Improved EEG-Based Demographic Prediction: Age and Gender from Neural Signals

by

Mojtaba Moodi

A thesis

presented to the University of Waterloo

in fulfillment of the
thesis requirement for the degree of

Master of Mathematics

in

Computational Mathematics

Waterloo, Ontario, Canada, 2025

Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Electroencephalography (EEG) provides a non-invasive window into brain function across the lifespan. Recent studies have shown that machine learning can predict individual demographics such as age and gender directly from EEG signals. For example, deep neural networks have estimated adult brain age from resting EEG with mean errors under 5-6 years (Khayretdinova et al., 2022; Zhang et al., 2024). Similarly, EEG-based gender classifiers can exceed 95% accuracy on single datasets (Niu et al., 2024; Issa et al., 2022). However, models often struggle to generalize across different EEG datasets or recording conditions due to variations in hardware, protocols, and population characteristics (Tveitstøl et al., 2025).

This thesis addresses EEG-based demographic prediction using modern machine learning and deep learning techniques, with a focus on convolutional neural networks (CNNs) and a large-scale pre-trained foundation model. We develop subject-independent classifiers to infer gender and age group from raw EEG data. Our approach leverages short one-second EEG segments as input, enabling fine-grained analysis of transient neural patterns. We design and evaluate several architectures: (i) standard CNN models trained separately for binary gender classification and three-class age-group classification, (ii) a multi-output CNN that jointly predicts age and gender in a multi-task learning framework, and (iii) a foundation model approach using the Large Brain Model (LaBraM), a recently proposed EEG transformer foundation model pre-trained on thousands of hours of multi-cohort EEG data, which we fine-tune for demographic prediction. Model performance is rigorously evaluated under both within-dataset (stratified cross-validation) and cross-dataset transfer scenarios (training and testing on different EEG datasets or conditions) to assess robustness and generalization.

Our experiments demonstrate that deep learning models can extract useful demographic information from even brief EEG segments, but also reveal important differences between tasks. In within-dataset evaluations, the CNN achieves up to 81.6% accuracy for gender and about 52% for age-group classification, while the fine-tuned LaBraM model attains higher accuracy, reaching 86.6% on gender. More importantly, the pre-trained foundation model shows superior generalization: for instance, when training on EEG recordings from cognitively active tasks and testing on independent resting-state data, LaBraM maintains

about 85% gender accuracy, compared to 75-80% for the CNN. Overall, gender prediction is considerably more accurate than age-based prediction in our study - all models perform better on gender (often by 30+ percentage points) than on age-group classification. LaBraM consistently outperforms the CNN baselines on gender detection and exhibits smaller performance drops under cross-dataset conditions, indicating that large-scale pretraining provides robust feature representations for sex-related neural patterns. In contrast, age-related EEG differences proved more subtle: the foundation model yields only modest gains (reaching 55% accuracy) over CNNs (50%) for age-group classification. Furthermore, we find that multi-task learning offers no significant benefit - the multi-output CNN performed on par with separate single-task models. Joint classification of combined age-gender categories (a 6-class problem) was the most challenging setting, with all approaches barely reaching 45-50% accuracy.

In summary, this work provides a comprehensive investigation of EEG-based demographic prediction using both custom CNNs and a fine-tuned foundation model. Our results establish new benchmarks for gender and age-group prediction from one-second EEG segments and highlight the promise of foundation models for improving cross-dataset reliability. At the same time, the findings underscore the challenge of age prediction from short EEG windows and motivate further research into capturing developmental EEG signatures.

Acknowledgements

I extend my sincere gratitude to my supervisor, Professor Ali Ghodsi, for his guidance and support throughout this work. I am also thankful to my family—especially my brother—for their steadfast support. I further acknowledge the researchers and institutions who provided the EEG datasets and open-source tools that enabled this study.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The recently introduced EEG2025 dataset, derived from the Healthy Brain Network project, provides a large-scale benchmark for EEG-based demographic modeling. It encompasses over 3,000 participants, each contributing a high-density EEG recording session with six distinct cognitive tasks. Participants range from early childhood to young adulthood (approximately 5-21 years old), and the dataset contains 1995 female and 1074 male participants, with each subject's age, sex, and handedness recorded. The tasks span both passive paradigms (resting-state with eyes open/closed, naturalistic movie viewing, visual surround suppression) and active cognitive challenges (contrast-change detection, sequence learning, and a symbol search test), covering a broad spectrum of neural function. All sessions were recorded with a 128-channel EEG system under consistent protocols, ensuring data uniformity across this multi-terabyte corpus. This unprecedented scale and diversity make EEG2025 an invaluable resource for training and evaluating models that predict age and gender from neural signals.

EEG-based demographic prediction also opens new real-world applications in user profiling and adaptive systems. Recent studies demonstrate the feasibility of such implicit profiling: (Gölz et al., 2023) classified individuals into age brackets using their EEG features with high accuracy, and (Issa et al., 2022) achieved over 95% accuracy in person-independent EEG gender recognition. With these capabilities, a system could automatically tailor the user experience based on inferred characteristics without requiring explicit input. For example, an educational application might adjust content difficulty or feedback style if the brain data suggests a child rather than an adult user. Similarly, a game or multimedia platform could modify its recommendations and interface if it detects an older versus younger neural profile. This type of passive demographic sensing enables personalization and usability improvements that go beyond traditional one-size-fits-all designs.

In human-computer interaction (HCI) scenarios, demographic-aware EEG models can enhance accessibility and personalization even further. For instance, a brain-computer inter-

face might adapt its calibration or control parameters for male vs. female users, accounting for known sex-based differences in EEG rhythms and microstate dynamics (Niu et al., 2024). Age predictions are likewise useful: an adaptive interface could interpret an elevated "brain age" from EEG as a sign of cognitive slowing or fatigue (Kang et al., 2024), and respond by simplifying the layout or providing additional guidance for an older user. Conversely, a virtual reality system could passively detect if a user is likely a child and then enable age-appropriate content filters or safety settings. By incorporating age and gender inference from brain signals, HCI systems become more context-aware - capable of subtly profiling users and tuning interactions to their neurocognitive profile - which ultimately contributes to more inclusive and adaptive user experiences.

Chapter 2

Background and Problem Definition

Electroencephalography (EEG) provides a non-invasive recording of the brain's electrical activity and is widely used in clinical diagnosis, developmental neuroscience, cognitive research, and brain-computer interfaces. Because EEG reflects both transient neural events and long-term physiological properties, recent studies have increasingly explored whether demographic attributes such as age and gender can be predicted directly from raw or minimally processed EEG data. This chapter provides relevant background on EEG signals, reviews known age- and gender-related neural differences, and defines the formal problem formulations for the demographic prediction tasks considered in this thesis.

2.1 Electroencephalography and Neural Signal Characteristics

EEG measures scalp potentials produced by synchronized post-synaptic activity of cortical pyramidal neurons. Signals are typically recorded from on the order of 16-128 electrodes arranged according to the international 10-20 system (or its high-density extensions). EEG signals exhibit oscillatory patterns across canonical frequency bands: δ (1-4 Hz), θ (4-8 Hz), α (8-12 Hz), β (12-30 Hz), and γ (> 30 Hz). These neural rhythms change systematically with development and aging. Infants and young children tend to show dominant low-frequency power and immature (slower) alpha rhythms, whereas alpha-band power and peak frequency increase through childhood and adolescence as the brain matures. In later adulthood, an overall slowing of the dominant alpha rhythm and changes in the aperiodic $1/f$ spectral slope have been observed as part of the normal aging process (Kang et al., 2024).

Modern deep learning approaches to EEG analysis often operate directly on raw multichannel time series segments, allowing models to automatically learn discriminative tempo-

ral, spectral, and spatial features. Alternatively, other approaches first extract intermediate representations—such as time-frequency spectrograms, sequences of microstates, or spectral entropy and connectivity measures—and then apply a classifier or regressor to these engineered features.

2.2 Demographic Effects in EEG

2.2.1 Age-related EEG properties

Multiple studies have demonstrated that EEG signals contain robust signatures of chronological age. Large-scale work such as Khayretdinova et al. (2022) showed that raw resting-state EEG can predict adult age with mean absolute error (MAE) below about six years across broad samples. Similarly, Zhang et al. (2024) trained a transformer-based architecture on over 18,000 overnight sleep EEG recordings and achieved mean absolute error near four years, approaching the performance of MRI-based brain-age estimates.

In pediatric populations, developmental EEG changes are even more pronounced. For example, An et al. (2025) reported that a multilayer perceptron trained on spectral power features could estimate infant and toddler age within approximately 90 days, indicating that early developmental stages leave distinctive, quantifiable signatures in spontaneous EEG.

Discrete age-group classification from EEG has likewise been demonstrated. Jusseaume and Valova (2022) achieved around 90% accuracy when assigning subjects to six age ranges using recurrent deep neural networks on a large clinical EEG dataset. Similarly, Gölz et al. (2023) showed that task-related EEG (from a Flanker inhibitory-control paradigm) could distinguish among six age brackets at levels significantly above chance.

2.2.2 Gender-related EEG properties

Gender differences in EEG arise from a complex interaction of hormonal, anatomical, and neurophysiological factors. A number of studies have shown that gender can be predicted reliably from both resting-state and task-based EEG recordings.

Microstate-based approaches are particularly effective. Niu et al. (2024) demonstrated that temporal EEG microstate sequence features combined with nonlinear complexity measures can generalize across datasets, achieving gender classification accuracies above 95%. CNN-based methods using time-frequency representations also reach very high accuracy. For example, Issa et al. (2022) reported approximately 95% gender classification accuracy using short-time Fourier transform (STFT) spectrograms as input to a stacked autoencoder and a convolutional classifier.

These results indicate that male and female EEG patterns are sufficiently consistent to be learned by models, although the physiological origins of these sex-related differences remain an active area of investigation.

