UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

UNIVERSITÄRE
PSYCHIATRISCHE
DIENSTE BERN

Universitäre Psychiatrische Dienste Bern (UPD),
Department of Old Age Psychiatry and Psychotherapy

Supervisor: Prof. Dr. Björn Rasch

Co-Supervisor: Dr. Marc Züst

Co-Supervisor: Dr. Ahmed Abdulkadir

Brain Age Prediction from Sleep Electrophysiology Using Foundation Model Principles

Master Thesis for the academic degree

Specialized Master of Science in Digital Neuroscience

at the Faculty of Science and Medicine, University of Fribourg

Hannah Portmann (19-211-549)

Bern, 07.07.2025

Abstract

Brain age gap, the difference between predicted biological and chronological age, shows promise as a biomarker for cognitive impairment, including mild cognitive impairment (MCI) and Alzheimer's disease. Given the connection between sleep, aging, and cognition, sleep electrophysiology offers a promising set of modalities for predicting brain age. This thesis aimed to explore whether brain age gap is increased in MCI and reduced after a multi-night phase-locked acoustic stimulation (PLAS) intervention. A brain age prediction model was developed by applying principles of foundation models, using a pretext task to extract features from electrophysiological signals, followed by a downstream age prediction task. Automated sleep staging served as the pretext task, achieving 83% accuracy and an F1 score of 0.82 on a broad dataset, and 49% accuracy and an F1 score of 0.45 on a smaller dataset with a narrower age range. Age prediction was performed using features from the pretext task, achieving Pearson correlations between predicted and chronological age of 0.54-0.78 and 0.03-0.36, which was considered insufficient performance. The results suggest that sleep staging may not provide age-representative features, and existing self-supervised approaches may be more suitable for this purpose. Due to this low predictive accuracy, the hypotheses were not tested.

Statement of Independence

I hereby certify that I have written this Master thesis independently without the help of third parties and without using any sources or aids other than those indicated.

07.07.2025, Hannah Portmann

Acknowledgments

I want to thank my supervisor, Prof. Dr. Björn Rasch from the University of Fribourg, for supervising this external Master thesis and for his support throughout my project.

I am especially grateful to Dr. Marc Züst (UPD Bern) and Dr. Ahmed Abdulkadir (UPD Bern and ZHAW) for co-supervising this thesis and for giving me the opportunity to work on this interesting project. Dr. Marc Züst supported me in understanding the neuroscientific background and the data. Dr. Ahmed Abdulkadir guided me through the development of machine learning and deep learning models and assisted me with all technical issues that came up. They both supported me throughout my thesis and provided assistence with understanding methodologies and interpreting findings.

I would also like to thank Dr. Marc Züst, Dr. Marina Wunderlin, and Dr. Céline Zeller for providing the data used in this thesis. Special thanks go to Dr. Marina Wunderlin for explaining the dataset and its specific variables and for patiently answering all my questions concerning the data.

Additionally, I would like to thank everyone from the group Züst - Dr. Marc Züst, Dr. Marina Wunderlin, Dr. Céline Zeller, Korian Wicki, Sinthujan Somasundaram, and Aaron Friedli - for warmly welcoming me in their team.

I also want to thank the Universitäre Psychiatrische Dienste Bern (UPD) for the opportunity to conduct my Master thesis at their institution and for providing access to their computational resources. My gratitude also goes to the IT support team at UPD - Kaurisanker Kirupananthan, Etienne Müller, and Elia Zimmermann - for their technical assistance.

Finally, I would like to thank everyone who supported me throughout this process, both academically and personally.

Contents

Acronyms

AASM	American Academy of Sleep Medicine
AD	Alzheimer's disease
AUC	area under the curve
CNN	convolutional neural network
EEG	electroencephalography
ELU	exponential linear unit
EMG	electromyography
EOG	electrooculography
MAE	mean absolute error
MCI	mild cognitive impairment
MoCA	Montreal Cognitive Assessment
MRI	magnetic resonance imaging
NREM	non-rapid eye movement
OSA	obstructive sleep apnea
PCA	principal component analysis
PLAS	phase-locked acoustic stimulation
PSG	polysomnography
RBF	radial basis function
REM	rapid eye movement
ROC	receiver operating characteristic
SC	Sleep Cassette
SSL	self-supervised learning
SVC	Support Vector Classification
SVM	Support Vector Machine
SVR	Support Vector Regression
SWA	slow-wave activity
SWS	slow-wave sleep
TANH	hyperbolic tangent
TST	total sleep time
UPD	Universitäre Psychiatrische Dienste Bern
WASO	wake after sleep onset

1 Introduction

With increasing global life expectancy, health conditions linked to aging, such as mild cognitive impairment (MCI) and dementia due to Alzheimer's disease (AD), are becoming more prevalent (World Health Organization, 2015). AD, the most common cause of dementia, is a progressive neurodegenerative disorder marked by severe cognitive impairments, as well as the presence of amyloid plaques and neurofibrillary tangles, whereas MCI is a transitional stage between normal aging and dementia, marked by memory complaints with preserved daily functioning (Kelley & Petersen, 2007). However, aging trajectories vary between individuals and can be influenced by genetic and environmental factors (Rando & Chang, 2012). Thus, chronological age, the time that has passed since birth, is not always a good indicator of age-related health risk and may not reflect biological age, which better captures the stage of age-related physiological decline (Muehlroth & Werkle-Bergner, 2020).

The brain is particularly vulnerable to aging. Age-related structural and functional changes are common and can increase the risk for neurodegenrative diseases (Anderton, 2002; Vanni et al., 2020). Especially disorders associated with cognitive impairment, such as MCI and dementia, pose an increasing challenge as they impair daily functioning, quality of life, and independence in older adults (Andersen et al., 2004; Christiansen et al., 2019; Reppermund et al., 2013) Yet, they remain severely underdiagnosed (Amjad et al., 2018; Bradford et al., 2009). This highlights the importance of reliable and accessible biomarkers of brain health to facilitate early detection and diagnosis of cognitive impairment and dementia. Such biomarkers could help close the diagnostic gap and allow early interventions aimed at slowing disease progression (Baecker et al., 2021). One promising biomarker is the brain age gap, the difference between predicted biological brain age and chronological age (Baecker et al., 2021). This thesis focuses on predicting brain age from sleep electrophysiology.

The following sections provide the theoretical background relevant to this thesis, including an overview of sleep physiology, measurement methods, and disorders, as well as the relationship of sleep with aging and cognition. Machine learning methods such as brain age prediction and MCI classification are introduced and existing research and potential applications are discussed. The concept of foundation models is explained and manual and automated sleep staging is presented. Based on this background, the objectives and hypotheses of this thesis are outlined.

1.1 Sleep and Its Role in Aging and Cognition

Sleep is a naturally recurring state characterized by reduced responsiveness to external stimuli, relative immobility, and a loss of consciousness (Rasch & Born, 2013; Tubbs et al., 2019). It is an important physiological process that occurs in regular intervals, regulated by homeostatic and circadian mechanisms (Hofman & Talamini, 2015; Rasch & Born, 2013). Sleep is essential

for various bodily systems and functions, playing a role in the immune system, cardiovascular health, hormone regulation, memory consolidation, and waste removal from the central nervous system via the glymphatic system (Baranwal et al., 2023). To investigate sleep, it must be quantified and assessed using different measures.

1.1.1 Measuring Sleep

Sleep can be assessed using objective measurements like polysomnography (PSG), which records brain electrical activity via electroencephalography (EEG), eye movements via electrooculography (EOG), and muscle tension of the chin via electromyography (EMG), and may also include additional measures such as respiration, oxygen saturation, or electrocardiography (ECG) (Hofman & Talamini, 2015). These measurements can be used to analyze sleep architecture by quantifying the sleep stages and their progression as well as total sleep time (TST) and the time it takes to fall asleep (sleep onset latency (SOL)) (Tubbs et al., 2019).

In contrast, subjective measures, such as sleep diaries and questionnaires, capture the subjective experience of sleep (Tubbs et al., 2019). Importantly, subjective perceptions of sleep quality do not always align with objective measures. For example, in insomnia, individuals often report poor sleep despite the PSG recording being normal (Carlson & Birkett, 2021; Rosa & Bonnet, 2000).

1.1.2 Sleep Stages

Human sleep alternates between two primary states, non-rapid eye movement (NREM) and rapid eye movement (REM) sleep, in cycles of around 90 minutes (Carlson & Birkett, 2021; Hofman & Talamini, 2015). Each stage is characterized by distinct patterns in the EEG signal along with some physiological differences.

Sleep EEG recordings also include periods of wakefulness, particularly before sleep onset or during short arousals over the course of the night. When a person is awake, the EEG signal shows predominantly alpha (8-12 Hz) and beta (12-30 Hz) activity. Alpha activity, a regular signal with medium frequency, occurs when a person is resting quietly. Beta activity consists of irregular low-amplitude waves and can be observed in active wake states (Carlson & Birkett, 2021; Tubbs et al., 2019).

At the onset of sleep, the brain transitions into NREM sleep, which can further be subdivided into the three stages N1, N2, and N3. Older manuals further divided N3 into two stages (Rechtschaffen & Kales, 1968), but this distinction is not made anymore in the newer American Academy of Sleep Medicine (AASM) manual (Iber, 2007). Across these three NREM stages, sleep becomes progressively deeper with EEG activity decreasing in frequency and increasing in amplitude (Tubbs et al., 2019). N1 marks the transition from wakefulness to sleep, characterized by mixed alpha and theta (4-7 Hz) activity in the EEG signal and slow rolling eye movements, making up for around 5% of total sleep (Hofman & Talamini, 2015; Tubbs et al., 2019). N2,

the most prevalent sleep stage at 45-55% of total sleep, includes periods of theta activity, characteristic sleep spindles and K-complexes, and increasing occurrence of delta waves (1-4 Hz). Sleep spindles are short bursts of waves at 12-15 Hz, whereas K-complexes are sudden large-amplitude waveforms (Carlson & Birkett, 2021; Hofman & Talamini, 2015; Tubbs et al., 2019). N3 sleep or slow-wave sleep (SWS) is the deepest sleep stage, dominated by high-amplitude low-frequency slow-wave activity (SWA) (0.5 - 4 Hz), including delta waves and slow oscillations ($<$ 1 Hz), and making up for 20-25% of total sleep (Baranwal et al., 2023; Rasch & Born, 2013).

These NREM stages are typically followed by REM sleep, forming a full sleep cycle. REM sleep shows a desynchronized low-amplitude EEG signal with occasional theta waves, showing some similarities with N1 sleep or wakefulness. It is characterized by rapid eye movements and muscle atonia and accounts for around 20% of sleep time (Baranwal et al., 2023; Carlson & Birkett, 2021; Hofman & Talamini, 2015). SWS dominates in the first half of the night, while REM sleep increases in the second half (Baranwal et al., 2023; Carlson & Birkett, 2021).

This general structure does not remain static throughout life, and both sleep architecture and EEG features are affected by the aging process.

1.1.3 Sleep and Aging

Sleep and brain aging are intrinsically and bidirectionally linked, meaning sleep affects the aging process, while aging in turn influences sleep patterns. Macrostructural sleep features undergo systematic changes with age. Older adults typically go to bed and wake up earlier, have reduced TST, more fragmented sleep with increased wake after sleep onset (WASO), and a longer SOL. This also results in lower sleep efficiency, the percentage of time spent asleep while in bed. Additionally, deeper N3 sleep and REM sleep decrease, while the lighter stages N1 and N2 increase (Mander et al., 2017; Ohayon et al., 2004). Microstructural features, observable in the sleep EEG signal, also change with age. Older adults show a decline in the density and amplitude of slow waves (Carrier et al., 2011), and sleep spindles during NREM show reduced duration, density, and amplitude (Mander et al., 2017; Purcell et al., 2017). Additionally, the temporal coupling between spindles and slow oscillations becomes disrupted with age (Helfrich et al., 2018). Both macro- and microstructural sleep changes can be detected via sleep EEG recordings, making it a useful modality for the detection of age-related functional brain changes.

Even though these changes are part of normal aging, they can also overlap or interact with pathological conditions. For example, the risk of developing sleep disorders increases with age.

1.1.4 Sleep Disorders

Sleep disorders increase in prevalence with age and often impair both sleep quality and daytime functioning. Around half of individuals over 65 years report some form of sleep complaints

(Cohen et al., 2022). Sleep disorders with increased incidence in aging include insomnia, sleep-disordered breathing, advanced sleep-wake phase disorder, sleep-related movement disorders, and REM sleep behavior disorder (Jaqua et al., 2023; Yaremchuk, 2018).

Insomnia is the most common sleep disorder and increases in prevalence in older adults (Cohen et al., 2022; Yaremchuk, 2018). Patients report difficulties initiating or maintaining sleep, often accompanied by fatigue, mood disturbances, or cognitive difficulties (Baranwal et al., 2023; Jaqua et al., 2023). Comorbid conditions or medications can increase the risk for insomnia in older adults (Jaqua et al., 2023; Yaremchuk, 2018). Diagnosis typically relies on subjective sleep measures, as patients report poor subjective sleep quality while objective sleep measurements often remain unobtrusive (Carlson & Birkett, 2021; Rosa & Bonnet, 2000).

Sleep-disordered breathing, particularly obstructive sleep apnea (OSA), also increases with older age (Russell & Duntley, 2011; Yaremchuk, 2018). OSA involves repeated obstructions of the upper airways during sleep, leading to oxygen desaturation, causing arousals and leading to daytime impairments (Baranwal et al., 2023; Russell & Duntley, 2011; Yaremchuk, 2018).

Advanced sleep-wake phase disorder is a circadian rhythm disturbance which is particularly common in older people, characterized by earlier sleep onset and wake times despite normal sleep quality. A common cause is a reduced exposure to external time cues, known as zeitgebers, that synchronize the circadian rhythm (Cohen et al., 2022).

Sleep-related movement disorders, which include restless leg syndrome and periodic limb movement disorder, also increase in prevalence with older age. Restless leg syndrome involves an urge to move the legs at rest, which often results in difficulties initiating sleep and frequent awakenings. Periodic limb movement disorder is characterized by repetitive limb movements during sleep that can disturb sleep, and often co-occurs with restless leg syndrome (Cohen et al., 2022; Yaremchuk, 2018).

REM sleep behavior disorder is characterized by patients acting out dreams due to loss of muscle atonia during REM sleep. It predominantly appears after age 50 and has been associated with neurodegenerative diseases (Cohen et al., 2022; Yaremchuk, 2018).

1.1.5 The Role of Sleep in Memory and Cognition

Age-related neurodegenerative diseases and cognitive decline are often accompanied by sleep disturbances. These disruptions may result from neurodegenerative processes but also further contribute to their progression (Bah et al., 2019; Mander et al., 2016). Sleep may thus serve as both a biomarker and a modifiable risk factor for cognitive impairment (Bah et al., 2019; Wennberg et al., 2017).

Disrupted sleep patterns frequently precede cognitive decline and worsen along the continuum from subjective cognitive decline to MCI to AD. Common changes include longer SOL, decreased sleep efficiency and TST, increased sleep fragmentation, and decreased SWS and REM sleep (D'Atri et al., 2021; Hu et al., 2017; Petit et al., 2004; Taillard et al., 2019;

Westerberg et al., 2012). Additionally, in the EEG signal, a decrease in spindle density and maximal amplitude, as well as reduced delta, theta, and sigma power during NREM sleep can be observed (Gorgoni et al., 2016; Taillard et al., 2019; Westerberg et al., 2012). Many of these alterations represent exaggerated forms of age-related sleep changes (Petit et al., 2004).

