[72] A. Dimofte, G. A. Bucagu, T. M. Ingolfsson, X. Wang, A. Cossettini, L. Benini, and Y. Li, "CEReBrO: Compact Encoder for Representations of Brain Oscillations Using Efficient Alternating Attention," 2025. [Online]. Available: https://arxiv.org/abs/2501.10885

[73] W. Jiang, L. Zhao, and B.-l. Lu, "Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI," Oct. 2023. [Online]. Available: https://openreview.net/forum?id=QzTpTRVtrP

[74] A. Tegon, T. M. Ingolfsson, X. Wang, L. Benini, and Y. Li, "FEMBA: Efficient and Scalable EEG Analysis with a Bidirectional Mamba Foundation Model," 2025. [Online]. Available: https://arxiv.org/abs/2502.06438

[75] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," 2017. [Online]. Available: https://arxiv.org/abs/1708.02002

[76] I. C, "The aasm manual for the scoring of sleep and associated events: Rules, terminology, and technical specification," 2007. [Online]. Available: https://cir.nii.ac.jp/crid/1370004237604151044

[77] P. Kligfield, L. S. Gettes, J. J. Bailey, R. Childers, B. J. Deal, E. W. Hancock, M. Gerard van Herpen, J. A. Kors, P. Macfarlane, D. M. Mirvis *et al.*, "Aha/acc/hrs scientific statement," 2007.

[78] M. M. Stecker, D. Sabau, L. Sullivan, R. R. Das, O. Selioutski, F. W. Drislane, T. N. Tsuchida, and W. O. Tatum, "American Clinical Neurophysiology Society Guideline 6: Minimum Technical Standards for EEG Recording in Suspected Cerebral Death," *Journal of Clinical Neurophysiology*, vol. 33, no. 4, pp. 324–327, Aug. 2016. [Online]. Available: https://journals.lww.com/00004691-201608000-00008

[79] R. B. Berry, R. Budhiraja, D. J. Gottlieb, D. Gozal, C. Iber, V. K. Kapur, C. L. Marcus, R. Mehra, S. Parthasarathy, S. F. Quan *et al.*, "Rules for scoring respiratory events in sleep: update of the 2007 aasm manual for the scoring of sleep and associated events: deliberations of the sleep apnea definitions task force of the american academy of sleep medicine," *Journal of clinical sleep medicine*, vol. 8, no. 5, pp. 597–619, 2012.

[80] J. Wang, S. Zhao, Z. Luo, Y. Zhou, H. Jiang, S. Li, T. Li, and G. Pan, "Cbramod: A criss-cross brain foundation model for eeg decoding," 2025. [Online]. Available: https://arxiv.org/abs/2412.07236

[81] "Graph Attention Based Spatial Temporal Network for EEG Signal Representation," in *Advances in Intelligent Systems and Computing*. Cham: Springer Nature Switzerland, 2024, pp. 286–298. [Online]. Available: https://link.springer.com/10.1007/978-3-031-47508-5_23

[82] M. Perslev, M. Jensen, S. Darkner, P. J. r. Jennum, and C. Igel, “U-Time: A Fully Convolutional Network for Time Series Segmentation Applied to Sleep Staging,” in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/57bafb2c2dfeefba931bb03a835b1fa9-Abstract.html

[83] F. Andreotti, H. Phan, N. Cooray, C. Lo, M. T. M. Hu, and M. De Vos, “Multichannel Sleep Stage Classification and Transfer Learning using Convolutional Neural Networks,” in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Honolulu, HI: IEEE, Jul. 2018, pp. 171–174. [Online]. Available: https://ieeexplore.ieee.org/document/8512214/

[84] G. Peng, K. Zhao, H. Zhang, D. Xu, and X. Kong, “Temporal relative transformer encoding cooperating with channel attention for EEG emotion analysis,” *Computers in Biology and Medicine*, vol. 154, p. 106537, Mar. 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0010482523000021

[85] Y. Yang, Q. Wu, Y. Fu, and X. Chen, “Continuous Convolutional Neural Network with 3D Input for EEG-Based Emotion Recognition,” in *Neural Information Processing*, L. Cheng, A. C. S. Leung, and S. Ozawa, Eds. Cham: Springer International Publishing, 2018, vol. 11307, pp. 433–443. [Online]. Available: https://link.springer.com/10.1007/978-3-030-04239-4_39

[86] Y. Geng, S. Shi, and X. Hao, “Deep learning-based EEG emotion recognition: a comprehensive review,” *Neural Computing and Applications*, vol. 37, no. 4, pp. 1919–1950, Feb. 2025. [Online]. Available: https://link.springer.com/10.1007/s00521-024-10821-y