2.2.3 Generalization challenges

Although EEG-based demographic models can achieve high accuracy within a single dataset, cross-dataset generalization remains an open problem. As Tveitstøl et al. (2025) showed, models trained on one EEG cohort often degrade significantly when evaluated on a different cohort, due to differences in recording hardware, population characteristics, noise profiles, and experimental protocols. This vulnerability underscores the importance of standardized preprocessing, carefully stratified data splits, and systematic evaluation of generalization in any EEG demographic prediction study.

2.3 Machine Learning Models for EEG Demographic Prediction

Several types of machine learning models have been applied to the EEG demographic prediction problem. Broadly, approaches fall into two categories: classical feature-based methods and modern deep learning architectures.

2.3.1 Classical approaches

Classical machine-learning methods rely on hand-engineered features such as band-specific spectral power, frequency band ratios, entropy measures, microstate statistics, and coherence (connectivity) metrics. Models such as support vector machines (SVMs), random forests, and logistic regression were used in early age and gender prediction studies, often achieving strong performance on curated datasets with clean signals and balanced classes (Al Zoubi et al., 2018; Wei et al., 2023).

2.3.2 Deep learning architectures

Recent work has shifted toward end-to-end deep learning models that learn directly from raw or minimally processed EEG. Key architectures include:

- **Convolutional Neural Networks (CNNs)** - CNN models operate directly on raw multi-channel EEG segments (or on time-frequency image representations). They learn

spatial-temporal filter kernels adapted to EEG data and have shown good performance for both age and gender classification tasks.

- **Recurrent Neural Networks (RNNs)** - Sequential models like LSTMs or GRUs are used when capturing the temporal dynamics of EEG is important. For instance, Jusseaume and Valova (2022) reported that a simple LSTM outperformed a CNN for multi-class age group classification on a clinical EEG dataset.

- **Transformers** - Transformer architectures can capture long-range temporal dependencies in EEG. The multi-flow Swin Transformer proposed by Zhang et al. (2024) set a new benchmark for EEG-based brain age estimation, significantly improving accuracy over prior approaches.

- **Foundation models** - Jiang et al. (2024) introduced a unified EEG foundation model called *Large Brain Model* (LaBraM), which was pre-trained on roughly 2,500 hours of EEG from about 20 datasets using a masked spectrum prediction objective. Such large-scale foundation models learn general-purpose neural representations that can be fine-tuned efficiently for specific downstream tasks, including demographic attribute prediction.

Overall, deep learning methods typically outperform classical feature-based approaches, particularly on large and heterogeneous EEG datasets where learned filters can capture complex signal variations that hand-crafted features may miss.

2.4 Problem Definition

We formally define the demographic prediction tasks addressed in this thesis.

2.4.1 Age regression (brain age prediction)

Given an EEG segment $x \in \mathbb{R}^{C \times T}$ with the participant's chronological age $y \in \mathbb{R}$, the goal is to learn a mapping

$$f_\theta : \mathbb{R}^{C \times T} \to \mathbb{R},$$

trained to minimize a regression loss (such as mean absolute error). The model's prediction $\hat{y}$ can be interpreted as the subject's "brain age."

2.4.2 Age-group classification

For certain analyses, we discretize age into $K = 3$ categories. We define the label set $\mathcal{Y} = \{0, 1, 2\}$ corresponding to child (age < 8.5 years), preteen ($8.5 \leq$ age ≤ 12.5 years), and teen (age > 12.5 years). The task is to learn a classifier

$$f_\theta : \mathbb{R}^{C \times T} \to \Delta^2,$$

where Δ^2 denotes the 3-class probability simplex (i.e., the output is a probability distribution over the three age groups).

2.4.3 Gender classification

Gender is treated as a binary variable $g \in \{0, 1\}$ representing female ($g = 0$) or male ($g = 1$). The model learns a mapping

$$f_\theta : \mathbb{R}^{C \times T} \to \Delta^1,$$

producing a two-class probability output. Training uses a binary classification loss (e.g., cross-entropy) to predict the correct gender from the EEG segment.

2.4.4 Joint demographic prediction

Joint prediction of age and gender from EEG can be approached in two ways:

- **Multi-task learning:** A single model is constructed with a shared feature-extraction backbone and two output heads (one for age or age-group, and one for gender). The model is trained to optimize both objectives simultaneously.
- **Combined multi-class classification:** Each possible (gender, age-group) pair is treated as a unique category, so that the problem becomes a single-stage classification with 6 classes (for example, "female teen" or "male child" as distinct labels).

Both formulations allow the model to capture any shared structure between the age and gender prediction tasks. In this thesis, we explore both multi-task and combined-category models for joint demographic prediction.

2.5 Summary

In summary, EEG signals carry rich temporal and spectral information that can reflect both short-term mental state and long-term individual characteristics. Recent studies (especially

since 2020) have demonstrated that modern deep learning models can achieve high accuracy on EEG-based age regression, age-group classification, and gender classification. Nonetheless, generalization across different datasets and recording conditions remains a central challenge, motivating careful dataset design and rigorous evaluation. These considerations define the tasks and methodological goals pursued in the remainder of this thesis.

Chapter 3

Related Work on EEG-Based Age and Gender Prediction

This chapter reviews the evolution of methodologies for predicting demographic attributes from electroencephalography (EEG). We classify the literature into three distinct paradigms: (i) classical feature-based machine learning, (ii) end-to-end deep learning architectures, and (iii) the emerging paradigm of large-scale foundation models. We specifically examine the physiological basis for these predictions and the persistent challenge of cross-dataset generalization.

3.1 Physiological Basis of Demographic EEG Signatures

The feasibility of predicting age and gender from EEG stems from robust neurophysiological changes that occur across the lifespan.

3.1.1 Age-Related Spectral Dynamics

Brain maturation and aging leave distinct fingerprints on the EEG power spectrum. In pediatric populations, the dominant alpha rhythm accelerates from approximately 3-4 Hz in infancy to 8-12 Hz in late adolescence. Concurrently, there is a global reduction in absolute power across low-frequency bands (δ, θ) due to cortical thinning and synaptic pruning (Kang et al., 2024). In adulthood, healthy aging is characterized by a "slowing" of the alpha peak frequency and a flattening of the aperiodic ($1/f$) spectral slope. Recent studies on high-density infant EEG have shown that these trajectories are sufficiently stable to allow for age

prediction with errors measured in days (Davoudi et al., 2025; An et al., 2025).

3.1.2 Gender Dimorphism

Sex differences in EEG are statistically significant, though subtler than age effects. Females typically exhibit higher absolute power in beta and gamma bands, a difference sometimes attributed to anatomical factors such as skull thickness (Kang et al., 2024). Additionally, Niu et al. (2024) identified sex-specific patterns in the temporal syntax of EEG microstates, suggesting distinct resting-state network dynamics.

3.2 Feature-Based Machine Learning Approaches

Prior to deep learning, research focused on "feature engineering"—the manual extraction of signal descriptors.

3.2.1 Spectral and Complexity Features

Al Zoubi et al. (2018) established a benchmark for age prediction by extracting a vast array of features, including spectral power, spectral entropy, and Higuchi fractal dimensions. Using a stack-ensemble model, they achieved a Mean Absolute Error (MAE) of 6.87 years. For gender classification, Issa et al. (2022) developed the Emotion and Gender Prediction (EGP) system using Short-Time Fourier Transform (STFT) features fed into a sparse autoencoder, achieving $> 95\%$ accuracy on the MAHNOB-HCI dataset. Similarly, Niu et al. (2024) combined microstate duration and occurrence parameters with Lempel-Ziv complexity to achieve high classification accuracy, validating the utility of nonlinear features.

3.2.2 Event-Related Potentials (ERPs)

While most work focuses on resting-state EEG, Gölz et al. (2023) demonstrated that ERP components (e.g., P300 latency/amplitude) from cognitive tasks could effectively distinguish between age groups, linking neural processing speed to demographic stage.

3.3 End-to-End Deep Learning Architectures

Deep learning models, particularly Convolutional Neural Networks (CNNs), automate feature extraction and have become the standard in the field.

3.3.1 Convolutional Neural Networks (CNNs)

Khayretdinova et al. (2022) applied 1D-CNNs to the large clinical TD-BRAIN dataset. Their analysis revealed that fusing "eyes-open" and "eyes-closed" resting states improved age prediction (MAE $\approx$ 6 years), highlighting the value of alpha reactivity. In the domain of sleep EEG, Zhang et al. (2024) achieved state-of-the-art performance (MAE 4.19 years) using a multi-flow architecture combining CNNs and Swin Transformers trained on $> 18,000$ overnight recordings. This work emphasizes that long-duration recordings (sleep) contain richer demographic information than short resting-state clips.

3.3.2 Pediatric and Clinical Applications

Deep learning has also been applied to developmental cohorts. Wei et al. (2023) used machine learning for sex and age classification in children, noting that while age grouping was feasible, sex classification remained challenging in prepubescent cohorts. Ju et al. (2025) developed the Auto-EEG-BrainAGE pipeline to quantify developmental deviations in children with Autism Spectrum Disorder (ASD).

3.3.3 The Generalization Gap

A critical limitation of current DL models is their lack of robustness. Tveitstøl et al. (2025) demonstrated that models trained on one EEG dataset often fail when tested on another (e.g., MAE increasing from 5 to 10+ years). This generalization gap is attributed to variations in recording hardware, reference schemes, and population demographics. Tveitstøl et al. (2025) found that broadband inputs generalized better than narrowband features, as they capture the robust aperiodic components of the signal.

3.4 Foundation Models: The Large Brain Model (LaBraM)

To overcome data scarcity and generalization issues, the field is adopting Foundation Models pre-trained on massive unlabeled datasets.