Several sleep features affected in cognitive decline are linked to cognition. Sleep spindles and slow oscillations play a role in sleep-dependent memory consolidation. According to the active system consolidation hypothesis, memory reactivation and transfer to long-term stores occur predominantly during SWS, depending on the coordinated interaction of neocortical slow oscillations, thalamo-cortical spindles, and hippocampal ripples. REM sleep may subsequently support the stabilization of these memories (Rasch & Born, 2013). Spindle loss has been linked to cognitive decline severity (Gorgoni et al., 2016), and impairment in slow oscillation-spindle coupling correlates with the degree of memory consolidation loss (Helfrich et al., 2018). Additionally, both SWS and REM sleep decrease with cognitive impairment (Westerberg et al., 2012), and their decrease has been associated with an increased risk of developing neurodegenerative diseases (Ibrahim et al., 2024).

SWS also enhances waste clearance through the glymphatic system, which removes accumulated waste products, such as amyloid-β, from the brain via cerebrospinal fluid circulation (Voumvourakis et al., 2023). Sleep deprivation, particularly reduced SWS, impairs glymphatic function and can lead to amyloid-β accumulation and an elevated risk for neurodegeneration (Bah et al., 2019; Ju et al., 2014; Voumvourakis et al., 2023). In turn, accumulated amyloid-β may further impair SWA, creating a vicious cycle (Bah et al., 2019).

Various sleep disorders are also linked to cognitive decline. OSA, and insomnia are both prevalent in and associated with an increased risk for MCI and AD, as well as worse cognitive function (Chang et al., 2013; Osorio et al., 2011; Wennberg et al., 2017; Yaffe et al., 2011).

Carpi et al. (2024) identified sleep fragmentation, decreased sleep efficiency, reduced REM sleep, increased light sleep, and sleep-disordered breathing as predictive of cognitive decline. Djonlagic et al. (2021) have also shown that different sleep metrics are predictive of cognitive performance and found that individuals with better cognitive performance tended to have sleep metric profiles commonly associated with younger, healthier individuals. Adra et al. (2023) developed sleep cognitive indices that correlated with specific cognitive scores, based on different sleep features, further underscoring sleep's predictive potential of cognitive decline.

In summary, sleep plays a critical role in cognitive health. While interventions aimed at improving sleep, especially SWS, hold promise for slowing cognitive decline and neurodegenerative processes (Bah et al., 2019; Mander et al., 2016; Wunderlin et al., 2020), objective sleep metrics could aid in early detection of cognitive decline, as sleep changes often precede clinical symptoms (Ju et al., 2014; Taillard et al., 2019; Wennberg et al., 2017). There is growing interest in developing accessible biomarkers that can quantify brain health. One such biomarker is the brain age gap, a measure of biological brain aging that has the potential to detect cognitive decline.

1.2 Brain Age Prediction

Brain age gap has become a highly investigated biomarker for assessing brain health, including the risk of cognitive decline and dementia. Machine learning is used to estimate biological brain age from structural or functional brain data, allowing calculation of the brain age gap, which is obtained by subtracting chronological age from predicted brain age, thereby quantifying deviations from typical brain aging (Baecker et al., 2021; Franke & Gaser, 2019). It represents a non-specific biomarker for detecting pathological brain aging due to a multitude of causes at the individual subject level. A positive brain age gap indicates accelerated aging, while a negative brain age gap suggests delayed aging (Baecker et al., 2021).

Brain age gap has shown promise for early detection, and diagnosis of neurodegenerative diseases (Franke et al., 2010). Additionally, it has the potential to be applied in making treatment decisions or tracking brain age over the course of interventions (Baecker et al., 2021).

Models are usually developed using healthy subjects, as their biological and chronological age should be equal. Model performance is typically assessed through the mean absolute error (MAE), which is the mean of absolute brain age gaps in years, therefore the absolute difference between predictions and labels, in healthy subjects (Baecker et al., 2021). The closer this is to 0, the more accurate the model usually is. However, this can depend highly on the age distribution of the test set. Additionally, Pearson's correlation coefficient (r) is commonly used to quantify the correlation between chronological and predicted age, with higher values indicating better model performance (Franke & Gaser, 2019).

Early brain age models relied on magnetic resonance imaging (MRI) scans, achieving MAEs below 5 years and Pearson correlations up to 0.97 (Franke & Gaser, 2019; Franke et al., 2010). Elevated MRI-based brain age gaps have been associated with a multitude of disorders, including MCI (Franke et al., 2012), AD (Franke et al., 2010, 2012), schizophrenia (Shahab et al., 2019), traumatic brain injury (Cole et al., 2015), human immunodeficiency virus (HIV) (Cole et al., 2017b), and type 2 diabetes mellitus (Franke et al., 2013). Additionally, they were also correlated with worse cognitive functioning and more severe clinical symptoms in subjects with cognitive impairment, MCI, or AD (Franke et al., 2012).

Recent advances increasingly used deep learning approaches to learn features directly from raw data, reducing the need for manual preprocessing and the bias of manual feature selection. However, these deep models require large and diverse datasets to be trained successfully (Cole et al., 2017a).

EEG-based brain age prediction offers a cost-effective, participant-friendly alternative that could capture functional brain changes and facilitate repeated or at-home measurements (Sun et al., 2019; D. Zhang et al., 2024). Resting-state EEG (rs-EEG) models have achieved MAEs around 6 years using both traditional machine learning and deep learning methods (Al Zoubi et al., 2018; Khayretdinova et al., 2022).

1.2.1 Sleep-EEG based Brain Age Prediction

Sleep EEG recordings provide a more stable and natural state than wake EEG, often with longer recordings. They could further allow to study the intrinsic effects of aging on brain activity (D. Zhang et al., 2024). Several studies have demonstrated their promise for brain age prediction. Sun et al. (2019) used manually extracted stage-averaged features in a regression model, achieving a MAE of 7.6 years and a correlation of 0.83. They also demonstrated an increased brain age gap in subjects with neurological and psychiatric diseases, hypertension, and diabetes. Ye et al. (2020) used the same prediction model and showed an increase in brain age gap from nondementia to MCI and further to dementia, thereby demonstrating the potential of brain age gap as a biomarker for progressing cognitive impairment. Brink-Kjaer et al. (2022) applied a convolutional neural network (CNN) and long-short term memory (LSTM) model on PSG data, reporting an MAE of 5.8 years and a correlation of 0.84. They suggest an association of a higher brain age gap with higher mortality, hypertension, type 2 diabetes mellitus, and heart-related comorbidities. Yook et al. (2022) used scalograms and sleep stage information in a CNN with dense connectivity, resulting in an MAE of 4.8 years and a correlation of 0.86 in cross-validation. A higher brain age gap was associated with cortical thinning, OSA, and insomnia. D. Zhang et al. (2024) used a multi-flow sequence-learning model on short sleep EEG signal segments and sleep stage hypodensity inputs, reporting a MAE of 4.19 years and a correlation of 0.97. They further demonstrated a slightly increased brain age gap in subjects with psychiatric and neurological disorders. Banville et al. (2024) estimated brain age from at-home EEG devices using vectorized filterbank covariance matrices with a Ridge Regression model, achieving an MAE of 7.57 years using sleep recordings.

Given the natural and age-related signals of sleep EEG recordings and the association of sleep electophysiology changes with cognitive decline, it represents a valuable modality for brain age prediction.

1.2.2 Brain Age Gap Application in PLAS Intervention

Brain age prediction holds promise for diagnosis, prognosis, and treatment decisions for age-related disorders, but also for tracking the effects of interventions that could potentially reduce brain age (Baecker et al., 2021). One potential intervention for cognitive decline is phase-locked acoustic stimulation (PLAS), which enhances SWS by presenting short acoustic stimuli time-locked to the positive peaks of slow oscillations, thereby increasing SWA (Ngo et al., 2013; Wunderlin et al., 2023).

Some improvements in overnight consolidation of episodic memory after a PLAS intervention have been suggested in younger adults (Wunderlin et al., 2021), while group-level effects in older adults remain limited. Nevertheless, multi-night PLAS interventions have been associated with improvements in memory and amyloid-β levels, indicating potential benefits (Wunderlin et al., 2023). Delayed effects have also been shown in adults with cognitive impairment (Zeller

et al., 2024). Investigating whether PLAS can modulate brain age gap may be of interest to demonstrate its preventive potential for cognitive decline and accelerated brain aging.

1.3 MCI Classification

Beyond age prediction, sleep EEG features could also be used to distinguish subjects with MCI from healthy controls, given the established association between cognitive impairment and sleep alterations (Taillard et al., 2019; Wennberg et al., 2017; Westerberg et al., 2012). While multiple studies focused on wake EEG recordings, only few have used sleep EEG data for MCI classification. Ye et al. (2023) used manually extracted sleep EEG features with different classical machine learning methods (logistic regression, Support Vector Machine (SVM), random forest) to classify healthy subjects from MCI and/or dementia. They achieved an area under the receiver operating characteristic curve (AUROC) of 0.73 for classifying MCI from healthy individuals. Geng et al. (2022) used slow wave and spindle features with SVM and gated recurrent unit (GRU) classifiers, resulting in a classification accuracy of 93.46%, an AUROC of 0.98, and a F1 score of 93.56%. The AUROC (from here on referred to as simply area under the curve (AUC)) reflects the ability of the model to distinguish between classes across different thresholds, and the F1 score is the harmonic mean of precision and recall (Fawcett, 2006). Haghayegh et al. (2025) focused on preclinical detection, classifying subjects that later developed cognitive impairment from subjects remaining cognitively healthy using a random forest classifier with an AUC of 0.76. These results highlight the potential of sleep EEG as a valuable modality for early detection of cognitive impairment.

1.4 Foundation Models and Pretext Tasks

With limited resources and data availability, directly predicting brain age from raw electrophysiological signals might be challenging. Although the model by Brink-Kjaer et al. (2022) has successfully predicted brain age directly from sleep PSG signals as input, it required a large amount of data for effective training. A promising alternative is the application of foundation model principles. Foundation models represent a new paradigm in artificial intelligence (AI), where models are typically trained on a pretext task using large, diverse datasets and self-supervised learning (SSL) to learn general-purpose representations which can then be adapted to a range of downstream tasks (Bommasani et al., 2022).

SSL pretext tasks are used to learn meaningful representations from the data, typically in the absence of large labeled training sets, reducing reliance on manual feature engineering (Albelwi, 2022; Rani et al., 2023). To facilitate a specific downstream task, the foundation model can be fine-tuned, or learned features can be extracted from it to be used in a secondary task (Rani et al., 2023). However, selecting an appropriate pretext task for a specific downstream task is crucial to ensure they work well together (Albelwi, 2022).

While foundation models have predominantly been used in natural language processing (e.g., ChatGPT) (Zhou et al., 2024), they have the potential to be useful for all kinds of modalities (Bommasani et al., 2022). Recent efforts started exploring foundation models for sleep data using self-supervised pretext tasks to learn general features from sleep measurements (C.-H. Lee et al., 2025; Ogg & Coon, 2024; Thapa et al., 2025). Some even tested brain age prediction as a downstream task, reporting MAEs of 7.3 to 15 years (Ogg & Coon, 2024; Thapa et al., 2025). Most also used sleep stage classification as a downstream task, achieving high accuracies (C.-H. Lee et al., 2025; Ogg & Coon, 2024; Thapa et al., 2025).

Automated sleep staging is a well-established task in sleep research that relies on the sleep EEG signal, which is known to change with age (Carrier et al., 2011; Helfrich et al., 2018; Mander et al., 2017; Purcell et al., 2017). Sleep staging may thus serve as a useful pretext task for brain age prediction. The next section provides background on both manual and automated sleep staging.

1.5 Manual and Automated Sleep Staging

Sleep is commonly scored manually into five stages following AASM guidelines, using at least three EEG channels, EOG, and EMG. However, manual scoring is time-consuming and subject to interrater variability, with agreement rates around 82% and Cohen's Kappa between 0.61 and 0.76 (Danker-Hopfe et al., 2009; Y. J. Lee et al., 2022). Cohen's Kappa is used to quantify agreement between predicted and true labels, accounting for agreement by change (Malafeev et al., 2018). N1 scoring consistently shows the lowest agreement, likely because the transition from Wake to N1 can be difficult to recognize, especially in subjects with little alpha activity, and it shows similarities to N2 in the absence of spindles or K complexes (Danker-Hopfe et al., 2009; Y. J. Lee et al., 2022; Silber et al., 2007). N3 agreement can also be reduced by the dependence on slow wave amplitude, which declines with age (Y. J. Lee et al., 2022; Muehlroth & Werkle-Bergner, 2020). Discretizing a continuous physiological process further complicates the detection of boundaries, especially within NREM sleep (Fiorillo et al., 2019; Muehlroth & Werkle-Bergner, 2020).

To reduce subjectivity and time expenditure, automated sleep staging methods have been developed. Early approaches used manually extracted features with classical machine learning models, while more recent methods apply deep learning directly to raw signals (Fiorillo et al., 2019; Sun et al., 2023). State-of-the-art models achieve up to 90% accuracy, Cohen's Kappa values above 0.80, and F1 scores of 85-90% (Gaiduk et al., 2023). Class imbalance presents a challenge which can lead to biased predictions that favor more represented classes, although balanced sampling helps mitigate this problem (Fiorillo et al., 2019). As in manual scoring, N1 is the most difficult stage to classify (Gunnarsdottir et al., 2020; Haghayegh et al., 2023; Malafeev et al., 2018), and confusion between N2 and N3 is also common (Gunnarsdottir et al., 2020; Stephansen et al., 2018; X. Zhang et al., 2020). These errors may again reflect the

artificial definition of stage boundaries, for instance, an epoch with 18% of SWA is scored as N2, while one with 20% is labeled N3 (Stephansen et al., 2018).

Generalizability across datasets is another challenge due to biological (e.g., sex, ethnicity) and technical (e.g., recording equipment) differences. Transfer learning has been proposed to adapt pretrained models to new domains (Ganglberger et al., 2024).

U-Sleep (Perslev et al., 2021) is a popular example of a generalizable model, as it performs well on different clinical cohorts and input channels without any additional retraining. It is a fully convolutional feedforward network trained on 15660 participants from 16 studies, achieving a global F1 score of 0.79 on an independent test set.

Nevertheless, supervised models inherit the limitations of manual scorings they learn from. They are tied to 30-second epoch scorings, which may not align with physiological transitions, and potentially inconsistent human labels (Fiorillo et al., 2019). Training on labels from a single rater further risks overfitting to that scorer's style, which is why consensus scorings from multiple raters should be preferred (Sun et al., 2023).

A potential improvement are hypnodensity graphs, which are probabilistic outputs representing the likelihood of each sleep stage per epoch. This representation could offer more information about sleep trends and may capture transitional epochs more accurately (Stephansen et al., 2018; Sun et al., 2023).

Unsupervised approaches offer another direction by detecting clusters of sleep states without predefined stage labels. Such pseudo-stages have shown similarities with manually scored hypnograms and could even be used for scoring (Agarwal & Gotman, 2001; Ferjani et al., 2020; Gath & Geva, 1989).