[87] Y.-T. Lan, W. Liu, and B.-L. Lu, “Multimodal Emotion Recognition Using Deep Generalized Canonical Correlation Analysis with an Attention Mechanism,” in *2020 International Joint Conference on Neural Networks (IJCNN)*. Glasgow, United Kingdom: IEEE, Jul. 2020, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/9207625/

[88] “Sleep Classification According to AASM and Rechtschaffen &amp; Kales: Effects on Sleep Scoring Parameters,” *Sleep*, Feb. 2009. [Online]. Available: https://academic.oup.com/sleep/article/32/2/139/2454306/Sleep-Classification-According-to-AASM-and

[89] J. Guan, W. Wang, P. Feng, X. Wang, and W. Wang, “Low-Dimensional Denoising Embedding Transformer for ECG Classification,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, ON, Canada: IEEE, Jun. 2021, pp. 1285–1289. [Online]. Available: https://ieeexplore.ieee.org/document/9413766/

[90] P. L. Nunez and R. Srinivasan, *Electric Fields of the Brain: The Neurophysics of EEG*. Oxford University Press, 2006, google-Books-ID: fUv54as56_8C.

[91] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal Transformer for Unaligned Multimodal Language Sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 6558–6569. [Online]. Available: https://www.aclweb.org/anthology/P19-1656

[92]

[93] Y. Luo, Y. Chen, A. Salekin, and T. Rahman, "Toward Foundation Model for Multivariate Wearable Sensing of Physiological Signals," 2024. [Online]. Available: https://arxiv.org/abs/2412.09758

[94] B. Surawicz, R. Childers, B. J. Deal, and L. S. Gettes, "AHA/ACCF/HRS Recommendations for the Standardization and Interpretation of the Electrocardiogram," *Journal of the American College of Cardiology*, vol. 53, no. 11, pp. 976–981, Mar. 2009. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0735109708041351

[95] W. Tao, C. Li, R. Song, J. Cheng, Y. Liu, F. Wan, and X. Chen, "EEG-Based Emotion Recognition via Channel-Wise Attention and Self Attention," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 382–393, Jan. 2023. [Online]. Available: https://ieeexplore.ieee.org/document/9204431/

[96] T. Xu, K. Zhao, Y. Hu, L. Li, W. Wang, F. Wang, Y. Zhou, and J. Li, "Transferable non-invasive modal fusion-transformer (NIMFT) for end-to-end hand gesture recognition," *Journal of Neural Engineering*, vol. 21, no. 2, p. 026034, Apr. 2024. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2552/ad39a5

[97] J. Zhang, L. Swinnen, C. Chatzichristos, W. Van Paesschen, and M. De Vos, "Learning Robust Representations of Tonic-Clonic Seizures With Cyclic Transformer," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 6, pp. 3721–3731, Jun. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10463103/

[98] J. Pradeepkumar, M. Anandakumar, V. Kugathasan, D. Suntharalingham, S. L. Kappel, A. C. De Silva, and C. U. S. Edussooriya, "Toward Interpretable Sleep Stage Classification Using Cross-Modal Transformers," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 32, pp. 2893–2904, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10623416/

[99] R. Casal, L. E. Di Persia, and G. Schlotthauer, "Temporal convolutional networks and transformers for classifying the sleep stage in awake or asleep using pulse oximetry signals," *Journal of Computational Science*, vol. 59, p. 101544, Mar. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1877750321002003

[100] Z. Yang, D. Wang, Z. Chen, M. Huang, N. Ono, M. Altaf-Ul-Amin, and S. Kanaya, "Exploring Feasibility of Truth-Involved Automatic Sleep Staging Combined with Transformer," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Houston, TX, USA: IEEE, Dec. 2021, pp. 2920–2923. [Online]. Available: https://ieeexplore.ieee.org/document/9669456/

[101] J. Zhu, Y. Feng, Q. Liu, H. Xu, Y. Miao, Z. Lin, J. Li, H. Liu, Y. Xu, and F. Li, "An Improved ConvNeXt With Multimodal Transformer for Physiological Signal Classification," *IEEE Access*, vol. 12, pp. 11 217–11 229, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10401880/

[102] K. M. Gunter, A. Brink-Kjær, E. Mignot, H. B. Sørensen, E. During, and P. Jennum, "SViT: A Spectral Vision Transformer for the Detection of REM Sleep Behavior Disorder," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 9, pp. 4285–4292, Sep. 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10172323/