3.4.1 LaBraM Architecture and Mechanism

The **Large Brain Model (LaBraM)**, introduced by Jiang et al. (2024), represents the state-of-the-art in this domain. LaBraM is a transformer-based model pre-trained on approximately 2,500 hours of EEG data from over 20 diverse datasets. Its key innovations include:

- **Vector-Quantized Neural Spectrum Prediction (VQ-NSP):** LaBraM employs a neural tokenizer that segments EEG into time-domain patches. The tokenizer is trained to reconstruct the *Fourier Spectrum* (amplitude and phase) of these patches into discrete neural codes. This forces the model to learn the oscillatory structure of neural signals (Jiang et al., 2024).
- **Masked Neural Code Prediction:** The model is pre-trained via a self-supervised objective where it must predict the neural codes of masked patches based on context, facilitating the learning of long-range temporal and spatial dependencies.

3.4.2 Utility for Demographic Prediction

Unlike CNNs trained from scratch, LaBraM leverages generic representations learned from massive data scales. This pre-training makes it theoretically more robust to the hardware and task variations investigated in this thesis. By fine-tuning LaBraM, we aim to exploit these universal neural features for accurate age and gender prediction from short (one-second) EEG segments, a challenging regime where feature-based models often struggle.

3.5 Summary

The literature reveals a clear methodological trajectory: early research relied on hand-crafted spectral and microstate features (Al Zoubi et al., 2018; Niu et al., 2024), which gave way to deep architectures like CNNs and RNNs capable of end-to-end learning (Khayretdinova et al., 2022; Jusseaume and Valova, 2022). While deep models have improved within-dataset accuracy, they suffer from a critical "generalization gap," often failing when tested on unseen cohorts or recording conditions (Tveitstøl et al., 2025). The recent emergence of foundation models (Jiang et al., 2024) offers a potential remedy through large-scale pre-training, yet their comparative advantage over lightweight baselines in specific demographic tasks remains under-explored.

This thesis addresses this precise gap. We use the large-scale HBN dataset as a rigorous testbed to benchmark these opposing paradigms—parameter-efficient CNNs versus massive pre-trained Transformers. By evaluating performance on strictly held-out tasks (active vs. passive) and short one-second segments, we explicitly isolate the value of foundation-model representations in overcoming the brittleness inherent to standard deep learning.

Chapter 4

Methodology

In this chapter we describe the end-to-end methodology used for EEG-based age and gender prediction. The pipeline consists of three main components: (1) data preprocessing and HDF5 dataset construction, (2) PyTorch datasets and data loaders with flexible target transforms and evaluation protocols, and (3) model architectures and experiment management, including both convolutional neural networks (CNNs) and fine-tuning of the LaBraM foundation model.

4.1 EEG Data Processing and Dataset Construction

Input Data Organization

The raw EEG data are organized in a `preprocessed_new` directory with a hierarchical structure at the dataset and participant level. Each release (e.g., `cmi_bids_R1`, `cmi_bids_R2`) contains multiple participants, and each participant folder includes

- a `demographics.csv` file with age and gender, and
- multiple NumPy arrays for EEG trials from active and passive tasks.

Active tasks are stored as files such as `ccd_data_trial_0.npy`, while passive tasks are stored in files such as `sus_data_trial_0.npy`. Each file has shape $(N, 60, T)$: N corresponds to N experimental runs, 60 electrodes, and T time points.

Segment Generation

We convert the trial-level arrays into fixed-length EEG segments suitable for supervised learning. The processing for each trial proceeds as follows:

1. For each of the runs in a trial, we discard the first 40 time points to remove potential start-up artefacts, resulting in $(60, 200)$ arrays.
2. One-second segments are taken directly as $(60, 200)$ "single-run" examples.
3. Four-second segments are created by concatenating four consecutive runs of the same participant and task type along the time dimension, producing $(60, 800)$ arrays.

This procedure is applied separately to active (CCD) and passive (SUS) tasks. Each segment is associated with metadata including the participant identifier, trial and run indices, task type (active or passive), age, and binary gender (0 = female, 1 = male).

HDF5 Storage and Splits

To support efficient streaming and large-scale training we store the processed segments in HDF5 files. For each segment length (e.g., `1s`, `2s`, `4s`) and split we obtain files of the form

`eeg_data_{split}_{segment_length}.h5`

where `split` $\in$ {`train`, `val`, `test`}. Within each HDF5 file we create two top-level groups:

- `active/` for active-task segments, and
- `passive/` for passive-task segments.

Each EEG segment is stored as an individual dataset (with gzip compression) and has a corresponding metadata group storing:

- `participant_id` (string),
- `gender` (integer: 0 or 1),
- `age` (float, in years),
- `task_type` ("active" or "passive"), and
- `task_number` (integer run index).

File-level attributes record the split name, segment length, total number of samples, and number of unique participants. The preprocessing step automatically produces consistent train/validation/ test splits so that all downstream experiments operate on the same partitioning.

Subject-wise splitting and class imbalance. The dataset exhibits a moderate class imbalance in gender labels, with an approximate female-to-male ratio of 2:1 across participants. To mitigate potential bias arising from this imbalance, we employed *stratified sampling* when constructing the training, validation, and test splits, preserving class proportions across all partitions.

Crucially, all dataset splits were performed strictly at the *subject level*. That is, EEG segments originating from a given participant were assigned exclusively to a single split (train, validation, or test), ensuring that no subject appears in more than one partition. This design prevents segment-level data leakage and ensures that reported performance reflects true subject-independent generalization rather than memorization of subject-specific patterns.

PyTorch Dataset and DataLoader

We implement an `EEGDataset` class that subclasses `IterableDataset` to stream data directly from HDF5 without loading the entire file into memory. The dataset constructor accepts parameters that control:

- **Task type filtering**: `task_type` $\in$ {`"active"`, `"passive"`, `"both"`}.
- **Target type**: `target_type` $\in$ {`"gender"`, `"age"`, `"both"`, `"combined"`}.
- **Participant filtering**: an optional list of `participant_id` values to include.
- **Shuffling** and random seed for reproducible iteration.

The dataset yields dictionaries of the form

```
{'eeg_data': (C,T), 'gender': g, 'age': a,
 'participant_ids': id, 'task_types': tt}
```

where $C = 60$ and $T \in \{200, 800\}$ depending on the segment length.

Target Transforms and Tasks

To support multiple prediction tasks on the same underlying dataset we define dedicated target transform functions:

- **Gender classification**: maps raw gender to class 0 (female) or 1 (male).
- **Age classification**: maps chronological age to one of three age groups:

$$0 : \text{age} < 8.5, \quad 1 : 8.5 \leq \text{age} \leq 12.5, \quad 2 : \text{age} > 12.5.$$

- **Age regression**: normalizes age to $[0, 1]$ for continuous prediction.
- **Combined gender+age classification**: maps each (gender, age-group) pair to a single 6-class label.

The transforms can be composed to create datasets for:

- binary gender classification,
- 3-class age-group classification,
- continuous age regression, and
- joint gender+age classification or multi-task outputs.

Cross-Task and Cross-Validation Protocols

The same HDF5 files support multiple evaluation protocols:

- **Standard train/val/test**: train and evaluate on segments from both task types.
- **Cross-task evaluation**: train on one task type (e.g., active) and test on the other (passive), using `task_type` filters.
- **Participant-level filtering**: hold out specific participants or sets of participants to study cross-subject generalization.

These protocols are later combined with the CNN experiment framework to realize baseline, cross-validation, and cross-task experiments.

4.2 Modeling and Experiment Framework

Model Selection Rationale

We select two distinct modeling paradigms to bracket the performance-robustness trade-off. First, we employ a custom **1D-CNN** as a lightweight supervised baseline. Its parameter efficiency makes it stable for training on limited data and well-suited for capturing local temporal features in short one-second windows without the risk of overfitting inherent in larger architectures.

Second, we evaluate the **LaBraM** foundation model to test the hypothesis that large-scale pre-training yields invariant representations. While computationally heavier, LaBraM's

exposure to diverse EEG corpora should theoretically provide robustness against distribution shifts (e.g., training on active tasks and testing on passive states)—a property typically lacking in CNNs trained from scratch.

Convolutional Baselines

As strong supervised baselines we adopt a modular CNN framework for EEG classification. The codebase follows a DRY design with:

- a `BaseEEGCNN` class that implements common convolutional and pooling blocks for EEG,
- task-specific subclasses `EEGGenderCNN` and `EEGAgeCNN`, and
- a model factory that instantiates the appropriate architecture given a configuration.

Both models operate on $(60, T)$ inputs and use several convolutional layers, non-linearities, pooling, and fully connected layers with dropout. The gender model outputs 2 logits (female/male); the age model outputs 3 logits corresponding to the predefined age groups.

Experiment Management

The CNN framework provides a high-level `Experiment` class that orchestrates data loading, training, evaluation, and reporting. Core configuration objects include:

- **DataConfig**: controls batch size, number of workers, cross-validation settings, and whether cross-task evaluation is used.
- **ModelConfig**: specifies the number of channels, dropout rates, and other architectural hyperparameters.
- **TrainingConfig**: defines the number of epochs, learning rate, early-stopping patience, and checkpoint frequency.
- **SystemConfig**: specifies output directories, device selection (CPU/GPU), and verbosity.

The framework supports twelve experiment types, grouped into:

- *Baseline* train/validation/test experiments for gender and age,
- *Cross-validation* experiments with 5-fold splits stratified by gender or age group,

- *Cross-task* experiments that train on active tasks and test on passive tasks (and vice versa), and
- *Cross-task cross-validation* experiments that combine both cross-task and cross-validation protocols.

For each experiment the trainer logs training and validation metrics, saves the best-performing model checkpoint, and writes aggregated results, confusion matrices, and plots to disk.

LaBraM Foundation Model Fine-Tuning

In addition to CNN baselines, we leverage the Large Brain Model (LaBraM), a foundation model pre-trained on $\sim$ 2,500 hours of EEG from around twenty datasets using a vector-quantized neural spectrum prediction objective. LaBraM consists of:

- a neural tokenizer trained to encode short EEG channel patches into discrete neural codes, and
- a transformer encoder trained to reconstruct masked neural codes, analogous to masked language modeling.