1.6 Objectives and Hypotheses

This thesis aimed to simulate the development of a foundation model for sleep electrophysiology data by applying a pretext task to learn generalizable features, which were then used for the downstream task of brain age prediction. This task is just one of many potential applications for a general sleep electrophysiology foundation model. Others could include sex prediction or the prediction of disorders, such as MCI or dementia.

Although chronological age labels are usually available for sleep EEG data and the problem of lacking labeling does not apply to brain age prediction, a pretext task may still be beneficial, as it could help extract biologically meaningful and relevant features from the data. Many models for automated sleep staging with strong performance already exist, and the high availability of sleep stage labels is an advantage. Because sleep architecture changes with age (Ohayon et al., 2004) and sleep staging relies on microstructural features, which also change with age (Carrier et al., 2011; Helfrich et al., 2018; Mander et al., 2017; Purcell et al., 2017), to detect the stages, sleep staging is likely to capture age-related characteristics in the EEG signal, making it a promising pretext task for brain age prediction.

An architecture derived from U-Sleep (Perslev et al., 2021) was chosen for the pretext task. It consists of an encoder module to extract abstract features, a following decoder module to upscale those features to the temporal resolution of the input, and a segment classifier that aggregates the high-frequency output to segments (Perslev et al., 2021). The bottleneck layer between the encoder and decoder has the lowest temporal resolution and is thus expected to contain the most general representations for transfer to downstream tasks. While existing work has explored SSL approaches for foundation models in sleep research (C.-H. Lee et al., 2025; Ogg & Coon, 2024; Thapa et al., 2025), the use of supervised sleep staging as a pretext task has, to my knowledge, not yet been examined. Its potential to learn general features makes it an interesting alternative to more complex self-supervised approaches.

In summary, the goal was to develop a brain age prediction model that imitates the principles of foundation models by utilizing automated sleep staging as a pretext task and leveraging features extracted from this task to implement the downstream task of age prediction. The model was intended to be evaluated on healthy adults and further on MCI patients and subjects who underwent PLAS intervention to compute their brain age gap before and after the intervention. The following hypotheses had been formulated:

1. MCI patients have a higher brain age gap compared to healthy subjects.

2. Brain age gap is reduced after a multi-night PLAS intervention.

However, the developed brain age model did not achieve sufficient accuracy to test the hypotheses reliably. Therefore, the thesis focuses on methodology, model evaluation, and the exploration of alternative approaches, including an MCI classification model. Potential improvements and future directions are discussed.

2 Methods

All analyses were performed on a server running Ubuntu 18.04 with an x86_64 CPU (48 cores) and three NVIDIA Quadro P5000 GPUs. The software environment included Python 3.6.9, TensorFlow 2.6.2, CUDA 11.2, and cuDNN 8. Primarily open-source libraries and tools were used, listed in Appendix A. To ensure reproducibility and consistet data splits and sampling, all scripts included a fixed random seed set to 42.

2.1 Data

This thesis used two datasets: the publicly available Sleep Cassette (SC) data and a locally collected dataset from the Universitäre Psychiatrische Dienste Bern (UPD).

2.1.1 SC Datasetset

The SC data, a subset of the Sleep-EDF database (Goldberger et al., 2000; Kemp et al., 2000; Mourtazaev et al., 1995), was used to develop and evaluate the pretext task and downstream models. The dataset consists of 153 PSG recordings from 78 healthy participants. Two nights were recorded per participant, except for three participants only one night was available. The subjects were between 25 and 101 years old (M $\pm$ SD: 58.79 $\pm$ 22.15 years), and included 37 men (47. 4%) and 41 women (52. 6%). Each recording includes two EEG channels (Fpz-Cz and Pz-Oz), one horizontal EOG channel, and a submental chin EMG channel, all sampled at a frequency of 100 Hz. Additional channels, such as respiration and temperature, were not used. Sleep stage annotations were available at 30-second epochs following the Rechtschaffen and Kales (1968) criteria, scoring the epochs with the labels Wake, REM sleep, S1, S2, S3, S4, M (movement time), and ? (not scored).

2.1.2 UPD Dataset

The UPD dataset was collected across three studies investigating the effect of acoustic stimulation on memory performance in older adults. Subset 19-01 includes data from 33 subjects, recruited initially as healthy subjects by Wunderlin et al. (2023). Subset 21-03 was collected by Zeller et al. (2024) and includes data from 15 subjects, classified initially as cognitively impaired. Subset 21-04 includes data from 16 subjects from the first data sample (19-01) who were retested one to two years later for the study by Wunderlin et al. (2024).

The combined dataset consisted of 64 data samples from 48 subjects, with each sample contributing one baseline night. Recordings from the adaptation night, during which participants adjust to sleeping in the lab, and experimental nights were not used to ensure consistency and avoid stimulation-related effects. Additionally, the participants' ages and Montreal Cognitive

Assessment (MoCA) scores (Nasreddine et al., 2005) were available. Cognitive status was reassessed for this thesis based on the MoCA scores (MCI: MoCA $\leq$ 25) (Nasreddine et al., 2005).

Summary statistics for the three sub-datasets and the cognitive groups in the final dataset are shown in Table 2.1 and Table 2.2. In the final UPD dataset, subjects who were remeasured in 21-04 are treated as distinct individuals, and their data is not combined with their recordings from 19-03, as their age has changed and they were reassessed using the MoCA.

Table 2.1: Summary statistics of the UPD dataset and its sub-datasets.

Data	Age M $\pm$ SD (years)	Age min–max (years)	Male / Female	MCI / Cognitively Healthy
19-01	69.18 $\pm$ 4.52	59 – 80	8 / 25	5 / 28
21-03	71.80 $\pm$ 5.14	63 – 79	10 / 5	11 / 4
21-04	70.06 $\pm$ 3.60	60 – 76	4 / 12	1 / 15
UPD	70.02 $\pm$ 4.52	59 – 80	22 / 42	17 / 47

Table 2.2: Summary statistics for cognitively healthy and MCI groups in the UPD dataset.

Group	Age M $\pm$ SD (years)	Age min–max (years)	Male / Female	MoCA M $\pm$ SD
Healthy	69.35 $\pm$ 4.58	59 – 80	12 / 35	28.00 $\pm$ 1.38
MCI	71.88 $\pm$ 3.87	63 – 78	7 / 10	23.24 $\pm$ 1.52

Recordings were obtained using a 128-channel EEG net referenced to Cz and sampled at 500 Hz. The data had already been preprocessed, including downsampling to 200 Hz, preprocessing using the PREP pipeline (Bigdely-Shamlo et al., 2015), and manual scoring following AASM guidelines (Iber, 2007) (Wunderlin et al., 2024; Zeller et al., 2024).

2.2 Data processing

To ensure comparability between datasets, UPD recordings were downsampled to 100 Hz, matching the frequency of the SC dataset. From the 128 available EEG channels in the UPD data, four were selected to approximate those in the SC data, as described in Table 2.3. No EMG channel was available for the UPD dataset, therefore, it was approximated as closely as possible by using an EEG channel.

Table 2.3: Channel Mapping between SC and UPD Datasets.

SC channels	UPD channels
Fpz-Cz	E15
Pz-Oz	E62 - E75
Horizontal EOG	E126 - E127
Submental Chin EMG	E120

After this, both datasets were processed similarly, as detailed below. Only segments relevant to sleep were retained by trimming all data that occurred more than 30 minutes before sleep onset and after awakening. Sleep stage labels were converted to numerical labels (0-4). In the SC dataset, stages S3 and S4 were combined into N3 for compatibility with the AASM scoring (Iber, 2007). The labels movement and not scored were masked, as they were not intended to be predicted.

To handle memory and loading time issues, a randomly chosen continuous segment of a fixed duration was extracted from each sample and saved. The segment length was defined by the shortest usable sleep duration across samples in each dataset, which was 5.5 hours for SC and 3 hours for UPD.

2.3 Correlational Analyses

To assess whether the macrostructural features in our data were predictive of or correlated with age, TST, as well as the absolute and relative durations of each sleep stage, were computed per participant and correlated with age for both datasets separately. Spearman's rank correlation coefficient (ρ) was used due to non-normal distributions, as indicated by the Shapiro-Wilk test ($p < 0.05$). P-values were computed for significance, and the significance threshold was adjusted for multiple testing using the Bonferroni correction ($\alpha = \frac{0.05}{11} \approx 0.0045$). Scatterplots with fitted linear regression lines were also generated.

For the SC dataset, each participant contributed a single data point by averaging the metrics across both nights where available, as these recordings occurred on consecutive nights. In contrast, for the UPD dataset, participants retested in subset 21-04 were treated as separate observations due to the time gap of one to two years and, thus, a change in age.

2.4 Pretext Task - Automated Sleep Staging

2.4.1 Data Preparation

Sleep staging was performed using only healthy subjects. Thus, the entire SC dataset and the healthy group only from the UPD dataset were used. Both datasets were split into training, validation, and test sets (70:15:15 ratio). For subjects with multiple nights of recordings, all recordings were placed in the same set, prioritizing a subject-wise split over an exact split ratio.

The training sets were cached and shuffled. From each sample, a 17.5-minute segment was selected at each training epoch. This sampling was implemented in the TensorFlow pipeline to ensure that different segments of each recording were encountered over the training process. To ensure exposure to rare sleep stages during training, each segment was centered around a timepoint sampled from a randomly selected sleep stage, based on the approach in Perslev et al. (2021). Sample weights of 0.1 were applied one second before and after each sleep stage transition to account for potential label imprecisions at boundaries. All other time points were weighted as 1. The labels were one-hot encoded, and the signal dimensions were realigned to ensure a uniform shape ([timepoints, channels]). A normalizer was fitted to the training sets to have a mean of zero and a standard deviation of one, and then applied to the training data. The training sets were then divided into batches of four samples.

For the validation sets, a completely random 17.5-minute segment was selected per sample per epoch, independent of sleep stage labels, to evaluate predictions of realistic sleep stage proportions. For testing, the full nights of sleep, with a maximum of 30 minutes of Wake before and after sleep, were used. These testing sets were preprocessed newly from the original data following the same steps as described in Section 2.2, except for the cutting to 5.5 or 3 hours. No sample weights were applied to the validation and test sets. All other preprocessing steps, including label encoding, normalization using the normalizer adapted to the training set, and batching, were also applied to the validation and test sets.

2.4.2 Network Architecture

The implemented network was heavily inspired by the architecture of U-Sleep (Perslev et al., 2021), a fully convolutional feed-forward deep neural network. The network consisted of 12 encoders, each including a convolutional layer, an exponential linear unit (ELU) activation, a batch normalization layer, and a max pooling layer. The intermediate representation between batch normalization and max pooling was saved as a residual connection. A bottleneck layer connecting the encoder and decoder included an additional convolutional layer and a batch normalization layer. To reconstruct the full resolution, the architecture also entailed 12 decoders. Each decoder comprised an up-sampling layer, a convolutional layer, an ELU activation, a batch normalization layer, and a residual connection from the corresponding encoder. A second

convolutional layer with ELU activation and another batch normalization layer followed. Finally, the model concluded with a classification head comprising three convolutional layers. The first included a hyperbolic tangent (TANH) activation, the second an ELU activation, and the last layer employed linear activation to output logits. The class with the highest values was then predicted. Unlike the original U-Sleep, this model outputs class logits at the input frequency of 100 Hz rather than at 30-second or otherwise defined epochs.

The architecture retained the kernel sizes and filter numbers from the original U-Sleep but was adapted to handle 4 channels and 17.5-minute segments sampled at 100 Hz as input. Full details of the model architecture are provided in Appendix B.

2.4.3 Training

The model was trained using the Adam optimizer (learning rate = 1×10^{-4}) and categorical cross-entropy loss for 500 epochs. Training progress was monitored via loss and accuracy on the training and validation sets. A separate 10000 epoch run was conducted to check whether the model could be improved by further training. This and additional substudies conducted during development and training are described in Appendix C.

2.4.4 Evaluation

Model performance was evaluated using categorical accuracy, cross-entropy loss, per-class and overall (class-averaged) F1 scores, and Cohen's Kappa (computed over 30-second aggregated predictions). Absolute and row-normalized confusion matrices were produced. To assess class discriminability, receiver operating characteristic (ROC) curves and corresponding AUC (also called AUROC) scores were computed. This included both one-vs-rest (each class against all others) and one-vs-one (each pair of classes) settings, using normalized probabilities for one-vs-one. ROC curves depict the trade-off between true positive and false positive rates across decision thresholds in binary classification. The AUC quantifies the overall discriminative ability, with a value of 1 indicating perfect discrimination and 0.5 representing random guessing (Fawcett, 2006).

2.4.5 Transfer Learning and Retraining

The model was first developed and trained using the SC dataset. Afterward, this pre-trained model was directly applied to the healthy UPD dataset, following the same evaluation procedure, except that Cohen's Kappa was not used as scoring epoch lengths were not consistent. As it was expected that the trained model would not generalize well to a completely different dataset, the model was also retrained from scratch on the UPD data with the same preprocessing, training, and evaluation procedure as described above.

2.5 Feature Extraction

To perform the downstream task of age regression, features were extracted from the bottleneck layer of the pretext task model, as this was expected to capture the most generalizable representations. To enable feature extraction during inference, a second model output was added to return the bottleneck features only when the model is not in training mode. The model thus produced two outputs: the final sleep stage predictions and the features from the bottleneck layer. A schema linking the pretext and downstream tasks by feature extraction is shown in Figure 2.1.

For feature extraction, the entire saved recordings (5.5 hours for SC, 3 hours for UPD) from all dataset splits were passed through the model. The same training, validation, and test splits as used during sleep stage training were retained. This resulted in 306 features with representations in the shape of [484, 306] for SC and [264, 306] for UPD, representing a temporal downsampling of factor 2^{12}. Slight deviations from this factor are due to zero-padding of signals with uneven lengths.

The extracted features were saved alongside the corresponding age, and for the UPD dataset also the corresponding MoCA scores.