[103] S. Duan, L. Wu, A. Liu, and X. Chen, "Alignment-Enhanced Interactive Fusion Model for Complete and Incomplete Multimodal Hand Gesture Recognition," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 4661–4671, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10323506/

[104] P.-L. Lee, S.-H. Chen, T.-C. Chang, W.-K. Lee, H.-T. Hsu, and H.-H. Chang, "Continual Learning of a Transformer-Based Deep Learning Classifier Using an Initial Model from Action Observation EEG Data to Online Motor Imagery Classification," *Bioengineering*, vol. 10, no. 2, p. 186, Feb. 2023. [Online]. Available: https://www.mdpi.com/2306-5354/10/2/186

[105] M. Goswami, K. Szafer, A. Choudhry, Y. Cai, S. Li, and A. Dubrawski, "MOMENT: A Family of Open Time-series Foundation Models," 2024. [Online]. Available: https://arxiv.org/abs/2402.03885

[106] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer, O. Hénaff, M. M. Botvinick, A. Zisserman, O. Vinyals, and J. Carreira, "Perceiver IO: A General Architecture for Structured Inputs &amp; Outputs," 2021. [Online]. Available: https://arxiv.org/abs/2107.14795

[107] H. Liu, S. Xu, Z. Zhao, L. Kong, H. Kamarthi, A. B. Sasanur, M. Sharma, J. Cui, Q. Wen, C. Zhang, and B. A. Prakash, "Time-MMD: Multi-Domain Multimodal Dataset for Time Series Analysis," 2024. [Online]. Available: https://arxiv.org/abs/2406.08627

[108] L. Zhu, Y. Ding, A. Huang, X. Tan, and J. Zhang, "MF-Net: a multimodal fusion network for emotion recognition based on multiple physiological signals," *Signal, Image and Video Processing*, vol. 19, no. 1, p. 58, Jan. 2025. [Online]. Available: https://link.springer.com/10.1007/s11760-024-03632-0

[109] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention Bottlenecks for Multimodal Fusion," in *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc., 2021, pp. 14 200–14 213. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/76ba9f564ebbc35b1014ac498fafadd0-Abstract.html

[110] Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong, and J. Huang, "Deep Multimodal Fusion by Channel Exchanging," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 4835–4845. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/339a18def9898dd60a634b2ad8fbbd58-Abstract.html

[111] T. Shaik, X. Tao, L. Li, H. Xie, and J. D. Velásquez, "A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom," *Information Fusion*, vol. 102, p. 102040, Feb. 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1566253523003561

[112] Q. Zhang, Y. Wei, Z. Han, H. Fu, X. Peng, C. Deng, Q. Hu, C. Xu, J. Wen, D. Hu, and C. Zhang, "Multimodal Fusion on Low-quality Data: A Comprehensive Survey," 2024. [Online]. Available: https://arxiv.org/abs/2404.18947

[113] M. Li, D. Yang, Y. Liu, S. Wang, J. Chen, S. Wang, J. Wei, Y. Jiang, Q. Xu, X. Hou, M. Sun, Z. Qian, D. Kou, and L. Zhang, "Toward Robust Incomplete Multimodal Sentiment Analysis via Hierarchical Representation Learning," 2024. [Online]. Available: https://arxiv.org/abs/2411.02793

[114] Z. Cheng, Z.-Q. Cheng, J.-Y. He, J. Sun, K. Wang, Y. Lin, Z. Lian, X. Peng, and A. Hauptmann, "Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning," 2024. [Online]. Available: https://arxiv.org/abs/2406.11161

[115] K. Hemker, N. Simidjievski, and M. Jamnik, "HEALNet: Multimodal Fusion for Heterogeneous Biomedical Data," 2023. [Online]. Available: https://arxiv.org/abs/2311.09115

[116] F. Yang, C. Feng, D. Wang, T. Wang, Z. Zeng, Z. Xu, H. Park, P. Ji, H. Zhao, Y. Li, and A. Wong, "NeuroBind: Towards Unified Multimodal Representations for Neural Signals," 2024. [Online]. Available: https://arxiv.org/abs/2407.14020

[117] A. Ballas and C. Diou, "Towards Domain Generalization for ECG and EEG Classification: Algorithms and Benchmarks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 8, no. 1, pp. 44–54, Feb. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10233054/

[118] J. Dong, H. Wu, H. Zhang, L. Zhang, J. Wang, and M. Long, "SimMTM: A Simple Pre-Training Framework for Masked Time-Series Modeling," 2023. [Online]. Available: https://arxiv.org/abs/2302.00861