To adapt LaBraM to our demographic prediction tasks we follow the original fine-tuning pipeline:

1. Apply the same preprocessing recommended for LaBraM (bandpass filtering between 0.1–75 Hz, 50 Hz notch, resample to 200 Hz, and scaling to μV units) when preparing inputs for the tokenizer.
2. Load a pre-trained LaBraM checkpoint along with the corresponding tokenizer weights and channel order.
3. Replace the original classification head with a new linear head appropriate for the downstream task (2-class gender or 3-class age group).
4. Fine-tune the model on our HDF5-based datasets using task-specific learning rates, layer-wise learning-rate decay, and standard optimization hyperparameters (AdamW with warmup and cosine decay).

Fine-tuning experiments are configured analogously to the CNN experiments, allowing direct comparison between foundation-model-based approaches and lightweight CNN baselines under identical data splits and evaluation protocols.

Chapter 5

Experiments and Results

This chapter presents the experimental methodology and results for demographic prediction using one-second EEG segments. We evaluate three modeling strategies: (i) a baseline convolutional neural network (CNN), (ii) a multi-output CNN with shared representation and two classification heads, and (iii) the LaBraM foundation model fine-tuned on our dataset. All experiments are performed separately for gender prediction (binary), age-group prediction (three classes), and combined prediction (six classes). We additionally assess cross-task generalization by evaluating models trained on active EEG segments and tested on passive segments, and vice versa.

We begin by describing the dataset characteristics using descriptive statistics and histograms. We then present results from CNN models under multiple validation conditions, followed by multi-output CNN experiments, LaBraM experiments, and a final comparative analysis.

5.1 Dataset Description

The dataset includes EEG recordings from 1,587 participants ages 5.0–21.9 (mean 10.5 $\pm$ 3.4). Figure 5.1 illustrates the overall participant age distribution. The distribution is right-skewed, with a larger concentration of participants between 7 and 12 years old, and a gradual decline in participant count among older adolescents.

Ethics, consent, and privacy considerations. The EEG data used in this study was obtained from the Healthy Brain Network (HBN) EEG dataset, a large, publicly released collection of high-density EEG recordings from children and young adults. The original data collection was approved by the Chesapeake Institutional Review Board, with written informed consent obtained from adult participants and consent from legal guardians along

with participant assent for minors. Participants were anonymized using a Global Unique Identifier (GUID) that does not share personally identifiable or protected health information. Given that EEG signals can contain rich biometric information, there is a risk that demographic or sensitive traits could be inferred if raw data are misused. In this work, we use only de-identified data and do not attempt to re-identify subjects or infer attributes beyond those provided, and we note that ethical use of human neuroimaging data requires respect for privacy, consent conditions, and consideration of fairness in downstream models Shirazi et al. (2024).

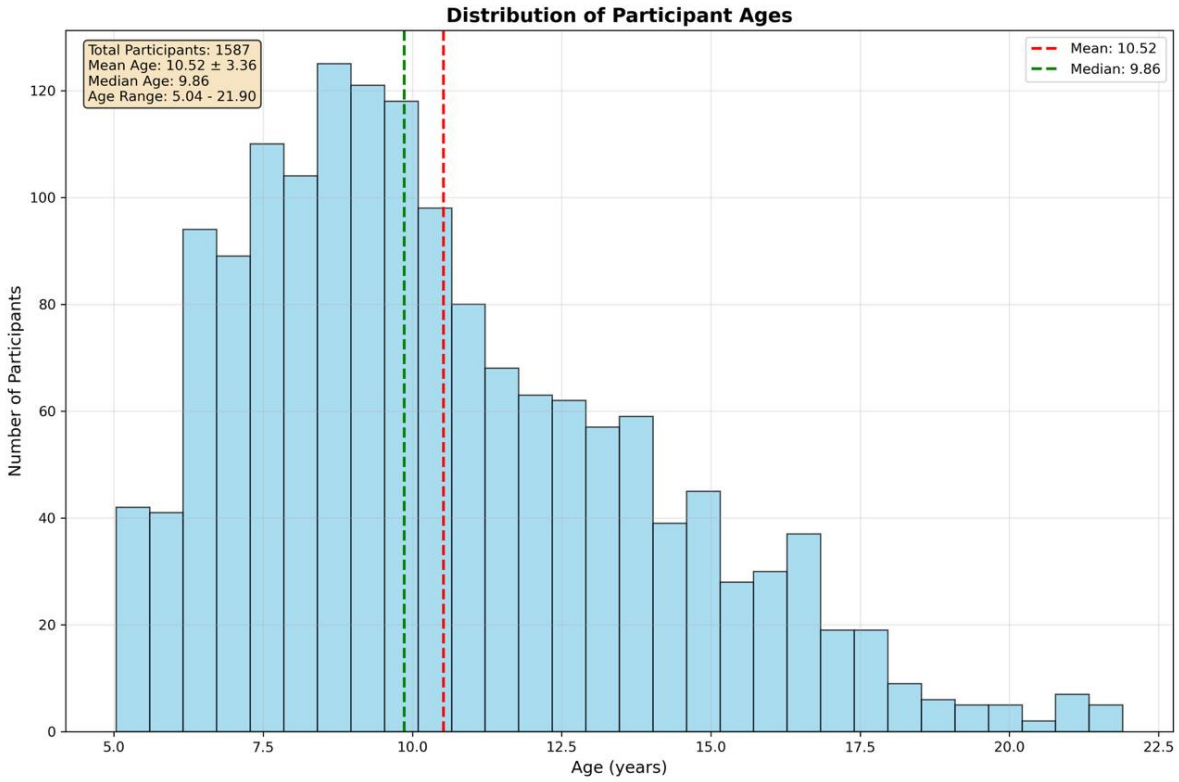

Figure 5.1: Distribution of participant ages. Vertical dashed lines show the mean (red) and median (green).

Figure 5.2 shows the age distributions for males ($n = 563$) and females ($n = 1024$). The dataset is moderately imbalanced in gender, with females representing approximately two-thirds of the sample. Both distributions follow similar shapes, but the female distribution shows a slightly higher density in the 8–12 year range. Because the sample is $\sim$2:1 female, we use class-balanced metrics (and class weights) to avoid inflated accuracy.

As shown in Figure **??**, the dataset exhibits a noticeable gender imbalance, with female participants comprising approximately two-thirds of the cohort. This imbalance motivates the use of stratified evaluation protocols and class-balanced metrics, as discussed in the

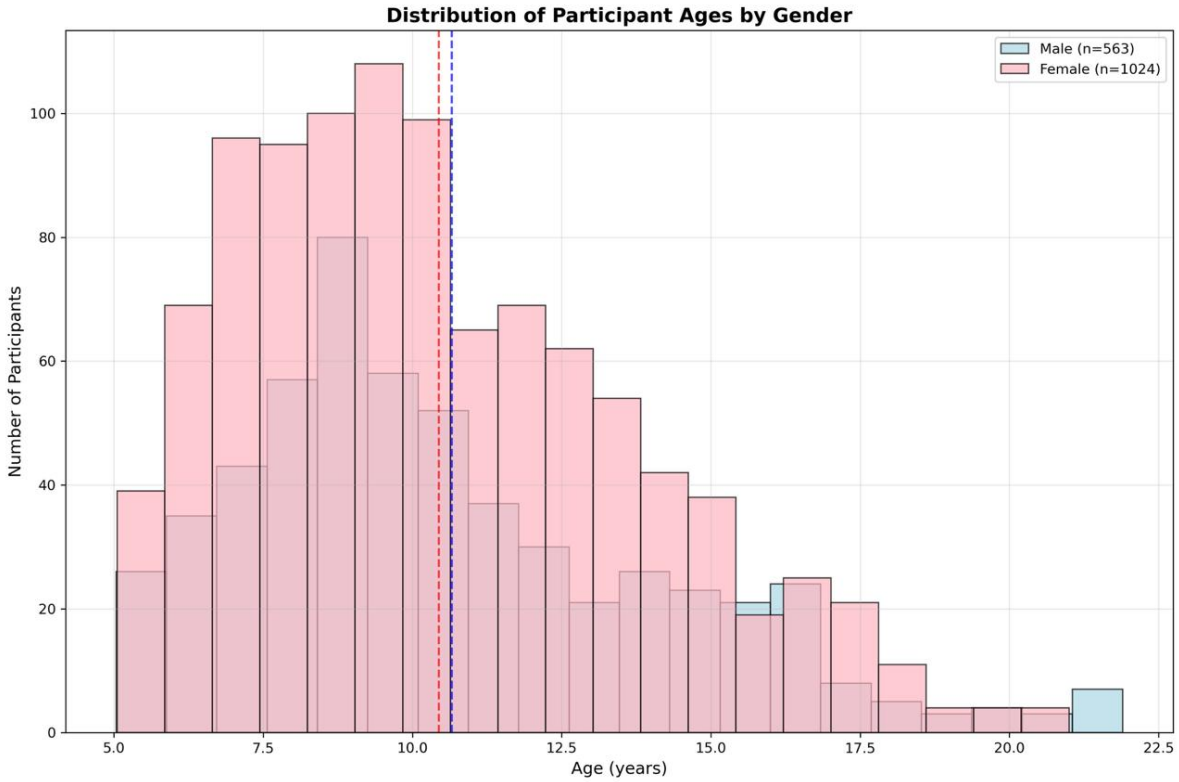

Figure 5.2: Distribution of participant ages by gender.

experimental setup.

5.2 CNN Experiments

We first evaluate a CNN trained on one-second EEG segments under several experimental conditions:

- **Baseline split:** 70% training, 15% validation, 15% testing.
- **Stratified 5-fold CV:** Subject-level stratification by target label.
- **Cross-task generalization:**
  - Train on active segments, test on passive.
  - Train on passive segments, test on active.
- **Prediction tasks:**
  - Gender (binary)

– Age-group (three classes)

– Combined (six classes: gender x age-group)

Figure 5.3 summarizes the CNN accuracies across all experiments. Gender prediction consistently achieves the highest accuracies (up to 0.866 under stratified cross-validation), whereas age-group classification performs substantially worse, with accuracies in the 0.45–0.55 range. Combined classification (six classes) shows the lowest performance, as expected from the increased task difficulty. See Appendix A for confusion matrices illustrating the clustering patterns of age-group misclassification under each experimental condition.