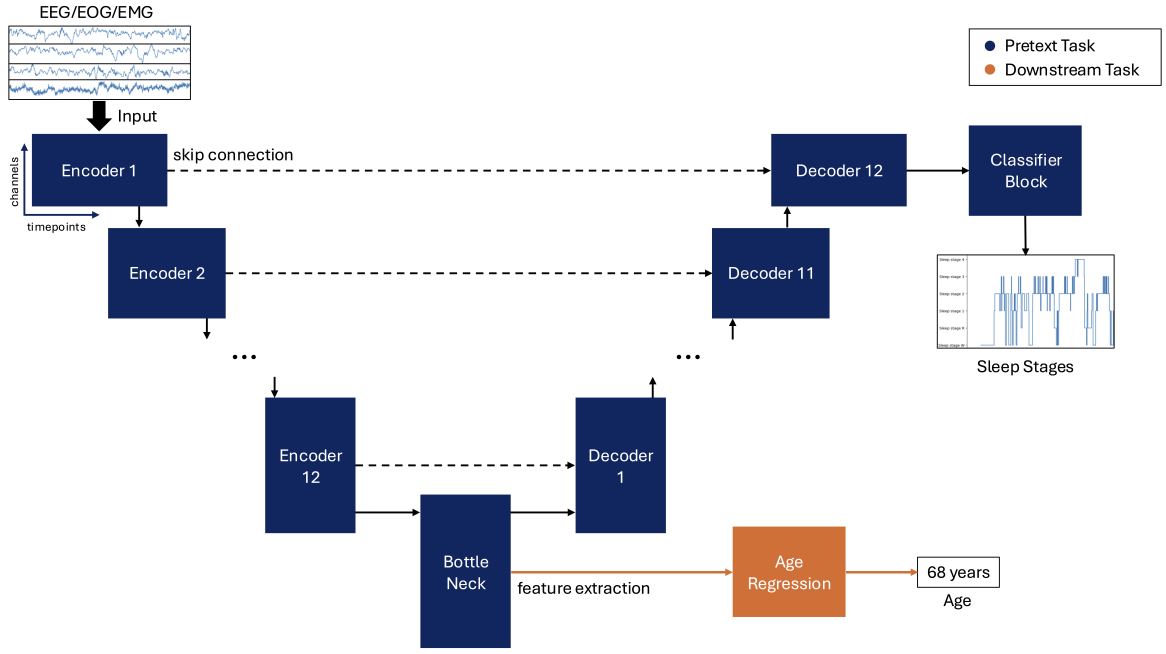

Figure 2.1: Complete model architecture illustrating the connection between the pretext task of sleep staging (blue) and the downstream task of age regression (orange) via feature extraction. The model adopts a U-shaped structure, with 12 encoder and decoder blocks dedicated to the pretext task of automated sleep stage classification. Skip connections between corresponding encoder and decoder blocks preserve temporal context. The bottleneck layer serves as a shared feature extractor whose representation is passed to the downstream task module for age regression. Solid arrows indicate the flow of data, while dashed arrows represent skip connections.

2.6 Downstream Tasks

2.6.1 Brain Age Prediction

The extracted features were used as input for traditional regression methods to predict age. Five different input representations were tested:

(1) **Flattened Features:** All features reshaped into a single vector.

(2) **Averaged Features:** Each of the 306 features averaged across time.

(3) **PCA on flattened features:** First 10 principal components after applying principal component analysis (PCA) for dimensionality reduction on flattened features.

(4) **PCA on averaged features:** First 10 principal components after applying PCA for dimensionality reduction on averaged features.

(5) **Top correlated features:** The 10 averaged features most strongly correlated with age.

Spearman's rank correlation was used for input five due to non-normality of most features (262/306 features, Shapiro-Wilk $p < 0.05$).

Three traditional regression methods were evaluated using their scikit-learn implementations:

(A) **Linear Regression**: Fits a linear function by minimizing mean squared error between predicted and true target values (Bishop, 2006).

(B) **Ridge Regression:** Adds L2 regularization to linear regression to reduce overfitting (Hoerl & Kennard, 1970).

(C) **Support Vector Regression (SVR):** Aims to find a function within a margin of tolerance around the actual targets while minimizing the prediction error. Four different kernels (linear, polynomial, radial basis function (RBF), and sigmoid) were applied to capture non-linear relationships by mapping the data into higher-dimensional feature spaces (Drucker et al., 1997).

Default hyperparameters were used except for the SVR kernels.

Model performance was evaluated using MAE and the correlation between the predicted and chronological ages. For this, Pearson's r was chosen as the primary metric due to its use in existing work and its suitability for linear relationships. However, as age was not normally distributed according to the Shapiro-Wilk test ($p < 0.05$), Spearman's rank correlation coefficient (ρ) was computed as an alternative correlation measure. Additionally, scatterplots and residual plots were generated for visual analysis.

In the SC dataset, models were trained on the training set. For each input type, the model with the lowest MAE and the one with the highest Pearson's r on the validation set were selected and subsequently evaluated on the test set, with those results being reported.

In the UPD dataset, due to the small sample size, leave-one-out cross-validation was applied. PCA and correlations for feature selection were applied newly for each split. Per input, the models with the lowest MAE and the highest Pearson's r across leave-one-out cross-validation were reported.

2.6.2 MCI Classification

MCI classification was performed on the UPD dataset only, as the SC dataset contained only healthy subjects. Both healthy and MCI subjects were now included, and MCI labels were used as a binary classification target. Again, leave-one-out cross-validation was used due to the limited sample size.

The same input types as in the age prediction task were used, except for the feature selection based on age correlations (input representation 5).

Five traditional classification algorithms were evaluated using their scikit-learn implementations:

(A) **Logistic Regression:** Linear model where class probabilities are estimated using the logistic function (Bishop, 2006).

(B) **Support Vector Classification (SVC):** Maximizes the margin between classes to find a decision boundary. Linear and RBF kernels were tested to enable linear and non-linear decision boundaries by mapping data into higher-dimensional spaces (Cortes & Vapnik, 1995).

(C) **Linear Discriminant Analysis (LDA):** Models the distribution of each class using Gaussian distributions and uses a linear combination of features that best separates the classes by maximizing between-class variance relative to within-class variance (Fisher, 1936).

(D) **Gaussian Naïve Bayes:** Probabilistic classifier based on Bayes' theorem modeling the likelihood of features using Gaussian distributions (Maron, 1961).

(E) **Random Forest:** Ensemble of decision trees returning the class predicted by majority voting (Breiman, 2001).

Default hyperparameters were used except for the SVC kernels, and due to the imbalanced class labels, depending on the method, either the class weights were set to balanced or the priors were computed and provided to the model.

Evaluation metrics included accuracy, AUC, and F1 score. Confusion matrices and ROC curves were also generated. For each input type, the model with the highest F1 score and the one with the highest AUC across leave-one-out cross-validation were reported.

3 Results

3.1 Correlational Analyses

SC Dataset

Table 3.1 shows Spearman's rank correlation coefficients (ρ) and p-values for correlations between sleep stage durations and percentages with age in the SC dataset (n = 78). Significant correlations (Bonferroni-adjusted $\alpha = 0.0045$) were found for N1, N3, and REM stages. N1 duration and percentage showed strong positive correlations with age. N3 duration and percentage exhibited moderate negative correlations, while REM duration and percentage showed weaker negative correlations with age. TST, Wake, and N2 measures showed weak, non-significant correlations.

Table 3.1: Spearman's correlation coefficient (ρ) and p-values for associations between sleep stage durations (minutes) and relative percentages (%) with age in the SC dataset (n = 78). An asterisk (*) indicates significance according to the Bonferroni-adjusted significance level ($\alpha = 0.0045$).

Feature	Spearman's ρ	p-value
Total sleep time	0.14	0.22
Wake duration	0.084	0.46
Wake %	0.033	0.77
N1 duration	0.65	1.73×10^{-10} *
N1 %	0.62	1.42×10^{-9} *
N2 duration	-0.055	0.63
N2 %	-0.11	0.33
N3 duration	-0.41	1.63×10^{-4} *
N3 %	-0.42	1.17×10^{-4} *
REM duration	-0.33	3.39×10^{-3} *
REM %	-0.39	4.71×10^{-4} *

UPD Dataset

As shown in Table 3.2, all correlations of sleep stage durations or percentages with age in the UPD dataset (n = 64) were weak to very weak, and none reached significance according to the Bonferroni-corrected significance level.

Table 3.2: Spearman's correlation coefficient (ρ) and p-values for associations between sleep stage durations (minutes) and relative percentages (%) with age in the UPD dataset (n = 64). No significant correlations were found according to the Bonferroni-adjusted significance level (α = 0.0045).

Feature	Spearman's ρ	p-value
Total sleep time	-0.14	0.26
Wake duration	0.061	0.63
Wake %	0.10	0.44
N1 duration	0.12	0.34
N1 %	0.22	0.078
N2 duration	-0.048	0.71
N2 %	0.038	0.76
N3 duration	-0.27	0.030
N3 %	-0.26	0.037
REM duration	-0.14	0.27
REM %	-0.095	0.45

Scatterplots depicting these relationships can be found in Appendix D.

3.2 Pretext Task

Training performance over 500 epochs on the training and validation sets is illustrated in Appendix E.

SC dataset

On the SC test set (n = 25), the model achieved a loss of 0.53, an accuracy of 0.83, and an overall F1 score of 0.82. Class-wise F1 scores were 0.92 for Wake, 0.84 for REM, 0.41 for N1, 0.84 for N2, and 0.80 for N3. Agreement of the predictions aggregated over 30 seconds with the manual scoring, regarded as true labels, resulted in a Cohen's Kappa of 0.77. The confusion matrices (Figure 3.1) show that Wake and REM stages were classified accurately. N2 and N3 also showed high accuracy, although N3 was occasionally misclassified as N2 and N2 as N1. N1 showed the weakest performance, correctly predicted in only 39% of cases, and often misclassified as Wake or N2.

The ROC curves and the corresponding one-vs-rest AUC values are presented in Figure 3.2. The lowest one-vs-rest AUC was 0.85 (N1 vs. Rest), with all others exceeding this. The one-vs-one AUC values (Table 3.3) were all above 0.90.

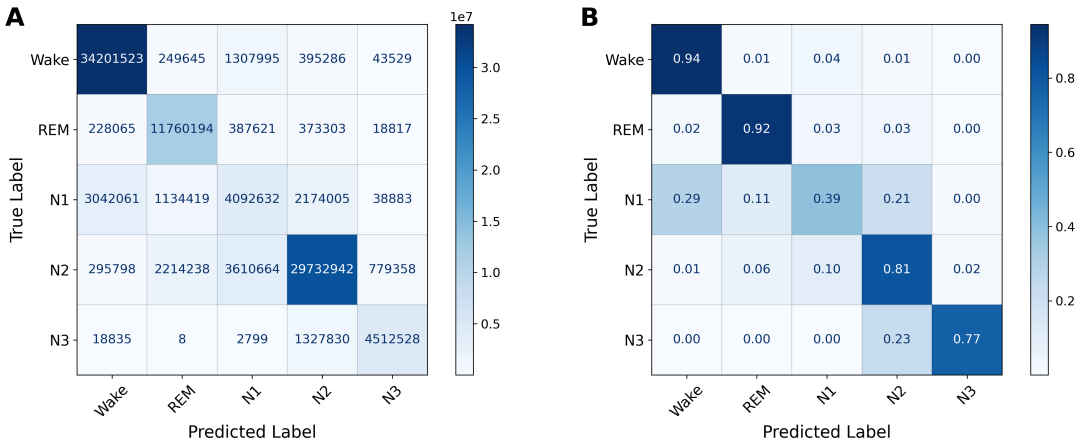

Figure 3.1: Confusion matrices from the SC test set. (A) Absolute counts of true versus predicted labels. (B) Row-normalized confusion matrix showing the percentage distribution of predicted classes for each true label.

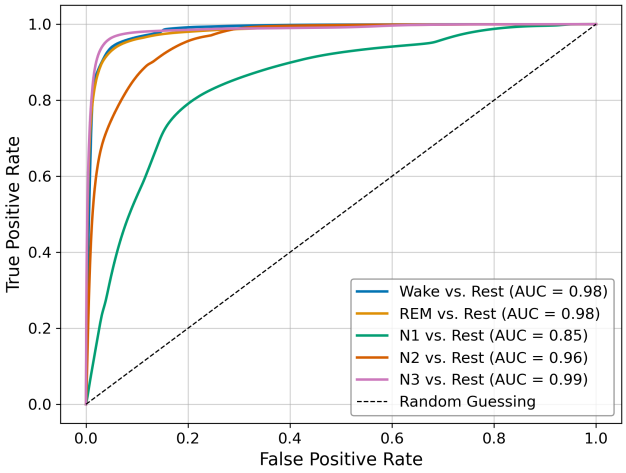

Figure 3.2: One-vs-rest ROC curves and corresponding AUC values for all sleep stages in the SC test set.

Table 3.3: One-vs-one AUC scores for all sleep stage pairs in the SC test set. Each cell reports the AUC for distinguishing the row class from the column class based on normalized one-vs-one probabilities summing to 1.

	Wake	REM	N1	N2	N3
Wake	–	1.00	0.94	1.00	1.00
REM		–	0.96	0.98	1.00
N1			–	0.91	0.99
N2				–	0.97
N3					–

UPD Dataset

Applying the SC pre-trained model directly to the healthy UPD dataset (n = 47) resulted in poor performance, with a loss of 1.56 and an accuracy of 0.23. As shown in the confusion matrices in Figure 3.3, the model almost exclusively predicted N1, regardless of the true class label. Due to this poor performance, no ROC and AUC analysis was conducted.

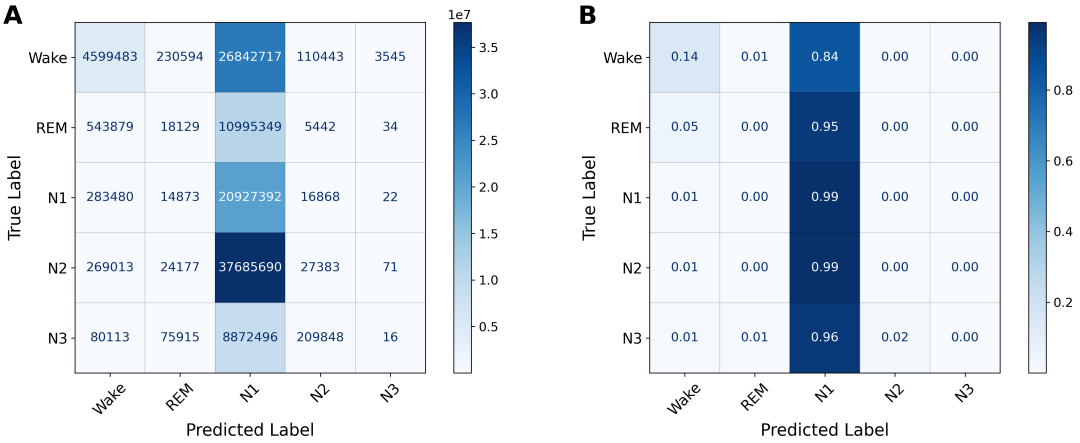

Figure 3.3: Confusion matrices from evaluating the SC pretrained model on the healthy UPD dataset. (A) Absolute counts of true versus predicted labels. (B) Row-normalized confusion matrix showing the percentage distribution of predicted classes for each true label.

After retraining the model on the UPD train dataset, performance on the test data (n = 8) improved, with a test loss of 1.36 and accuracy of 0.49. The overall F1 score was 0.45, and class-wise F1 scores were 0.71 for Wake, 0.44 for REM sleep, 0.22 for N1, 0.54 for N2, and 0.000095 for N3. The confusion matrices (Figure 3.4) show that Wake and REM were often correctly predicted, while N1 was frequently misclassified as Wake or REM. N2 was correctly classified in slightly over half of the cases. N3 was rarely predicted and often misclassified as N2.