[119] R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. E. Elger, "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state," *Physical Review E*, vol. 64, no. 6, p. 061907, Nov. 2001. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.64.061907

[120] K.-K. Tseng, C. Wang, Y.-F. Huang, G.-R. Chen, K.-L. Yung, and W.-H. Ip, "Cross-Domain Transfer Learning for PCG Diagnosis Algorithm," *Biosensors*, vol. 11, no. 4, p. 127, Apr. 2021. [Online]. Available: https://www.mdpi.com/2079-6374/11/4/127

[121] H. Liu, T. Lou, Y. Zhang, Y. Wu, Y. Xiao, C. S. Jensen, and D. Zhang, "EEG-Based Multimodal Emotion Recognition: A Machine Learning Perspective," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–29, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10443946/

[122] G. Nie, J. Zhu, G. Tang, D. Zhang, S. Geng, Q. Zhao, and S. Hong, "A Review of Deep Learning Methods for Photoplethysmography Data," 2024. [Online]. Available: https://arxiv.org/abs/2401.12783

[123] F. D. Pup and M. Atzori, "Applications of Self-Supervised Learning to Biomedical Signals: A Survey," *IEEE Access*, vol. 11, pp. 144 180–144 203, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10365170/

[124] F. Krones, U. Marikkar, G. Parsons, A. Szmul, and A. Mahdi, "Review of multimodal machine learning approaches in healthcare," 2024. [Online]. Available: https://arxiv.org/abs/2402.02460

[125] K. McKeen, L. Oliva, S. Masood, A. Toma, B. Rubin, and B. Wang, "ECG-FM: An Open Electrocardiogram Foundation Model," 2024. [Online]. Available: https://arxiv.org/abs/2408.05178

[126] J. Song, J.-H. Jang, B. T. Lee, D. Hong, J.-m. Kwon, and Y.-Y. Jo, "Foundation Models for Electrocardiograms," 2024. [Online]. Available: https://arxiv.org/abs/2407.07110

[127] S. Heesoo, L. Sangseok, and L. Sael, "Cross-Attention Model for Multi-modal Bio-Signal Processing," in *2022 IEEE International Conference on Big Data and Smart Computing (BigComp)*. Daegu, Korea, Republic of: IEEE, Jan. 2022, pp. 43–46. [Online]. Available: https://ieeexplore.ieee.org/document/9736473/

[128] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A Compact Convolutional Network for EEG-based Brain-Computer Interfaces," 2016. [Online]. Available: https://arxiv.org/abs/1611.08024

[129] C.-Y. Yang, P.-C. Chen, and W.-C. Huang, “Cross-Domain Transfer of EEG to EEG or ECG Learning for CNN Classification Models,” *Sensors*, vol. 23, no. 5, p. 2458, Feb. 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/5/2458

[130] “Learning Topology-Agnostic EEG Representations with Geometry-Aware Modeling.” [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/a8c893712cb7858e49631fb03c941f8d-Abstract-Conference.html

[131] C. Qin, R. Yang, W. You, Z. Chen, L. Zhu, M. Huang, and Z. Wang, “EEGUnity: Open-Source Tool in Facilitating Unified EEG Datasets Towards Large-Scale EEG Model,” 2024. [Online]. Available: https://arxiv.org/abs/2410.07196

[132] Y. Zhang, J. Gao, Z. Tan, L. Zhou, K. Ding, M. Zhou, S. Zhang, and D. Wang, “Data-Centric Foundation Models in Computational Healthcare: A Survey,” 2024. [Online]. Available: https://arxiv.org/abs/2401.02458

[133] A. Gjølbye, L. Skerath, W. Lehn-Schiøler, N. Langer, and L. K. Hansen, “SPEED: Scalable Preprocessing of EEG Data for Self-Supervised Learning,” 2024. [Online]. Available: https://arxiv.org/abs/2408.08065

[134] M. Ferrante, T. Boccato, and N. Toschi, “Towards neural foundation models for vision: Aligning EEG, MEG and fMRI representations to perform decoding, encoding and modality conversion,” Mar. 2024. [Online]. Available: https://openreview.net/forum?id=nxoKCdmteM

[135] A. Keutayeva and B. Abibullaev, “Data Constraints and Performance Optimization for Transformer-Based Models in EEG-Based Brain-Computer Interfaces: A Survey,” *IEEE Access*, vol. 12, pp. 62 628–62 647, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10509679/