Cross-task generalization results show cross-task accuracy drops by ∼5–10% for gender and ∼10–15% for age relative to within-task experiments. For gender, cross-task accuracy decreases by 5–10%, while for age-group tasks the losses are more severe. These results indicate that the distributional differences between active and passive EEG segments are significant enough to degrade classifier performance.

To further quantify age-group classification performance across evaluation protocols, we compute the macro-F1 score for each condition. Table 5.1 summarizes these macro-F1 results for baseline, cross-task, and cross-validation age-classification experiments using one-second EEG segments.

Table 5.1: Macro-F1 scores for age-group classification across experimental conditions (1s segments).

Experiment	**Macro-F1**
Baseline	**0.4862**
Cross-task (active→passive)	**0.4601**
Cross-task CV (active→passive, age-stratified)	**0.4454**
Cross-task CV (passive→active, age-stratified)	**0.4901**
Cross-task (passive→active)	**0.4915**
Age-stratified CV (within-task)	**0.5165**

These macro-F1 scores complement the accuracy metrics reported in Figure 5.3, offering a balanced assessment of multi-class performance and revealing where systematic misclassification occurs between age groups.

5.3 Multi-Output CNN Experiments

To investigate whether shared representations benefit demographic prediction, we implement a multi-output CNN with two separate classification heads: one for gender and one for age-

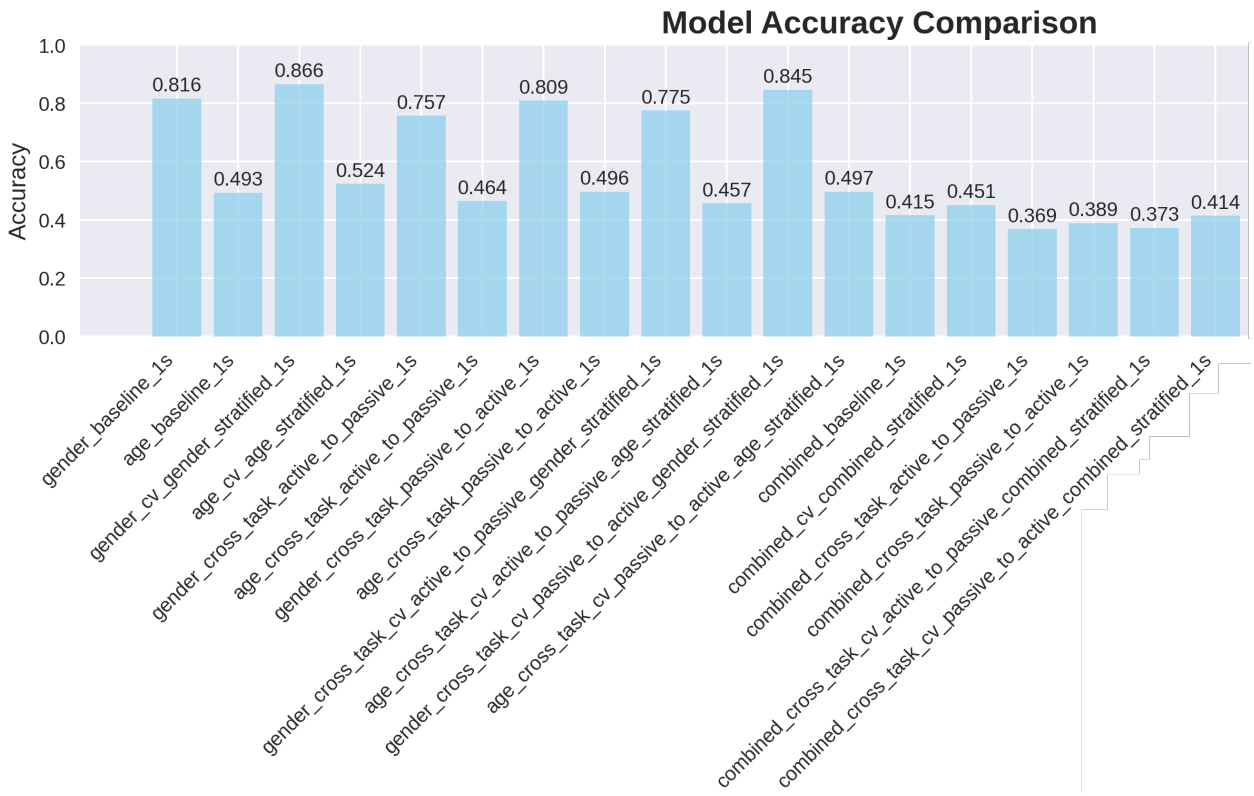

Figure 5.3: CNN accuracy comparison across baseline, cross-validation, and cross-task experiments using one-second EEG segments.

group. The backbone CNN is identical to the single-task model. The model is trained jointly using a weighted sum of the gender and age losses (equal weights).

Figure 5.4 presents the results. Gender prediction remains high (0.768–0.838 across experiments), comparable to or slightly below the single-task CNN. Age-group prediction shows moderate performance (0.459–0.494), again consistent with the difficulty of the task. We observe that the shared representation does not substantially degrade gender performance, but it does not improve age prediction either. This suggests that the age- and gender-relevant EEG features may be partially distinct, limiting the benefits of multi-task learning in this setting.

5.4 LaBraM Experiments

LaBraM is a transformer pretrained on ∼2,500 hours from ∼20 datasets using vector-quantized spectrum prediction and masked-code objectives. We fine-tune LaBraM on our dataset using the same one-second segmentation and evaluate it under the same scenarios as the CNN.

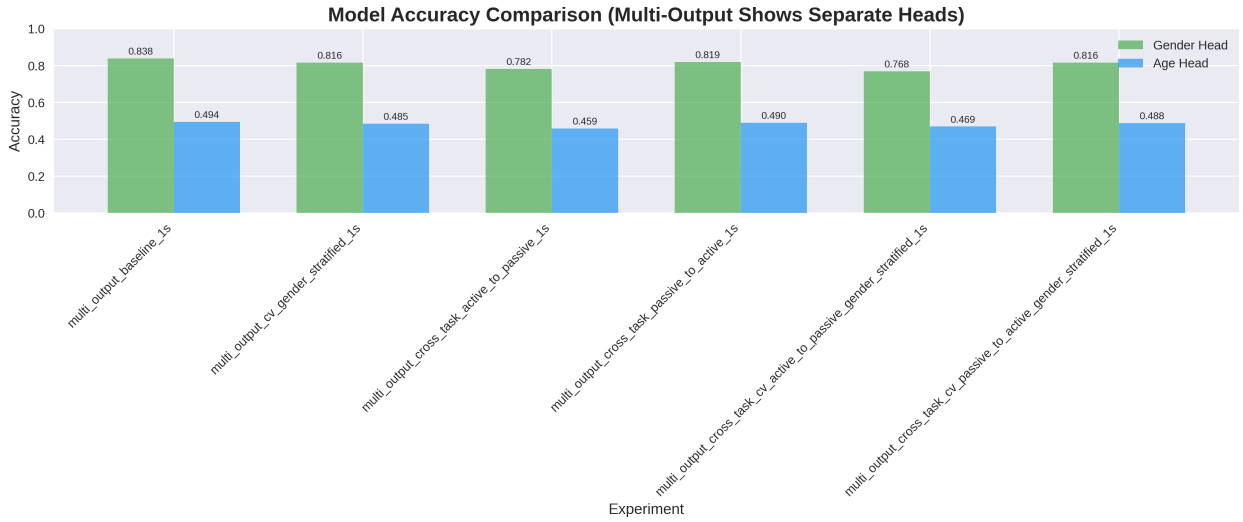

Figure 5.4: Accuracy comparison for multi-output CNN (gender head vs. age head).

Figure 5.5 shows the results. Gender prediction achieves the highest accuracies among all models, with up to 0.866 and 0.863–0.862 in cross-task settings. Notably, LaBraM exhibits smaller performance drops in cross-task evaluation compared with CNN models.

Age-group classification performance remains moderate (0.509–0.555), comparable to CNN models, indicating that large-scale pretraining greatly benefits gender discrimination but does not dramatically enhance age-group prediction on this dataset.

Combined-class prediction remains the most challenging, with LaBraM achieving 0.417–0.459 accuracy, roughly on par with CNN performance.

5.5 Comparative Analysis

Table 5.2 summarizes the performance of all models grouped by task category, **demonstrating that the foundation model (LaBraM) consistently outperforms CNN baselines, particularly in cross-task generalization.** CNN baseline results are taken from the standard 70–15–15 split, whereas stratified CV and cross-task results reflect their respective conditions. LaBraM results correspond to the one-second fine-tuned experiments.

The overall findings can be summarized as follows:

- Gender prediction is consistently easier than age-group or combined prediction.
- Cross-task generalization is challenging for CNNs, with notable drops in accuracy, especially for age-group classification.