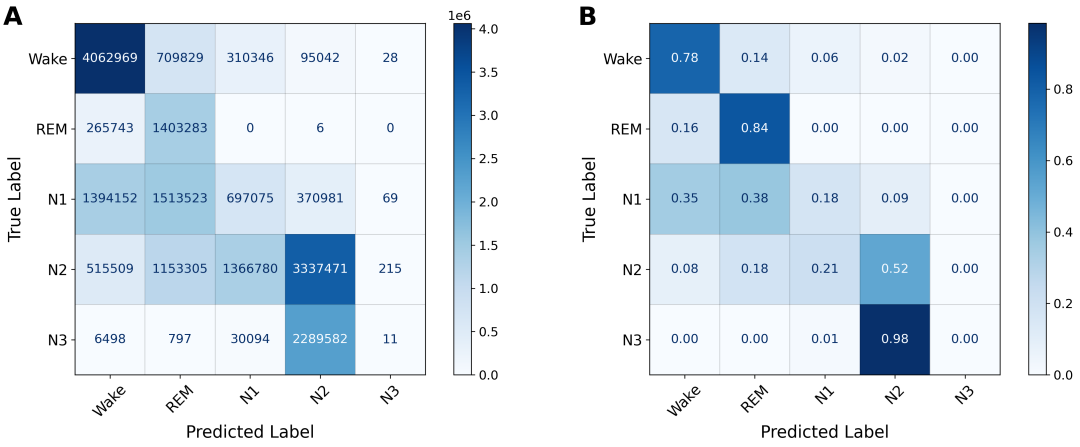

Figure 3.4: Confusion matrices from the UPD test set after retraining the model on the UPD training set. (A) Absolute counts of true versus predicted labels. (B) Row-normalized confusion matrix showing the percentage distribution of predicted classes for each true label.

ROC curve analysis (Figure 3.5) showed high one-vs-rest AUC values over 0.75, except for N1 with a value of 0.58. One-vs-one AUCs (Table 3.4) also exceeded 0.75 for all class pairs except N2 vs. N3, which was notably lower at 0.14.

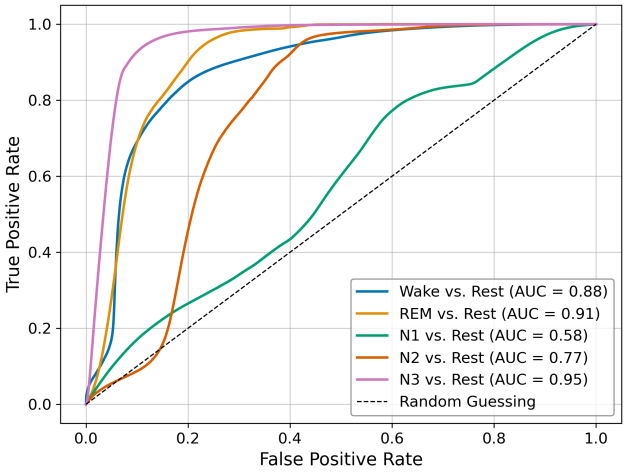

Figure 3.5: One-vs-rest ROC curves and corresponding AUC values for all sleep stages in the UPD test set.

Table 3.4: One-vs-one AUC scores for all sleep stage pairs in the UPD test set. Each cell reports the AUC for distinguishing the row class from the column class based on normalized one-vs-one probabilities summing to 1.

	Wake	REM	N1	N2	N3
Wake	–	0.92	0.78	0.95	1.00
REM		–	0.83	0.99	1.00
N1			–	0.85	0.98
N2				–	0.14
N3					–

3.3 Downstream Tasks

3.3.1 Brain Age Prediction

SC Dataset

Table 3.5 presents the performance of models selected based on their validation set (n = 21) performance, either by lowest MAE or highest Pearson's r, and evaluated on the test set (n = 25).

Table 3.5: Evaluation results on the SC test set for age prediction models selected according to their validation set performance, using either minimum MAE or maximum Pearson's r as selection criteria. When both MAE and r are mentioned as the selection metric, the same model produced the lowest MAE and the highest Pearson's r. Default hyperparameters were used: for Ridge Regression, *alpha*=1.0; for SVR, *C*=1.0 and *gamma*=`scale`.

#	Input Features	Model	Selection Metric	MAE	r	ρ
1	(1) Flattened	(C) SVR (kernel=sigmoid)	MAE & r	17.36	0.54	0.62
2	(2) Averaged	(C) SVR (kernel=rbf)	MAE	20.53	0.57	0.58
3	(2) Averaged	(B) Ridge Regression	r	14.47	0.67	0.66
4	(3) PCA Flattened	(C) SVR (kernel=sigmoid)	MAE	18.50	0.51	0.38
5	(3) PCA Flattened	(A) Linear Regression	r	18.50	0.51	0.38
6	(4) PCA Averaged	(C) SVR (kernel=rbf)	MAE & r	15.49	0.78	0.76
7	(5) Correlated Averaged	(C) SVR (kernel=rbf)	MAE & r	16.93	0.73	0.70

The chronological age distribution in both the validation and test sets was non-normal, according to the Shapiro-Wilk test ($p < 0.05$). Predicted age distributions varied by model. The 10 features most strongly correlated with age all achieved a Spearman's ρ of over 0.50, with the highest at 0.54.

Model 3 (Ridge Regression on averaged features) achieved the lowest MAE on the test set, with 14.47 years and correlations of r = 0.67 and ρ = 0.66. Model 6 (SVR with RBF kernel on PCA averaged features) obtained the highest correlations of r = 0.78 and ρ = 0.76 with a slightly higher MAE of 15.49 years.

Figure 3.6 (Model 3) and Figure 3.7 (Model 6) show predicted ages against chronological ages and residuals against chronological ages. Model 6 clearly underestimated the age of younger people, while the age of older people was systematically overestimated. The residuals of Model 3 are slightly more evenly distributed.

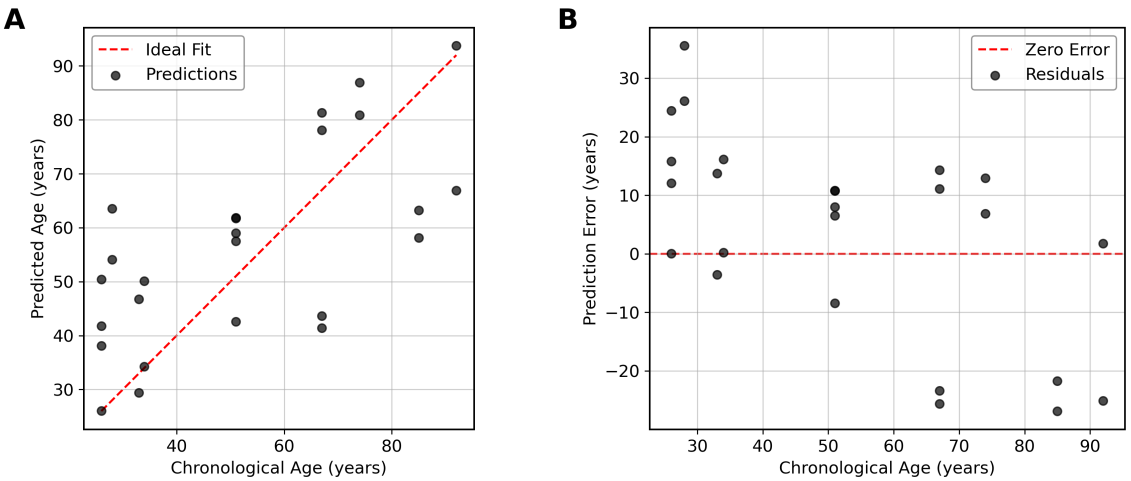

Figure 3.6: Scatterplots showing the performance of Model 3 (Ridge Regression with averaged feature input) on the SC test set (n = 25). (A) Predicted age plotted against chronological age. (B) Residuals (prediction error) plotted against chronological age, showing the deviation of model predictions from actual age across the age range.

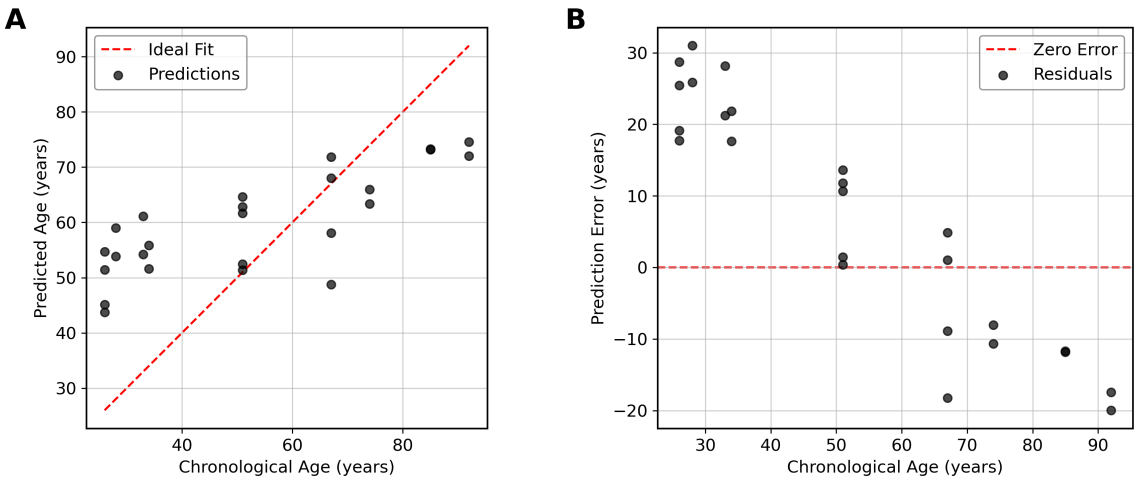

Figure 3.7: Scatterplots showing the performance of Model 6 (RBF SVR with PCA averaged feature input) on the SC test set (n = 25). (A) Predicted age plotted against chronological age. (B) Residuals (prediction error) plotted against chronological age, showing the deviation of model predictions from actual age across the age range.

UPD Data

Table 3.6 shows the leave-one-out cross-validation results (n = 47). Model selection was based on lowest MAE or highest Pearson's r per feature type during leave-one-out cross-validation.

Table 3.6: Leave-one-out cross-validation results from the UPD healthy set for age prediction models selected based on minimum MAE or maximum Pearson's r as selection criteria. When both MAE and r are mentioned as the selection metric, the same model produced the lowest MAE and the highest Pearson's r. Default hyperparameters were used: for Ridge Regression, *alpha*=1.0; for SVR, *C*=1.0 and *gamma*=`scale`.

#	Input Features	Model	Selection Metric	MAE	r	ρ
1	(1) Flattened	(C) SVR (kernel=sigmoid)	MAE	3.30	0.22	0.18
2	(1) Flattened	(B) Ridge Regression	r	3.73	0.25	0.19
3	(2) Averaged	(C) SVR (kernel=rbf)	MAE	3.47	0.024	0.10
4	(2) Averaged	(B) Ridge Regression	r	5.23	0.11	0.12
5	(3) PCA Flattened	(B) Ridge Regression	MAE & r	3.14	0.34	0.25
6	(4) PCA Averaged	(C) SVR (kernel=linear)	MAE & r	3.31	0.36	0.32
7	(5) Correlated Averaged	(C) SVR (kernel=linear)	MAE & r	3.46	0.14	0.17

The chronological ages in this set of healthy subjects followed a normal distribution according to the Shapiro-Wilk test (p = 0.21). The distribution of predicted ages varied across models. The top-correlated features with age achieved a ρ of around 0.3, depending on the leave-one-out split.

Model 5 (Ridge Regression on PCA flattened features) had the lowest MAE at 3.14 years, with correlations of r = 0.34 and ρ = 0.25. Model 6 (SVR with a linear kernel on PCA averaged features) showed the highest correlation coefficients, with r = 0.36 and ρ = 0.32, and a comparable MAE at 3.31 years.

Figure 3.8 (Model 5) and Figure 3.9 (Model 6) show similar bias patterns as in the SC dataset, with an underestimation of younger and overestimation of older subjects. Most predictions cluster around the mean age.

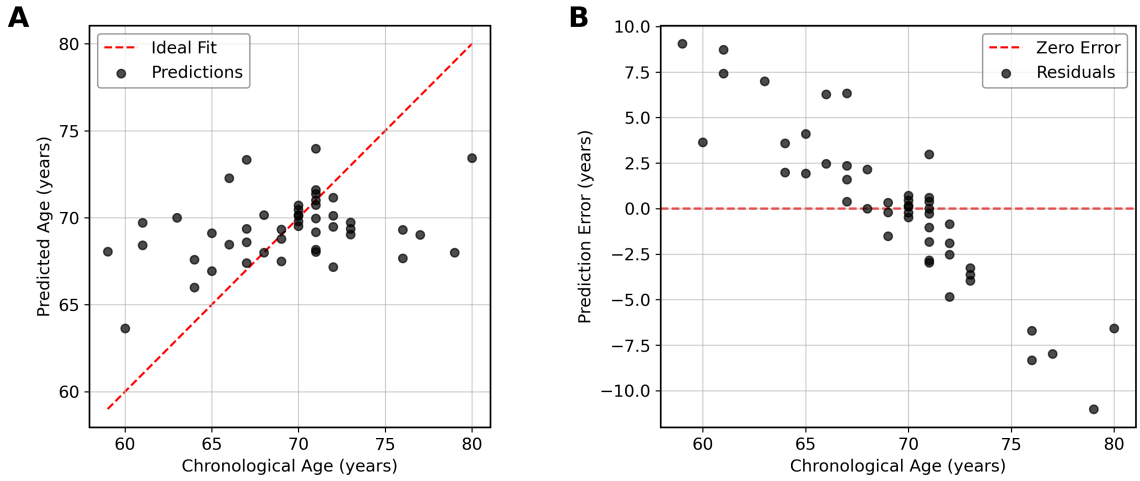

Figure 3.8: Scatterplots showing the cross-validation performance of Model 5 (Ridge Regression with PCA flattened feature input) on the UPD dataset (n = 47). (A) Predicted age plotted against chronological age. (B) Residuals (prediction error) plotted against chronological age, showing the deviation of model predictions from actual age across the age range.

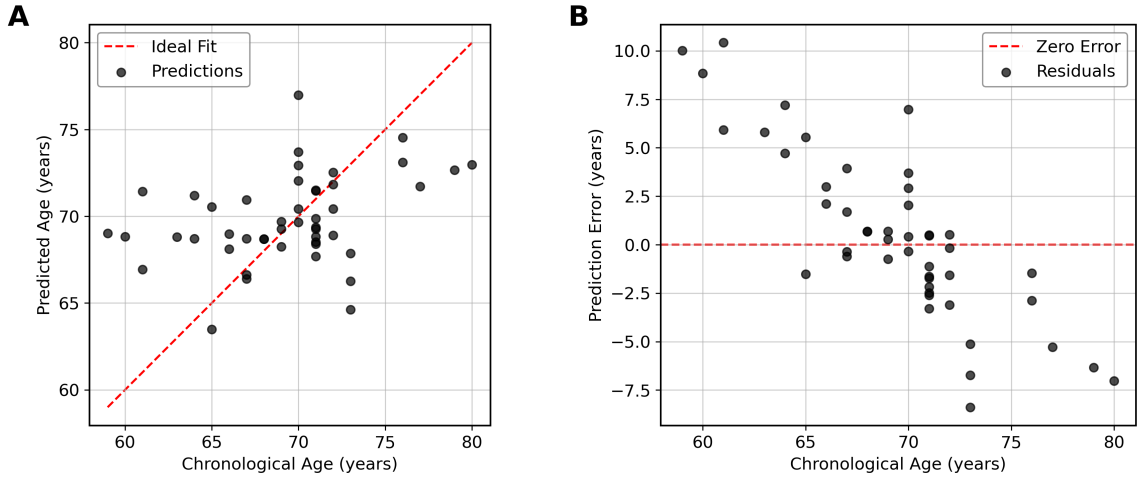

Figure 3.9: Scatterplots showing the cross-validation performance of Model 6 (Linear SVR with PCA averaged feature input) on the UPD dataset (n = 47). (A) Predicted age plotted against chronological age. (B) Residuals (prediction error) plotted against chronological age, showing the deviation of model predictions from actual age across the age range.