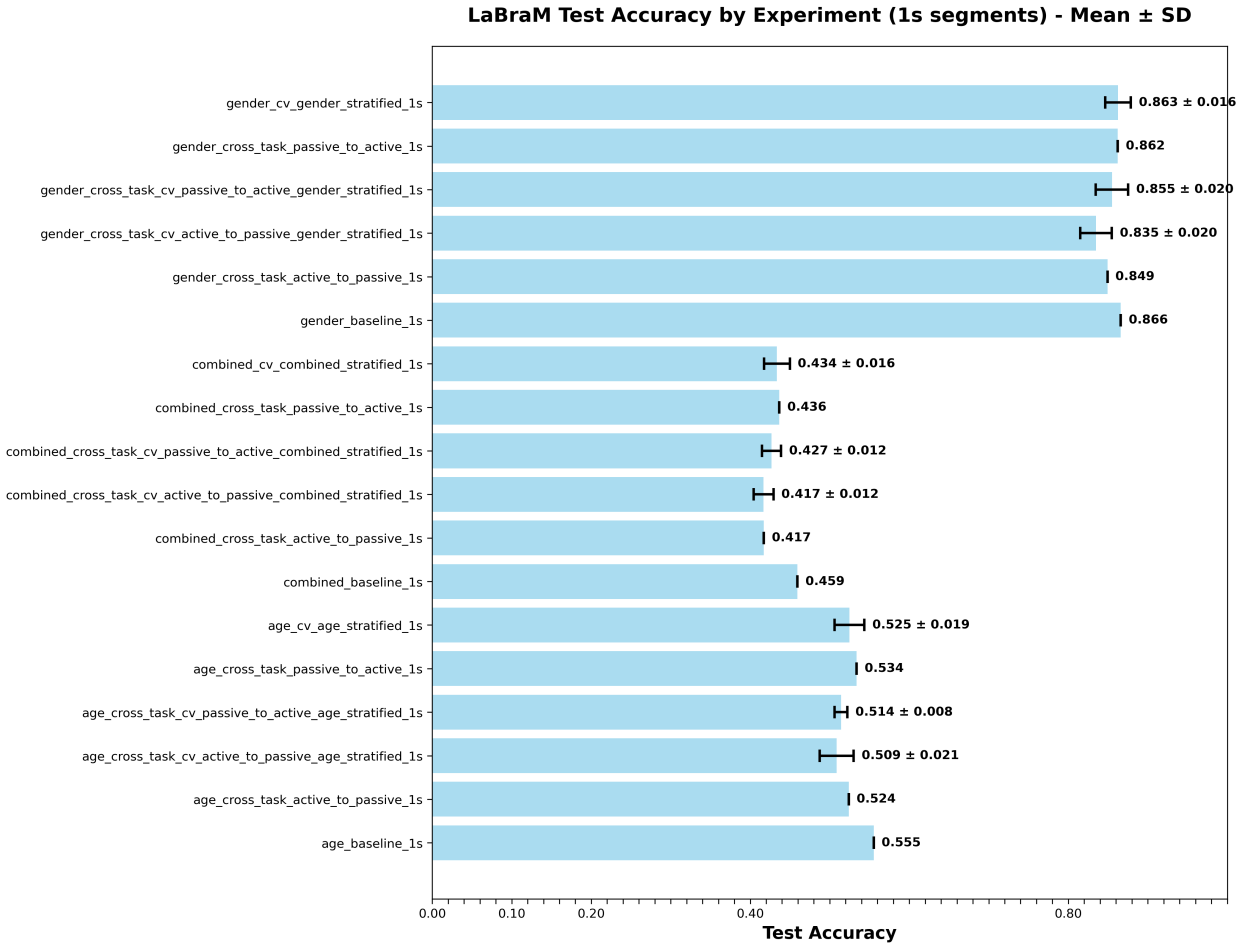

Figure 5.5: LaBraM test accuracy across baseline, cross-validation, and cross-task experiments using one-second EEG segments.

- LaBraM substantially improves gender accuracy and exhibits greater robustness under cross-task conditions.
- Age-group prediction remains difficult for all models, even with large-scale pretraining.
- Multi-output CNNs offer no significant advantage over single-task CNNs for either task.

5.6 Summary

This chapter presented an extensive evaluation of demographic prediction from one-second EEG segments. While CNN models provide a reasonable baseline, their performance is sensitive to task shifts between active and passive segments. Multi-output CNNs do not provide noticeable improvements over single-task variants. The LaBraM foundation model

Table 5.2: Comparison of CNN, Multi-Output CNN, and LaBraM accuracies across task categories (one-second segmentation). A → P: trained on active tasks, tested on passive; P → A: trained on passive tasks, tested on active. Values represent test accuracy.

Task Type	**CNN Baseline**	**CNN (CV)**	**Multi-Output CNN**	**LaBraM**	**LaBraM (CV)**	**Train/Val/Test (segments)**
Gender Baseline	0.816	0.866	0.838	0.866	0.863 ± 0.016	813.5k/174.5k/170k
Gender Cross-Task A → P	0.757	0.775	0.782	0.849	0.835 ± 0.020	252.2k/120.3k/118.3k
Gender Cross-Task P → A	0.809	0.845	0.819	0.862	0.855 ± 0.020	561.3k/54.2k/51.7k
Age-Group Baseline	0.493	0.524	0.494	0.555	0.525 ± 0.019	813.5k/174.5k/170k
Age Cross-Task A → P	0.464	0.457	0.459	0.524	0.509 ± 0.021	252.2k/120.3k/118.3k
Age Cross-Task P → A	0.496	0.497	0.490	0.534	0.514 ± 0.008	561.3k/54.2k/51.7k
Combined Baseline (6-Class)	0.415	0.451	–	0.459	0.434 ± 0.016	813.5k/174.5k/170k
Combined Cross-Task A → P	0.369	0.373	–	0.417	0.417 ± 0.012	252.2k/120.3k/118.3k
Combined Cross-Task P → A	0.389	0.414	–	0.437	0.427 ± 0.012	561.3k/54.2k/51.7k

demonstrates clear advantages in gender prediction and robustness but offers only modest gains in age-group classification. These findings suggest that gender-related EEG features are more readily captured by deep models, whereas age-group differences may require longer temporal context, richer input representations, or additional modalities.

Chapter 6

Conclusion

6.1 Summary of Findings

This thesis investigated the robustness of EEG-based demographic prediction using the Large Brain Model (LaBraM) compared to standard Convolutional Neural Networks (CNNs) on the massive HBN dataset. Our experiments yielded three definitive conclusions regarding the decoding of age and gender from short (one-second) neural signals.

- **Gender is a Stable, "Instantaneous" Signal:** Biological sex was classified with high accuracy (**86.6%** for LaBraM) even from one-second segments. Crucially, this signal proved highly robust to domain shifts; performance remained high ($\sim$86%) even when training on active cognitive tasks and testing on passive resting states. This suggests sex-related spectral features are intrinsic and orthogonal to cognitive state.
- **Age is a Dynamic, Context-Dependent Signal:** In contrast, age-group prediction plateaued at **55.5%** accuracy. While LaBraM outperformed the CNN baseline (52.4%), the overall difficulty indicates that developmental age is likely encoded in longer-scale temporal dynamics or low-frequency rhythms that are lost in one-second snapshots.
- **Foundation Models Enable Robustness:** The fine-tuned LaBraM consistently outperformed CNNs in cross-task generalization. While CNN performance dropped significantly (e.g., to 46.4% for age) when the testing domain changed, LaBraM maintained stability (52.4%). This confirms that large-scale pre-training effectively learns invariant neural representations that resist superficial distribution shifts.

6.2 Theoretical Implications and Limitations

Our findings challenge the reliance on simple supervised learning for neuro-demographics and highlight the specific utility of foundation models.

Limitations and scope. While this work evaluates cross-task generalization between active and passive EEG conditions, all experiments are conducted within a single cohort and dataset. As such, the reported results reflect *within-dataset* transfer rather than true cross-dataset generalization. This distinction is important, as cross-dataset transfer additionally entails robustness to differences in recording hardware, preprocessing pipelines, and population characteristics, which are not explicitly assessed in the present study. We note that LaBraM benefits from large-scale pretraining on multiple EEG datasets; however, in this work it is fine-tuned and evaluated solely on the same active–passive task splits used for all other models.

- **The Generalization Gap:** The failure of CNNs to generalize across tasks underscores the fragility of models trained on limited data from scratch. LaBraM's success validates the "Foundation Model Shift" in EEG, showing that self-supervised pre-training on diverse corpora captures a stable manifold of brain dynamics.
- **Temporal Granularity Constraint:** The primary methodological limitation was the **one-second input window**. This choice, while computationally efficient, likely filtered out critical low-frequency (< 4 Hz) and infra-slow oscillations that characterize neurodevelopment.
- **Biological vs. Chronological Age:** The difficulty in predicting chronological age may reflect the "Brain Age Gap" in the HBN cohort, which is enriched for neurodevelopmental disorders.

6.3 Future Directions

To bridge the gap between current performance and clinical utility, future research must move beyond short-segment classification.

1. **Hierarchical Temporal Modeling:** Integrating LaBraM embeddings into sequence models (e.g., LSTMs or Longformers) to aggregate information over 30–60 seconds would capture the dynamic stability of neural rhythms required for accurate age estimation.

2. **Pediatric-Specific Pre-training:** Pre-training foundation models exclusively on pediatric datasets would better capture the high-amplitude slow waves and unique spectral trajectories of the developing brain.

3. **From Demographics to Diagnostics:** The most promising application is not predicting age itself, but using the *deviation* between predicted and true age (Brain Age Gap) as a biomarker for the psychopathology factors (e.g., ADHD, anxiety) available in the HBN dataset.