3.3.2 MCI Classification

Table 3.7 summarizes the best-performing models for MCI classification on the full UPD dataset (n = 64), selected based on either the highest F1 score or AUC during leave-one-out cross-validation.

Table 3.7: Leave-one-out cross-validation results from the UPD dataset for MCI classification models selected based on maximum AUC and maximum F1. When both F1 and AUC are mentioned in the selection metric column, the same model produced the highest F1 score and highest AUC. Default hyperparameters were used except for class weights or priors: for Logistic Regression, *penalty*=`l2`, *C*=1.0, *class_weight*=`balanced`; for SVR, *C*=1.0, *class_weight*=`balanced`; for Gaussian Naïve Bayes, *priors* were set based on the observed class frequencies.

#	Input Features	Model	Selection Metric	Accuracy	AUC	F1
1	(1) Flattened	(B) SVC (kernel=linear)	AUC	0.73	0.49	0.00
2	(1) Flattened	(A) Logistic Regression	F1	0.63	0.37	0.20
3	(2) Averaged	(D) Gaussian Naïve Bayes	AUC & F1	0.64	0.56	0.41
4	(3) PCA Flattened	(D) Gaussian Naïve Bayes	AUC	0.66	0.57	0.21
5	(3) PCA Flattened	(A) Logistic Regression	F1	0.55	0.50	0.36
6	(4) PCA Averaged	(D) Gaussian Naïve Bayes	AUC & F1	0.70	0.54	0.30

Model 3 (Gaussian Naïve Bayes on averaged features) achieved the highest F1 score of 0.41 and the second-best AUC of 0.56 with an accuracy of 0.64. Even though this model outperformed others, the ROC curve (Figure 3.10 A) lies only slightly above chance, and the confusion matrix (Figure 3.10 B) reveals a tendency to misclassify both MCI and healthy cases.

While Model 4 showed a slightly higher AUC (0.57), its F1 score was considerably lower (0.21), making Model 3 a better choice. However, with an F1 below 0.5 and AUC barely above 0.5, even the best model showed weak performance.

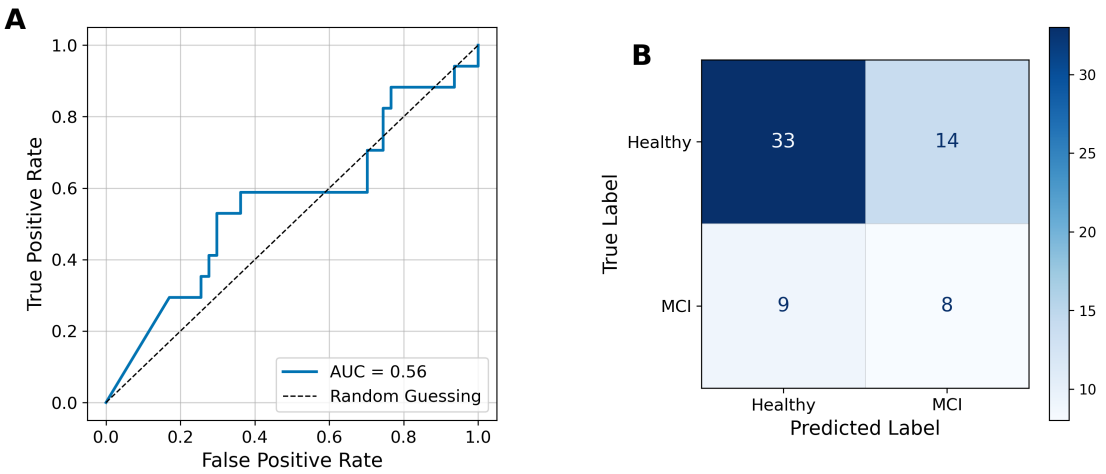

Figure 3.10: Plots showing the cross-validation performance of Model 3 (Gaussian Naïve Bayes with averaged feature input) for MCI classification on the UPD dataset (n = 47). (A) ROC Curve with indicated AUC value. (B) Confusion matrix with absolute values of true versus predicted labels.

4 Discussion

4.1 Correlational Analyses

Correlations between sleep architecture features and age were tested for both datasets. All correlations in the UPD dataset were weaker and non-significant compared to those found in the SC dataset. The smaller sample size of the UPD dataset likely contributed to this, as it may have reduced statistical power. Additionally, due the narrower age range in the UPD dataset, the age-related differences were possibly smaller. The significant correlations observed in the SC dataset are consistent with previous findings in the literature. Conforming with the meta-analysis by Ohayon et al. (2004), a positive correlation between age and N1 sleep and negative correlations between age and both N3 and REM sleep were observed. However, some expected relationships reported in the literature, specifically positive correlations between age and Wake (corresponding to WASO), N2, and TST (Ohayon et al., 2004), were not observed in the SC dataset. These results show that some sleep architectural features reflect age-related changes, supporting their relevance for brain age prediction. Examining correlations between age and microstructural sleep features could have provided additional insight, given prior evidence of age-related changes (Carrier et al., 2011; Helfrich et al., 2018; Mander et al., 2017; Purcell et al., 2017). However, this was beyond the scope of this thesis.

4.2 Pretext Task

The sleep staging model trained on the SC dataset demonstrated solid performance, indicating that the model effectively learned relevant features from the data. It achieved an overall accuracy of 83%, lower than state-of-the-art models, which report accuracies exceeding 90% (Gaiduk et al., 2023). However, as accuracy is highly sensitive to class distribution, the overall F1 score of 0.82 provides a more informative measure. This performance is comparable to the mean F1 score of 0.79 in the original U-Sleep model, with class-wise F1 scores also being similar (Perslev et al., 2021). Other top-performing models achieved F1 scores between 0.85 and 0.95 (Gaiduk et al., 2023). The Cohen's Kappa of 0.70 indicates substantial agreement (Landis & Koch, 1977), comparable to interrater agreement for manual scoring, typically ranging from 0.68 to 0.76 (Danker-Hopfe et al., 2009; Y. J. Lee et al., 2022). While direct comparison to human-level performance is limited, since the developed model is evaluated against its own training labels, Cohen's Kappa remains a useful metric for comparison against other automated models, some of which reported values above 0.80 (Gaiduk et al., 2023). However, the values reported by Gaiduk et al. (2023) are results from the highest performing recent models, whereas many other recent models also reported evaluation metrics in the same range as the implemented model on SC data (Ganglberger et al., 2024; Gunnarsdottir et al., 2020; X. Zhang et al., 2020).

The classification of N1 showed the lowest performance, often being misclassified as N2 or Wake. This has also been observed in prior studies (Gunnarsdottir et al., 2020; Haghayegh et al., 2023; Perslev et al., 2021; X. Zhang et al., 2020). N1 is a transitional stage between Wake and N2, and shares some EEG features with both of those stages (Hofman & Talamini, 2015). Additionally, N1 is underrepresented compared to N2 and Wake, leading to a lower prior probability, which can further impact classification. Moreover, N1 also has the lowest rate of manual interrater agreement, likely due to the difficulty in identifying transitions to and from N1 (Danker-Hopfe et al., 2009; Y. J. Lee et al., 2022). This interrater variability could further explain the problem of N1 classification and may also impact the supervised learning performance, as the manual labels might not always be physiologically accurate.

Some misclassifications of N3 as N2 also occurred, likely due to overlapping characteristics of deeper NREM stages, particularly the gradual increase in delta wave activity (Hofman & Talamini, 2015). These difficulties have also been reported in other automated sleep staging studies (Gunnarsdottir et al., 2020; Stephansen et al., 2018).

Despite some class confusion, ROC curves and AUC values indicated strong class separability. The lowest one-vs-rest AUC was 0.85 for N1, with all other classes achieving higher values. One-vs-one AUC values were also high (> 0.90), with the lowest between Wake/N1 and N1/N2, again reflecting the transitional state of N1. Only few existing studies looked at ROC curves and corresponding AUCs. Ganglberger et al. (2024) report AUC values from 0.84 for N1 to 0.96 for Wake and REM, thus, in a similar range as what was found in sleep staging on the SC dataset. The high AUC values suggest that the model captures useful class-specific representations from the SC data.

When the SC-trained model was directly applied to the UPD dataset without retraining, model performance dropped significantly, and almost exclusively N1 was predicted. This is likely due to considerable differences between the datasets. The UPD dataset utilized a different EEG system with different electrode setup and did not include an EMG channel. Furthermore, the UPD dataset included only older participants. These differences may have reduced generalizability, which is a known challenge of transfer learning when domains are not similar enough, for example, when recording equipment or biological factors differ (Ganglberger et al., 2024; Zhuang et al., 2021). Due to the small dataset sizes and these substantial differences, fine-tuning was not pursued and the whole model was retrained from scratch on the UPD dataset.

The retrained model on the UPD dataset showed lower accuracy (49%) and F1 scores (overall F1: 0.45) than on the SC dataset and in existing approaches (Gaiduk et al., 2023; Perslev et al., 2021) but performance was still above chance level for five classes. The confusion matrix confirmed that different sleep stages were predicted, contrary to the results of direct transfer. Most one-vs-rest AUC values exceeded 0.75, except for N1, suggesting that the model still learned distinguishable patterns.

As with the SC data, classification of N1 remained challenging and was often misclassified as Wake or REM, highlighting the transitional state of N1 from Wake and the lower prior

probability of N1 again. Misclassification of N1 as REM has also been observed in existing studies (Gunnarsdottir et al., 2020; Haghayegh et al., 2023) and might be explained by some similarities of the lower-amplitude EEG signal including theta activity in both stages (Carlson & Birkett, 2021).

N3 was rarely predicted and often misclassified as N2. In addition to the reasons for this class confusion mentioned before, the age range of the UPD dataset likely played a role. Age-related reductions in slow-wave amplitude make N3 less distinct and a clear-cut distinction between N2 and N3 challenging (Carrier et al., 2011; Muehlroth & Werkle-Bergner, 2020). This could affect manual and automated scoring in older individuals and further complicate training from human labels (Y. J. Lee et al., 2022). In spite of these misclassifications, the high AUC of 0.95 for N3-vs-rest suggests potential for improving N3 classification via recalibration of thresholds.

Despite reduced performance, the ROC curves and AUC values indicate that the UPD trained model still captures distinguishable sleep stage patterns, supporting its suitability for feature extraction for use in downstream tasks.

A general problem of supervised sleep staging is its reliance on manual labels, which are known to be inconsistent, especially for transitional stages, such as N1 (Danker-Hopfe et al., 2009; Y. J. Lee et al., 2022). Fixed 30-second epochs may not align well with true physiological transitions, leading to inaccuracies in the labels the model learns from (Fiorillo et al., 2019). Consequently, model performance is limited by label quality. Alternatives like hypnodensity graphs, which display probabilities of the different sleep stages, provide more continuous representations about sleep trends (Anderer et al., 2022; Stephansen et al., 2018; Sun et al., 2023). While the implemented model outputs such probabilities, visualizing them in this format was not explored in this thesis.

The discussed results were obtained without hyperparameter tuning. Although optimization of hyperparameters or finding the optimal point for stopping training might have improved sleep staging accuracy, the aim was to use this model as a pretext task for brain age prediction rather than to optimize sleep staging.

4.3 Downstream Tasks

4.3.1 Brain Age Prediction

The best-performing age prediction models on the SC dataset achieved MAEs of 14.47 and 15.49 years, which are higher than those reported in previous studies, ranging from 4.19 to 7.6 years, while correlations were lower, with Pearson's r of 0.67 and 0.78 compared to 0.83 to 0.97 in prior work (Banville et al., 2024; Brink-Kjaer et al., 2022; Sun et al., 2019; Yook et al., 2022; D. Zhang et al., 2024).

When looking at individual models, it can be observed that more complex models often underperformed, suggesting that model capacity alone does not explain the results. For example,

Ridge Regression (Model 3) outperformed RBF SVR (Model 2) despite being simpler, and sigmoid SVR (Model 4) performed no better than linear regression (Model 5) on PCA-reduced inputs. These results indicate that the used features likely lack the information to support the modeling of complex relationships.

The tested models performed even worse on the UPD dataset, although the MAEs of 3.14 and 3.31 years were lower than in the SC dataset and in existing models (Banville et al., 2024; Brink-Kjaer et al., 2022; Sun et al., 2019; Yook et al., 2022; D. Zhang et al., 2024). This was due to the narrow age range, which leads to smaller errors when the predictions are close to the mean, and does not indicate better model performance, as also highlighted by de Lange et al. (2022). The Pearson, but also Spearman, correlations between predicted and chronological ages were weak, with a maximum r of 0.36, lower than in the SC dataset and in existing models (Brink-Kjaer et al., 2022; Sun et al., 2019; Yook et al., 2022; D. Zhang et al., 2024). The performance of all models on the UPD dataset was unsatisfactory, and no apparent differences by model choice or input features can be highlighted.

The choice of evaluation metrics in brain age prediction is crucial for meaningful comparison across studies, as each metric has inherent strengths and limitations. This trade-off supports the use of multiple complementary metrics. For example, Pearson's r between predicted and chronological age could be misleadingly high if predictions are systematically shifted by a constant offset. As chronological age was not normally distributed according to the Shapiro-Wilk test, Spearman's rank correlation coefficient was additionally computed, capturing monotonic relationships. However, both metrics yielded similar results across models. In an ideal case in healthy subjects, predicted age should match chronological age, resulting in a linear and monotonically increasing relationship, and we would expect both to be high. Additionally, MAE was included for comparability with existing studies. In retrospect, including the coefficient of determination (R^2) could have provided additional insight into the explained variance, as used in other studies (Cole et al., 2017a; de Lange et al., 2022). Importantly, de Lange et al. (2022) emphasizes that brain age prediction metrics depend highly on the dataset and cannot always directly be compared between different studies.

The sample sizes likely limited the models' performances, and larger datasets with a broader age distribution than in the UPD dataset would likely be necessary for improvement. The study by de Lange et al. (2022) highlighted that both Pearson's r and MAE are typically lower in datasets with narrower age ranges and that performance metrics can improve with an increased sample size. Existing studies used significantly larger datasets for model training, often from multiple cohorts. For example, Yook et al. (2022) used 1259 PSGs and D. Zhang et al. (2024) even trained their model on 17294 PSGs. Furthermore, they also utilized datasets with larger age ranges, similar to or even broader than those of the SC dataset (Brink-Kjaer et al., 2022; Sun et al., 2019; D. Zhang et al., 2024). To my current knowledge, no study has attempted to predict age from a dataset with such a small age range as the UPD dataset.

Furthermore, most existing highly accurate studies directly predicted brain age from raw or engineered EEG features. Yook et al. (2022) used scalograms as input, Sun et al. (2019) used manually selected sleep features, Banville et al. (2024) used covariance matrices, and Brink-Kjaer et al. (2022) and D. Zhang et al. (2024) predicted age directly from raw PSG and EEG measurements. These studies did not rely on pretext tasks.

In an abstract supplement, Ganglberger et al. (2022) proposed a U-Net-like deep neural network to predict sleep stages, age, and neuropsychological test scores. Their architecture shares similarities with the model in this thesis, although they employed a single model that predicted multiple outcomes instead of a two-step pretraining approach. They reported a Cohen's Kappa of 0.54 for sleep staging and a Pearson's r of 0.60 between predicted and chronological age, both slightly lower than the performance achieved by the model presented here on the SC dataset. However, to the best of my knowledge, no article has been published on this model.