Appendix A

Confusion Matrices for Age-Group Classification

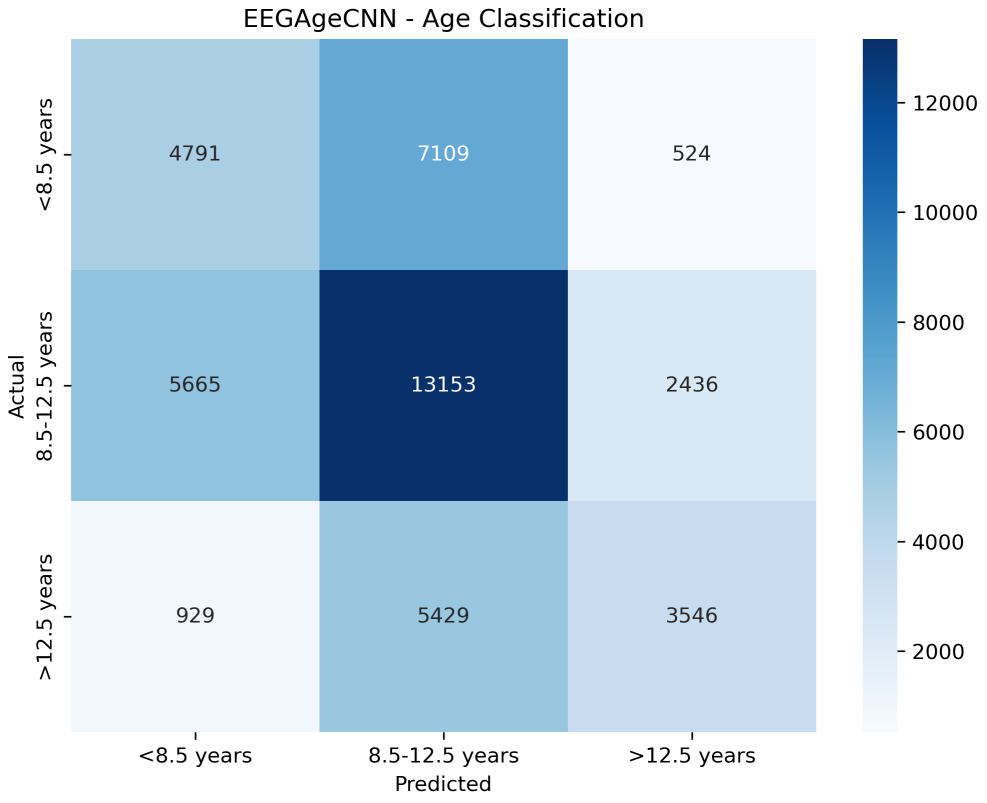

Figure A.1: Confusion matrix for age-group classification — baseline (1s).

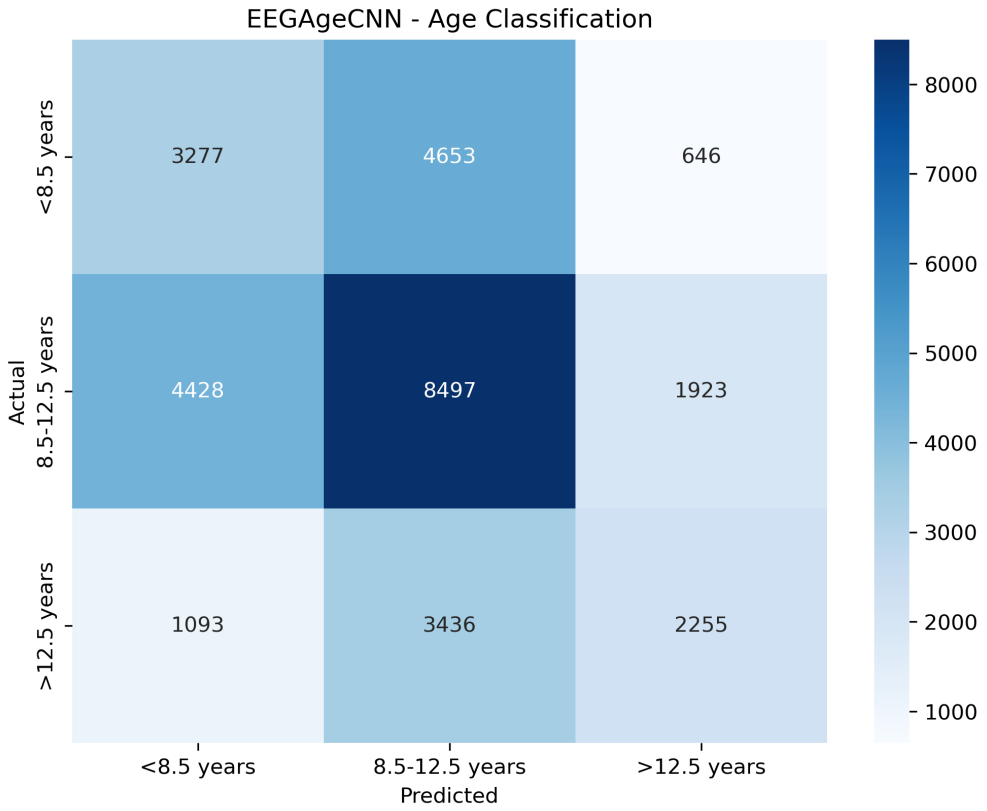

Figure A.2: Confusion matrix for age-group classification — cross-task active→passive (1s).

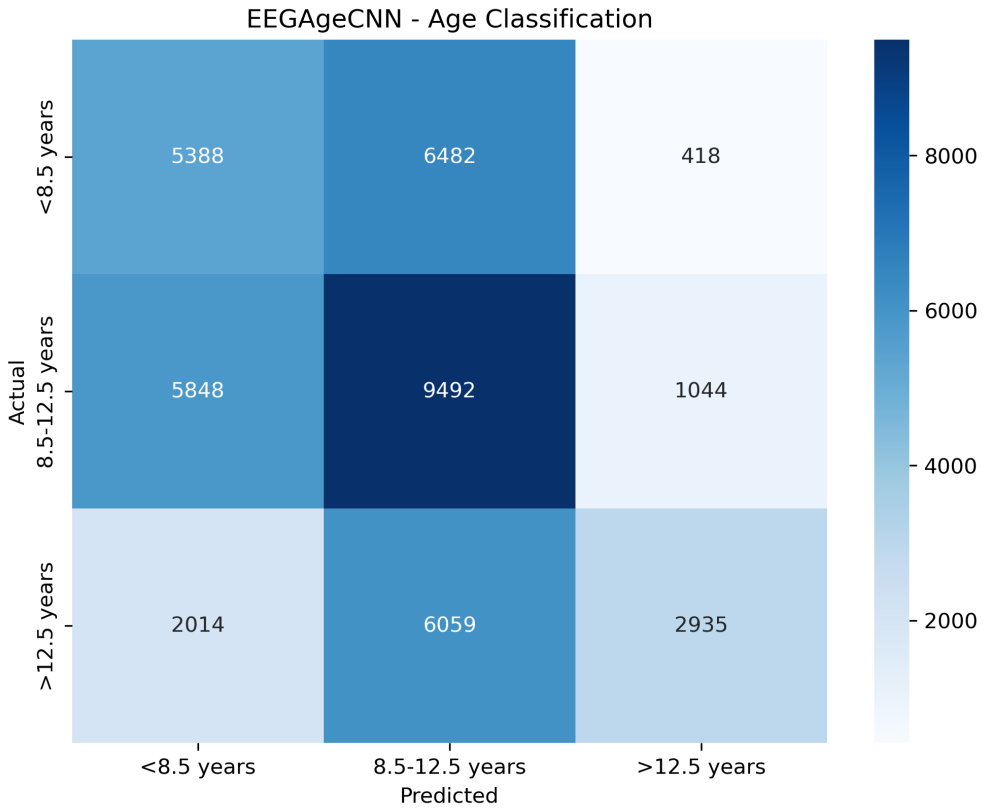

Figure A.3: Confusion matrix — cross-task CV active→passive, age-stratified (1s).

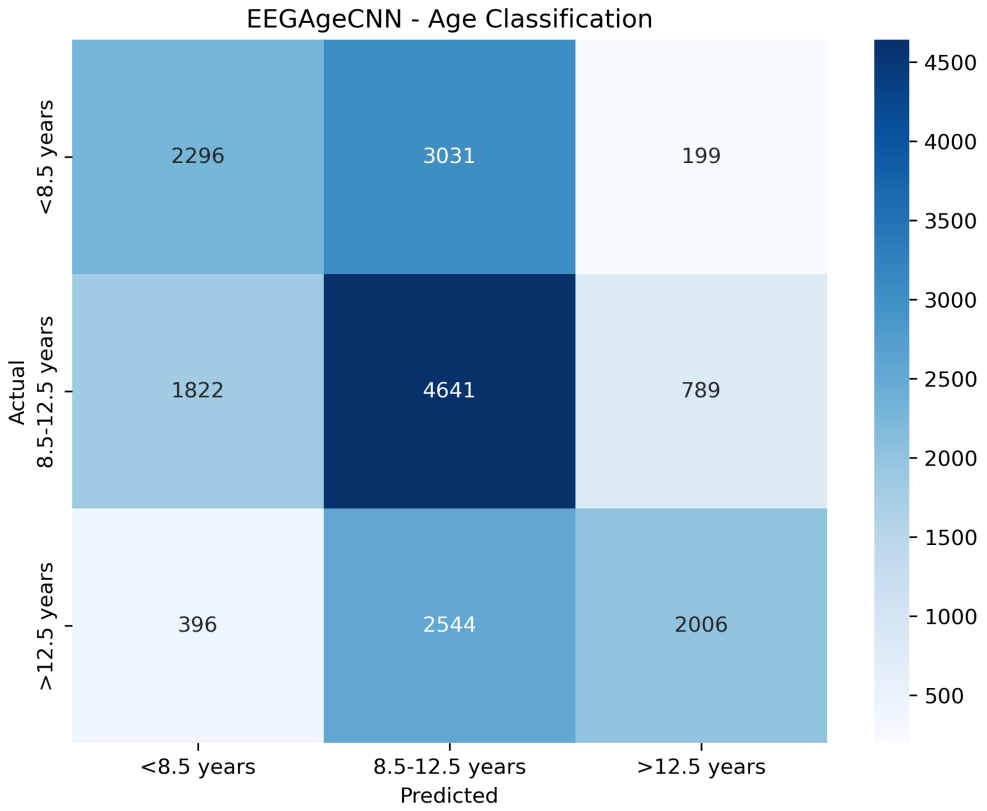

Figure A.4: Confusion matrix — cross-task CV passive→active, age-stratified (1s).