A notable pattern across both datasets was the systematic bias in predictions, with models regressing to the mean. Models tended to overestimate the age of younger subjects and underestimate that of older subjects. This problem has also been observed in other brain age prediction models (Brink-Kjaer et al., 2022; Cole et al., 2017a). While some previous work has explicitly corrected for this bias (Yook et al., 2022), this was not done in this thesis. However, de Lange et al. (2022) reported that such corrections can inflate performance metrics for poorly performing models, and uncorrected model results should be inspected to evaluate model validity.

It would be valuable to explore what predicted brain age and brain age gap from sleep electrophysiology can reveal about sleep itself, particularly in contrast to estimates derived from rs-EEG or MRI. Both D. Zhang et al. (2024) and Sun et al. (2019) report that Wake and light sleep are associated with a higher brain age gap, while deep and REM sleep are linked to a lower one. Brink-Kjaer et al. (2022) further found a link between an increased brain age gap and sleep fragmentation metrics. Yook et al. (2022) observed significantly higher brain age gaps in groups with OSA, insomnia, and comorbid insomnia and OSA, and a higher brain age gap was further associated with an increased risk of developing these disorders. However, this association was not consistently replicated by other studies (Sun et al., 2019; D. Zhang et al., 2024). Sleep-disordered breathing and insomnia, which have both also been linked with cognitive decline and dementia (Chang et al., 2013; Osorio et al., 2011; Wennberg et al., 2017; Yaffe et al., 2011), have been the focus of existing research, while the association of other sleep disorders with brain aging remains unexplored.

Existing studies have also shown that an elevated brain age gap, derived from sleep EEG signals, correlates with conditions involving cognitive decline, including MCI and AD (Ye et al., 2020). This supports the use of sleep-derived brain age gap as a biomarker for the risk of neurodegeneration and cognitive impairment, and also the usage of similar features for MCI classification as for brain age prediction.

Due to the limited performance of the brain age prediction models, especially on the UPD dataset, the predefined hypotheses were not tested. Robust testing would require a more accurate brain age prediction model as a foundation. However, MCI classification was explored as a second downstream task.

4.3.2 MCI Classification

The overall performance of the MCI classification task was poor across all models. Accuracy was moderate, while AUC values and F1 scores were consistently low, indicating a limited ability to discriminate between individuals with MCI and healthy individuals. High accuracy in some models, such as Model 1, was misleading, as classifying all subjects as healthy results in an accuracy of 73%. Therefore, metrics like F1 score and AUC provide a more meaningful assessment of performance and should be prioritized. Among the tested classifiers, Logistic Regression and Gaussian Naïve Bayes performed best, though their results remained poor.

Focusing on Model 3, which resulted in the best performance, sensitivity is low as only eight MCI samples were correctly classified, whereas nine were misclassified as healthy. Model 3 achieved an accuracy of 64% and an F1 score of 0.41, reflecting the model's limited ability to classify the MCI class. This tendency to over-predict the healthy class could be due to the limited number of MCI samples, while almost double that number is available for healthy subjects. However, the low AUC value of 0.56 suggests that imbalance alone cannot fully explain the poor performance. Model 3, and also other tested models, might have failed to learn meaningful patterns from the available features to distinguish these two classes, suggesting that the features extracted from sleep staging might not sufficiently capture MCI characteristics.

The AUC score remained close to chance level, indicating that the model struggled to find a meaningful decision boundary and was likely underfitting the data. While more complex models could potentially improve performance, they would likely require more training data, especially from the MCI class.

In contrast, existing studies have reported better results for MCI classification. For instance, Ye et al. (2023) achieved an AUC of 0.73 and Geng et al. (2022) an AUC of 0.98 and a F1 score of 0.94. Haghayegh et al. (2025) predicted the later development of cognitive impairment with an AUC of 0.76. These studies generally utilized larger and/or more balanced datasets. Geng et al. (2022) used a small but perfectly balanced dataset of 40 subjects, Haghayegh et al. (2025) used data from 281 women, and Ye et al. (2023) used a larger cohort of over 8000 subjects. These studies all used manually selected features as input.

For this second downstream task, again, no model achieved satisfactory performance, leaving the optimal choice of model and input features inconclusive.

4.3.3 Disussion Across Downstream Tasks

For both downstream tasks, classical shallow machine learning approaches were used. Model choice reflects a trade-off between bias and variance, where simpler models risk underfitting, leading to high bias, while more complex models might overfit and have high variance (Geman et al., 1992). With more data, more complex approaches, such as deep learning, could potentially improve performance. However, given the limited sample size, using such methods would not have been appropriate, as it would have risked overfitting. This limitation affects age regression and MCI classification more than the pretext task sleep staging. The reason is that age regression and MCI classification rely on a single label per subject, resulting in fewer training examples. In contrast, sleep staging benefits from continuous labels recorded at high temporal resolution (100 Hz) throughout entire nights, providing a large number of labeled data points that better support complex model training.

The rather limited performance on both downstream tasks suggests that the features learned during sleep staging may not sufficiently capture age- or cognition-related patterns. They might be too specific to sleep stage classification and not generalize well to other tasks. Limited correlation between the averaged learned features and age during feature selection could support this claim. More generalizable feature representations could potentially be learned using SSL, which leverages the internal structure of the data rather than relying on human-annotated labels. As noted by Bommasani et al. (2022), a pretext task constrains the downstream capabilities. Zhou et al. (2024) also highlights the difficulty of matching a pretext task to a downstream task. Thus, if sleep stage related patterns poorly reflect aging or MCI in the electrophysiology signal, performance will remain limited.

Another restriction arises from the feature processing approach. By averaging features or reducing them to 10 principal components, essential temporal dynamics may have been lost. This dimensionality reduction may have oversimplified the data, further limiting model performance. Exploring temporally aware downstream models, such as RNNs or LSTMs, might capture age- or cognition-related dynamics within extracted features better.

Recent advances have started to explore foundation models for sleep research, using large-scale datasets and SSL to learn generalizable representations that transfer well to a different downstream tasks. SleepFM used multimodal contrastive learning on multi-modal sleep data from 65000 participants and achieved strong results across downstream tasks such as age prediction (MAE: 7.33 years, r = 0.88), prediction of future development of MCI (AUC: 0.84), gender classification, sleep stage classification (F1: 0.69 - 0.78), and apnea diagnosis (AUC: 0.90) (Thapa et al., 2025). C.-H. Lee et al. (2025) combined contrastive learning with masked prediction and also incorporated temporal context, achieveing high accuracies on sleep-stage classification and apnea detection. Ogg and Coon (2024) used a masked transformer for pretext training, reaching an MAE of 15 years for age prediction, comparable to the results of my model on the SC dataset, and good performance on sleep staging. The focus was thus on brain age

prediction and apnea detection, illustrating the clinical potential of foundation models for sleep data, which could be expanded to detection of other (sleep) disorders. These studies highlight the effectiveness of SSL, whether based on contrastive learning, masked prediction, or their combination, in learning generalizable features from sleep data. Such approaches show promise for building general-purpose foundation models for sleep electrophysiology measurements that can be fine-tuned for various tasks in research and clinical applications.

In summary, no combination of model and feature input clearly outperformed others across datasets or downstream tasks and the results remain inconclusive regarding the best model and input choice. Given the small sample sizes, especially problematic for non-linear methods, larger datasets would likely be needed for more complex models to reach their full potential.

4.4 Limitations

A limitation of this study is the relatively small amount of training data, as only two datasets with a total of 220 nights of recordings were used. This is significantly less than in other brain age prediction studies, which have used between 1200 and 13300 nights, often from multiple cohorts (Brink-Kjaer et al., 2022; Sun et al., 2019; Yook et al., 2022). Prior work with MRI and rs-EEG has shown that the number of training samples had a strong influence on the accuracy of age prediction (Al Zoubi et al., 2018; Franke et al., 2010), suggesting that model performance could possibly benefit from more training data.

The two available datasets were used independently for training and evaluation of each model rather than being merged. Various studies working with sleep EEG data improved generalization by combining datasets (Brink-Kjaer et al., 2022; Perslev et al., 2021; D. Zhang et al., 2024), however, differences in recording systems and preprocessing steps between the SC and UPD datasets might limit the benefits of such an approach.

Additionally, the same data was used for the development and evaluation of both the pretext and downstream tasks in both datasets. To avoid bias, a separate hold-out set should be used exclusively for final evaluation.

No data augmentation techniques were implemented during model training. Augmentation methods, like adding noise, simulating missing channels, or inverting signals, have been effective in existing studies (Perslev et al., 2021; D. Zhang et al., 2024) and may be useful for mitigating small dataset size and improving model robustness and generalization.

Furthermore, the models were trained using only four channels. This was a constraint of the SC dataset, whereas the UPD dataset offered additional channels that were not used. Including more channels could have increased spatial resolution and provided additional information. Nonetheless, using a reduced set of channels makes the approach more applicable to low-density or home-use EEG systems, and existing sleep EEG studies have also only used a limited number of two to six input channels (Banville et al., 2024; Sun et al., 2019; Yook et al., 2022; D. Zhang et al., 2024).

A significant limitation of the UPD dataset was the narrow age range and small overall sample size. The limited age variance constrains the learning of age prediction models and may have limited the ability to detect age-related trends, as changes in sleep architecture may have become more pronounced across a broader age range. Due to the limited size of the UPD dataset, leave-one-out cross-validation was used instead of evaluation on an independent test set. While this is common in smaller datasets, using an independent test set would have allowed for a more robust assessment of model performance. These dataset characteristics likely contributed to the poor performance observed in the UPD models.

In addition, the number of participants with MCI was small, and their MoCA scores were often near the diagnostic cutoff. Including participants with a broader age spectrum and more pronounced cognitive impairment could have improved both brain age regression and MCI classification.

Additionally, no hyperparameter optimization was conducted for the downstream models. While preliminary tests suggested that variations had minimal impact, a systematic search (e.g., grid search) could still lead to performance gains, even if unlikely to change the overall conclusions.

Moreover, the brain age regression models did not include constraints to prevent implausible predictions of negative ages. Introducing output constraints, as Sun et al. (2019) did, could help improve models in future implementations.

Finally, the deep sleep staging model lacked interpretability, which is typical for such deep learning models, making it difficult to understand what the learned features represent and how transferable they are to downstream tasks (Sun et al., 2023).

4.5 Reflections and Improvements

This project revealed some challenges of applying foundation model principles to sleep electrophysiology signals for brain age prediction, while the literature review highlighted its potentials.

Some limitations could have been mitigated by combining data from multiple cohorts, such as publicly available sleep EEG datasets, to increase diversity and cover broader age ranges. This would enable larger-scale training and allow learning curves to assess whether model performance saturates or continues to improve with more data.

For the pretext task, I would, in hindsight, increase the duration of the reduced sample weights, which were now applied only one second before and after a stage transition, which is rather short for 30-second epochs. However, as performance on the SC dataset was already good, I do not expect this change to make a large difference.

Given more time, I would have explored different pretext tasks beyond sleep staging. SSL methods, such as contrastive learning or masked prediction, especially using transformer architectures, may allow better feature extraction from sleep EEG data. Zhou et al. (2024) highlights that many successful foundation models are based on a transformer architecture, utilizing atten-

tion mechanisms. These approaches are more aligned with current trends in foundation model development and may improve performance in downstream tasks, such as brain age prediction and MCI classification, and could additionally be tested on other downstream tasks.

Moreover, the approach of model evaluation of the downstream task, testing multiple models and input types without systematic optimization, may have made the process less efficient and introduced bias. Focusing on one or two well-justified model architectures and input feature transformations, and conducting hyperparameter tuning could have been more effective and led to better interpretable results. Additionally, selecting and reporting only the best-performing models based on leave-one-out cross-validation introduces bias, as both model selection and performance evaluation were carried out on the same dataset. This issue could have been avoided using approaches such as nested cross-validation.

Furthermore, to facilitate the interpretation of MAE in datasets with varying age ranges, I would suggest comparing the obtained MAE to that of a dummy predictor that always predicts the median of the dataset, as Banville et al. (2024) did.

Once a reliable model is developed, it could be used to test the hypotheses about the relationship between MCI and increased brain age gap, as well as the effect of interventions like PLAS on brain age gap, as originally planned.

4.6 Future Directions and Implications

While the results of the use of automated sleep staging as a pretext task for brain age prediction were limited, the general idea of developing foundation models for clinical signals such as sleep EEG recordings remains relevant. Ultimately, a large-scale foundation model for sleep electrophysiology recordings could be trained on diverse datasets and evaluated across multiple downstream tasks, similar to what SleepFM already did with multimodal sleep data (Thapa et al., 2025). Having reliable foundation models in the medical field could facilitate research and clinical diagnosis, even for people who are not machine learning professionals. Bommasani et al. (2022) discusses multiple opportunities for foundation models in healthcare and biomedical research, highlighting the potential of multimodal medical foundation models.

Additionally, improving the biomarker of brain age gap could make it useful for further research or clinical applications. While it has already been demonstrated that brain age gap from sleep EEG is increased in MCI (Ye et al., 2020), it would still be valuable to replicate these findings using different models and datasets. This could validate the use of sleep EEG based brain age gap as a biomarker for early detection of cognitive impairment. Furthermore, it has been suggested that brain age gap could be used at different stages of diseases, such as assessing general brain health, detecting disorders early, making prognostic predictions, supporting differential diagnosis, making treatment decisions, and monitoring potential reduction of brain age by interventions (Baecker et al., 2021). Therefore, testing these possibilities, for example, the effect of interventions such as PLAS on brain age, presents a relevant future research direction.

5 Conclusion

This thesis aimed to develop and evaluate a foundation model-based approach to predict brain age from sleep electrophysiology data. The objective was to implement an automated sleep staging pretext task and use the learned features to predict brain age. It was intended to then test the effect of MCI and a PLAS intervention on brain age gap. The proposed pretext task successfully predicted sleep stages, achieving good results on the SC dataset and moderate results on the UPD dataset. The downstream task of brain age prediction using the learned representations from the pretext task did not result in accurate predictions of age, producing models that were insufficient for testing the effect of MCI and PLAS on brain age gap. This thesis tested a new principle, using supervised automated sleep staging as a pretext task to learn general features, which showed to be unsuitable for this data and the chosen downstream tasks.

Some limitations should be noted. The models were trained and validated on two datasets with limited sizes, with one also having a narrow age range, and no hyperparameter optimization was applied. Using self-supervised pretext tasks for building robust foundation models for sleep data is a promising future research direction. Moreover, testing the effect of interventions for cognitive decline on brain age gap should also be a focus. Overall, the potential of foundation models in the medical field is demonstrated, with the example of sleep electrophysiology as an input modality. Additionally, brain age gap shows promise as a biomarker for early detection of cognitive impairment and monitoring of targeted interventions in the future.

Bibliography

Adra, N., Dümmer, L. W., Paixao, L., Tesh, R. A., Sun, H., Ganglberger, W., Westmeijer, M., Da Silva Cardoso, M., Kumar, A., Ye, E., Henry, J., Cash, S. S., Kitchener, E., Leveroni, C. L., Au, R., Rosand, J., Salinas, J., Lam, A. D., Thomas, R. J., & Westover, M. B. (2023). Decoding information about cognitive health from the brainwaves of sleep. *Scientific Reports*, *13*(1), Article 11448. https://doi.org/10.1038/s41598-023-37128-7.