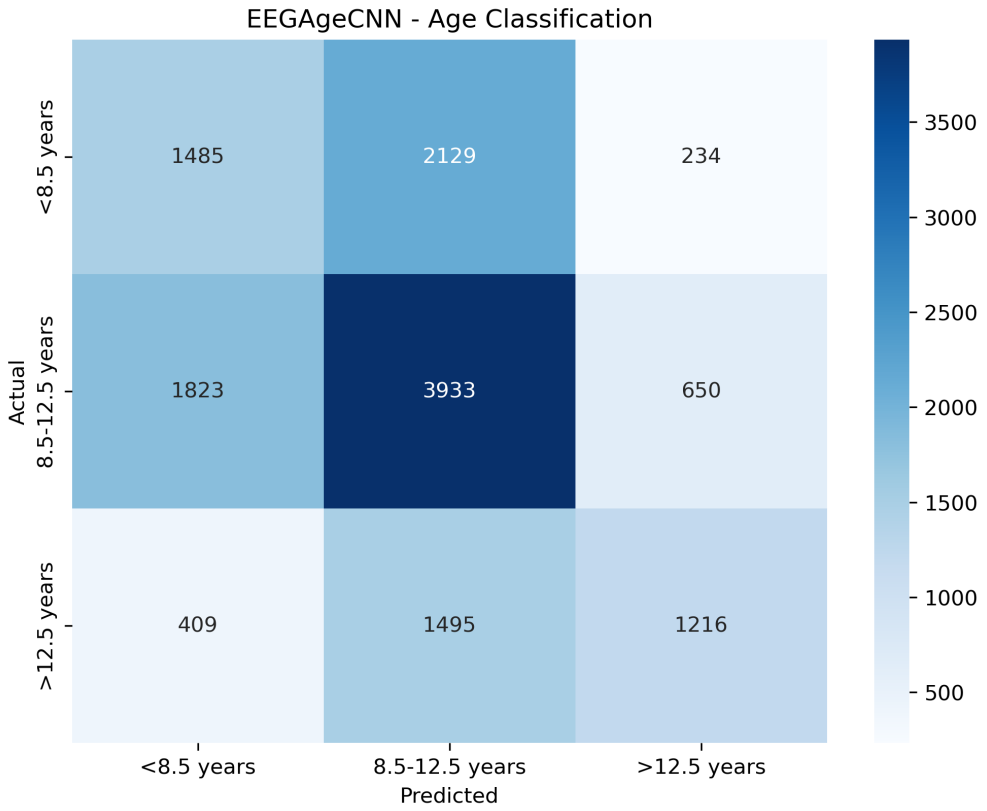

Figure A.5: Confusion matrix — cross-task passive→active (1s).

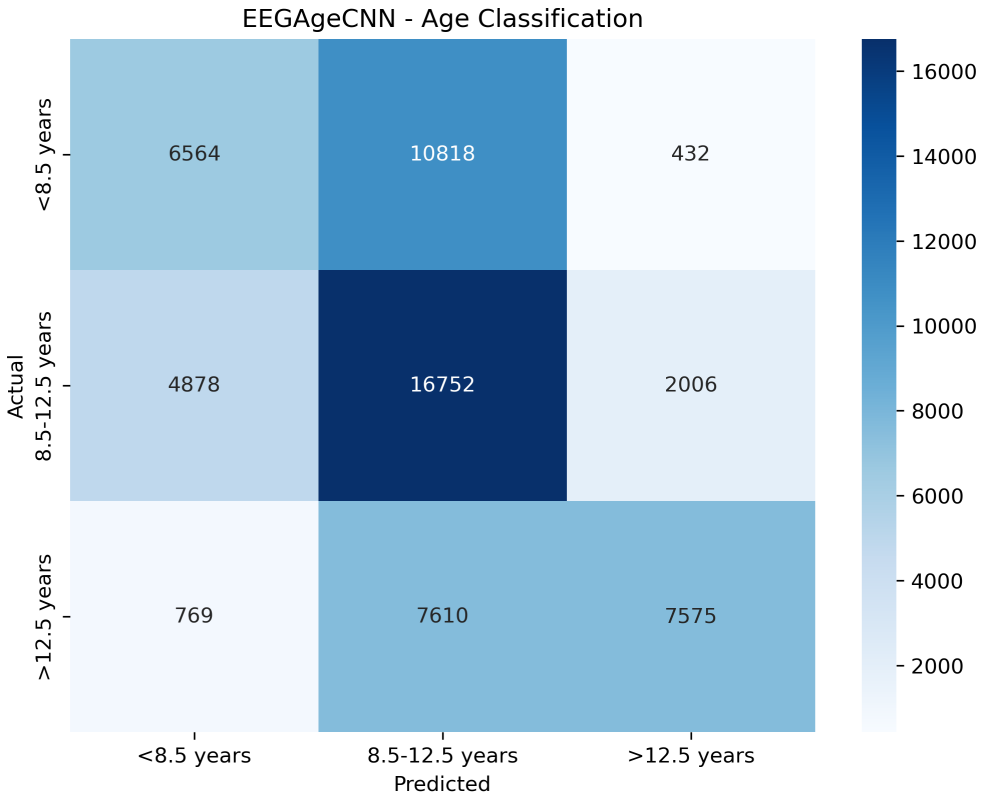

Figure A.6: Confusion matrix — age-stratified CV within-task (1s).

Bibliography

Omar Al Zoubi, Chorong K. Wong, Rayus T. Kuplicki, Hung-Wen Yeh, Atefeh Mayeli, Hazem Refai, et al. Predicting age from brain EEG signals–a machine learning approach. *Frontiers in Aging Neuroscience*, 10:184, 2018. doi: 10.3389/fnagi.2018.00184.

Weiwei An et al. EEG-based brain age prediction in infants–toddlers: Implications for early detection of neurodevelopmental disorders. *Developmental Cognitive Neuroscience*, 71: 101493, 2025. doi: 10.1016/j.dcn.2024.101493.

Saeideh Davoudi, Gabriela Lopez Arango, Florence Deguire, Inga Sophie Knoth, Fanny Thebault-Dagher, Rebecca Reh, Laurel J. Trainor, Janet F. Werker, and Sarah Lippé. Electroencephalography estimates brain age in infants with high precision: Leveraging advanced machine learning in healthcare. *NeuroImage*, 312:121200, 2025. doi: 10.1016/j.neuroimage.2025.121200. URL `https://www.sciencedirect.com/science/article/pii/S1053811925002034`.

Christian Gölz, Eva-Maria Reuter, Stephanie Fröhlich, Julian Rudisch, Ben Godde, Solveig Vieluf, and Claudia Voelcker-Rehage. Classification of age groups and task conditions provides additional evidence for differences in electrophysiological correlates of inhibitory control across the lifespan. *Brain Informatics*, 10(1):11, 2023. doi: 10.1186/s40708-023-00190-y.

Haitham Issa, Qinmu Peng, Sali Issa, Xinge You, Ruijiao Peng, and Jing Wang. Person-independent emotion and gender prediction (EGP) system using EEG signals. *The International Arab Journal of Information Technology*, 19(4):629–637, 2022. doi: 10.34028/iajit/19/4/7.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. *arXiv*, page 2405.18765, 2024. doi: 10.48550/arXiv.2405.18765. preprint.

Yi Ju, Tong Zhao, Zaifen Gao, Wenguang Hu, Jiejian Luo, Nian Cheng, Chunli Liu, Yuwu Jiang, Bo Hong, Taoyun Ji, and Yuxiang Yan. Brain age prediction model based on electroencephalogram signal and its application in children with autism spectrum disorders. *Frontiers in Neurology*, 16:1605291, 2025. doi: 10.3389/fneur.2025.1605291.

Pierre Jusseaume and Ilinka Valova. Brain age prediction/classification through recurrent deep learning with electroencephalogram recordings of seizure subjects. *Sensors*, 22(21): 8112, 2022. doi: 10.3390/s22218112.

Jae-Hwan Kang, Jang-Han Bae, and Young-Ju Jeon. Age-related characteristics of resting-state electroencephalographic signals and the corresponding analytic approaches: A review. *Bioengineering*, 11(5):418, 2024. doi: 10.3390/bioengineering11050418.

Mariam Khayretdinova, Alexey Shovkun, Vladislav Degtyarev, Andrey Kiryasov, Polina Pshonkovskaya, and Ilya Zakharov. Predicting age from resting-state scalp EEG signals with deep convolutional neural networks on TD-BRAIN dataset. *Frontiers in Aging Neuroscience*, 14:1019869, 2022. doi: 10.3389/fnagi.2022.1019869.

Yanxiang Niu, Xin Chen, Yuansen Chen, Zixuan Yao, Xuemei Chen, Ziquan Liu, Xiangyan Meng, Yanqing Liu, Zongya Zhao, and Haojun Fan. A gender recognition method based on EEG microstates. *Computers in Biology and Medicine*, 173:108366, 2024. doi: 10.1016/j.compbiomed.2024.108366.

Seyed Yahya Shirazi, Alexandre Franco, and Mauricio et al. Scopel Hoffman. Hbn-eeg: The fair implementation of the healthy brain network eeg datasets. *bioRxiv*, 2024. doi: 10.1101/2024.10.03.615261.

Trygve Tveitstøl et al. Assessing the robustness of deep learning based brain age prediction models across multiple eeg datasets. *bioRxiv*, page 2025.05.20.655022, 2025. doi: 10.1101/2025.05.20.655022. preprint.

Lan Wei, John C. McHugh, and Catherine Mooney. A machine learning approach for sex and age classification of paediatric EEGs. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–4, 2023. doi: 10.1109/EMBC40787.2023.10341120.

Di Zhang, Yichong She, Jinbo Sun, Yapeng Cui, Xuejuan Yang, Xiao Zeng, and Wei Qin. Brain age estimation from overnight sleep electroencephalography with multi-flow sequence learning. *Nature and Science of Sleep*, 16:879–896, 2024. doi: 10.2147/NSS.S463495.