Agarwal, R., & Gotman, J. (2001). Computer-assisted sleep staging. *IEEE Transactions on Biomedical Engineering*, *48*(12), 1412–1423. https://doi.org/10.1109/10.966600.

Al Zoubi, O., Ki Wong, C., Kuplicki, R. T., Yeh, H.-w., Mayeli, A., Refai, H., Paulus, M., & Bodurka, J. (2018). Predicting age from brain EEG signals—a machine learning approach. *Frontiers in Aging Neuroscience*, *10*, Article 184. https://doi.org/10.3389/fnagi.2018.00184.

Albelwi, S. (2022). Survey on self-supervised learning: Auxiliary pretext tasks and contrastive learning methods in imaging. *Entropy*, *24*(4), Article 551. https://doi.org/10.3390/e24040551.

Amjad, H., Roth, D. L., Sheehan, O. C., Lyketsos, C. G., Wolff, J. L., & Samus, Q. M. (2018). Underdiagnosis of dementia: An observational study of patterns in diagnosis and awareness in US older adults. *Journal of General Internal Medicine*, *33*(7), 1131–1138. https://doi.org/10.1007/s11606-018-4377-y.

Anderer, P., Ross, M., Cerny, A., & Shaw, E. (2022). Automated scoring of sleep and associated events. In T. Penzel & R. Hornero (Eds.), *Advances in the diagnosis and treatment of sleep apnea: Filling the gap between physicians and engineers* (pp. 107–130, Vol. 1384). Springer. https://doi.org/10.1007/978-3-031-06413-5_7.

Andersen, C. K., Wittrup-Jensen, K. U., Lolk, A., Andersen, K., & Kragh-Sørensen, P. (2004). Ability to perform activities of daily living is the main factor affecting quality of life in patients with dementia. *Health and Quality of Life Outcomes*, *2*, 52. https://doi.org/10.1186/1477-7525-2-52.

Anderton, B. H. (2002). Ageing of the brain. *Mechanisms of Ageing and Development*, *123*(7), 811–817. https://doi.org/10.1016/S0047-6374(01)00426-2.

Baecker, L., Garcia-Dias, R., Vieira, S., Scarpazza, C., & Mechelli, A. (2021). Machine learning for brain age prediction: Introduction to methods and clinical applications. *EBioMedicine*, *72*, Article 103600. https://doi.org/10.1016/j.ebiom.2021.103600.

Bah, T. M., Goodman, J., & Iliff, J. J. (2019). Sleep as a therapeutic target in the aging brain. *Neurotherapeutics*, *16*(3), 554–568. https://doi.org/10.1007/s13311-019-00769-6.

Banville, H., Jaoude, M. A., Wood, S. U., Aimone, C., Holst, S. C., Gramfort, A., & Engemann, D.-A. (2024). Do try this at home: Age prediction from sleep and meditation with large-scale low-cost mobile EEG. *Imaging Neuroscience*, *2*, 1–15. https://doi.org/10.1162/imag_a_00189.

Baranwal, N., Yu, P. K., & Siegel, N. S. (2023). Sleep physiology, pathophysiology, and sleep hygiene. *Progress in Cardiovascular Diseases*, *77*, 59–69. https://doi.org/10.1016/j.pcad.2023.02.005.

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K.-M., & Robbins, K. A. (2015). The PREP pipeline: Standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics*, *9*, Article 16. https://doi.org/10.3389/fninf.2015.00016.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer. https://www.microsoft.com/en-us/research/publication/pattern-recognition-machine-learning/.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., . . . Liang, P. (2022). On the opportunities and risks of foundation models [arXiv]. https://doi.org/10.48550/arXiv.2108.07258.

Bradford, A., Kunik, M. E., Schulz, P., Williams, S. P., & Singh, H. (2009). Missed and delayed diagnosis of dementia in primary care: Prevalence and contributing factors. *Alzheimer Disease & Associated Disorders*, *23*(4), 306–314. https://doi.org/10.1097/WAD.0b013e3181a6bebc.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324.

Brink-Kjaer, A., Leary, E. B., Sun, H., Westover, M. B., Stone, K. L., Peppard, P. E., Lane, N. E., Cawthon, P. M., Redline, S., Jennum, P., Sorensen, H. B. D., & Mignot, E. (2022). Age estimation from sleep studies using deep learning predicts life expectancy. *npj Digital Medicine*, *5*(1), Article 103. https://doi.org/10.1038/s41746-022-00630-9.

Carlson, N. R., & Birkett, M. (2021). Sleep and biological rhythms. In *Foundations of behavioral neuroscience* (10th Edition, pp. 195–221). Pearson Education Limited.

Carpi, M., Fernandes, M., Mercuri, N. B., & Liguori, C. (2024). Sleep biomarkers for predicting cognitive decline and Alzheimer's disease: A systematic review of longitudinal studies. *Journal of Alzheimer's Disease*, *97*(1), 121–143. https://doi.org/10.3233/JAD-230933.

Carrier, J., Viens, I., Poirier, G., Robillard, R., Lafortune, M., Vandewalle, G., Martin, N., Barakat, M., Paquet, J., & Filipini, D. (2011). Sleep slow wave changes during the middle years of life. *European Journal of Neuroscience*, *33*(4), 758–766. https://doi.org/10.1111/j.1460-9568.2010.07543.x.

Chang, W.-P., Liu, M.-E., Chang, W.-C., Yang, A. C., Ku, Y.-C., Pai, J.-T., Huang, H.-L., & Tsai, S.-J. (2013). Sleep apnea and the risk of dementia: A population-based 5-year follow-up study in Taiwan. *PLOS ONE*, *8*(10), Article e78655. https://doi.org/10.1371/journal.pone.0078655.

Christiansen, L., Sanmartin Berglund, J., Lindberg, C., Anderberg, P., & Skär, L. (2019). Health-related quality of life and related factors among a sample of older people with cognitive impairment. *Nursing Open*, *6*(3), 849–859. https://doi.org/10.1002/nop2.265.

Cohen, Z. L., Eigenberger, P. M., Sharkey, K. M., Conroy, M. L., & Wilkins, K. M. (2022). Insomnia and other sleep disorders in older adults. *Psychiatric Clinics of North America*, *45*(4), 717–734. https://doi.org/10.1016/j.psc.2022.07.002.

Cole, J. H., Leech, R., Sharp, D. J., & Alzheimer's Disease Neuroimaging Initiative. (2015). Prediction of brain age suggests accelerated atrophy after traumatic brain injury. *Annals of Neurology*, *77*(4), 571–581. https://doi.org/10.1002/ana.24367.

Cole, J. H., Poudel, R. P. K., Tsagkrasoulis, D., Caan, M. W. A., Steves, C., Spector, T. D., & Montana, G. (2017a). Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*, *163*, 115–124. https://doi.org/10.1016/j.neuroimage.2017.07.059.

Cole, J. H., Underwood, J., Caan, M. W., De Francesco, D., van Zoest, R. A., Leech, R., Wit, F. W., Portegies, P., Geurtsen, G. J., Schmand, B. A., Schim Van Der Loeff, M. F., Franceschi, C., Sabin, C. A., Majoie, C. B., Winston, A., Reiss, P., Sharp, D. J., & COBRA collaboration. (2017b). Increased brain-predicted aging in treated HIV disease. *Neurology*, *88*(14), 1349–1357. https://doi.org/10.1212/WNL.0000000000003790.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297. https://doi.org/10.1007/BF00994018.

Danker-Hopfe, H., Anderer, P., Zeitlhofer, J., Boeck, M., Dorn, H., Gruber, G., Heller, E., Loretz, E., Moser, D., Parapatics, S., Saletu, B., Schmidt, A., & Dorffner, G. (2009). Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *Journal of Sleep Research*, *18*(1), 74–84. https://doi.org/10.1111/j.1365-2869.2008.00700.x.

D'Atri, A., Scarpelli, S., Gorgoni, M., Truglia, I., Lauri, G., Cordone, S., Ferrara, M., Marra, C., Rossini, P. M., & De Gennaro, L. (2021). EEG alterations during wake and sleep in mild cognitive impairment and Alzheimer's disease. *iScience*, *24*(4), Article 102386. https://doi.org/10.1016/j.isci.2021.102386.

de Lange, A.-M. G., Anatürk, M., Rokicki, J., Han, L. K. M., Franke, K., Alnaes, D., Ebmeier, K. P., Draganski, B., Kaufmann, T., Westlye, L. T., Hahn, T., & Cole, J. H. (2022). Mind the gap: Performance metric evaluation in brain-age prediction. *Human Brain Mapping*, *43*(10), 3113–3129. https://doi.org/10.1002/hbm.25837.

Djonlagic, I., Mariani, S., Fitzpatrick, A. L., Van Der Klei, V. M. G. T. H., Johnson, D. A., Wood, A. C., Seeman, T., Nguyen, H. T., Prerau, M. J., Luchsinger, J. A., Dzierzewski, J. M., Rapp, S. R., Tranah, G. J., Yaffe, K., Burdick, K. E., Stone, K. L., Redline, S., & Purcell, S. M. (2021). Macro and micro sleep architecture and cognitive performance in older adults. *Nature Human Behaviour*, *5*(1), 123–145. https://doi.org/10.1038/s41562-020-00964-y.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. J., & Vapnik, V. (1997). Support vector regression machines. *Neural Information Processing Systems*, *9*, 155–161.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

Ferjani, R., Rejeb, L., & Said, L. B. (2020). Unsupervised sleep stages classification based on physiological signals. In Y. Demazeau, T. Holvoet, J. M. Corchado, & S. Costantini (Eds.), *Advances in practical applications of agents, multi-agent systems, and trustworthiness. The PAAMS collection* (pp. 134–145, Vol. 12092). Springer. https://doi.org/10.1007/978-3-030-49778-1_11.

Fiorillo, L., Puiatti, A., Papandrea, M., Ratti, P.-L., Favaro, P., Roth, C., Bargiotas, P., Bassetti, C. L., & Faraci, F. D. (2019). Automated sleep scoring: A review of the latest approaches. *Sleep Medicine Reviews*, *48*, Article 101204. https://doi.org/10.1016/j.smrv.2019.07.007.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

Franke, K., & Gaser, C. (2019). Ten years of BrainAGE as a neuroimaging biomarker of brain aging: What insights have we gained? *Frontiers in Neurology*, *10*, Article 789. https://doi.org/10.3389/fneur.2019.00789.

Franke, K., Gaser, C., & Alzheimer's Disease Neuroimaging Initiative. (2012). Longitudinal changes in individual BrainAGE in healthy aging, mild cognitive impairment, and Alzheimer's disease. *GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry*, *25*(4), 235–245. https://doi.org/10.1024/1662-9647/a000074.

Franke, K., Gaser, C., Manor, B., & Novak, V. (2013). Advanced BrainAGE in older adults with type 2 diabetes mellitus. *Frontiers in Aging Neuroscience*, *5*, Article 90. https://doi.org/10.3389/fnagi.2013.00090.

Franke, K., Ziegler, G., Klöppel, S., Gaser, C., & Alzheimer's Disease Neuroimaging Initiative. (2010). Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters. *NeuroImage*, *50*(3), 883–892. https://doi.org/10.1016/j.neuroimage.2010.01.005.

Gaiduk, M., Serrano Alarcón, Á., Seepold, R., & Martínez Madrid, N. (2023). Current status and prospects of automatic sleep stages scoring: Review. *Biomedical Engineering Letters*, *13*(3), 247–272. https://doi.org/10.1007/s13534-023-00299-3.

Ganglberger, W., Adra, N., Sun, H., Nasiri, S., Nassi, T., Landolt, H.-P., Huber, R., Thomas, R., & Westover, M. (2022). Predicting age, cognitive scores, and sleep stages from sleep EEG with a multi-task deep neural network using the Framingham Heart Study [Abstract]. *Sleep Medicine*, *100*, S35. https://doi.org/10.1016/j.sleep.2022.05.107.

Ganglberger, W., Nasiri, S., Sun, H., Kim, S., Shin, C., Westover, M. B., & Thomas, R. J. (2024). Refining sleep staging accuracy: Transfer learning coupled with scorability models. *Sleep*, *47*(11), 1–11. https://doi.org/10.1093/sleep/zsae202.

Gath, I., & Geva, A. (1989). Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*(7), 773–780. https://doi.org/10.1109/34.192473.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*(1), 1–58. https://doi.org/10.1162/neco.1992.4.1.1.

Geng, D., Wang, C., Fu, Z., Zhang, Y., Yang, K., & An, H. (2022). Sleep EEG-based approach to detect mild cognitive impairment. *Frontiers in Aging Neuroscience*, *14*, Article 865558. https://doi.org/10.3389/fnagi.2022.865558.

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., Mietus, J. E., Moody, G. B., Peng, C. K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, *101*(23), e215–e220. https://doi.org/10.1161/01.CIR.101.23.e215.

Gorgoni, M., Lauri, G., Truglia, I., Cordone, S., Sarasso, S., Scarpelli, S., Mangiaruga, A., D'Atri, A., Tempesta, D., Ferrara, M., Marra, C., Rossini, P. M., & De Gennaro, L. (2016). Parietal fast sleep spindle density decrease in Alzheimer's disease and amnesic mild cognitive impairment. *Neural Plasticity*, *2016*(1), Article 8376108. https://doi.org/10.1155/2016/8376108.

Gunnarsdottir, K. M., Gamaldo, C., Salas, R. M., Ewen, J. B., Allen, R. P., Hu, K., & Sarma, S. V. (2020). A novel sleep stage scoring system: Combining expert-based features with the generalized linear model. *Journal of Sleep Research*, *29*(5), Article e12991. https://doi.org/10.1111/jsr.12991.

Haghayegh, S., Herzog, R., Bennett, D. A., Redline, S., Yaffe, K., Stone, K. L., Ibáñez, A., & Hu, K. (2025). Predicting future risk of developing cognitive impairment using ambulatory sleep EEG: Integrating univariate analysis and multivariate information theory approach [Advance Online Publication]. *Journal of Alzheimer's Disease*. https://doi.org/10.1177/13872877251319742.

Haghayegh, S., Hu, K., Stone, K., Redline, S., & Schernhammer, E. (2023). Automated sleep stages classification using convolutional neural network from raw and time-frequency electroencephalogram signals: Systematic evaluation study. *Journal of Medical Internet Research*, *25*, Article e40211. https://doi.org/10.2196/40211.

Helfrich, R. F., Mander, B. A., Jagust, W. J., Knight, R. T., & Walker, M. P. (2018). Old brains come uncoupled in sleep: Slow wave-spindle synchrony, brain atrophy, and forgetting. *Neuron*, *97*(1), 221–230. https://doi.org/10.1016/j.neuron.2017.11.020.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634.

Hofman, W. F., & Talamini, L. M. (2015). Normal sleep and its neurophysiological regulation. In R. R. Watson (Ed.), *Modulation of sleep by obesity, diabetes, age, and diet* (pp. 25–32). Academic Press. https://doi.org/10.1016/B978-0-12-420168-2.00004-1.

Hu, M., Zhang, P., Li, C., Tan, Y., Li, G., Xu, D., & Chen, L. (2017). Sleep disturbance in mild cognitive impairment: A systematic review of objective measures. *Neurological Sciences*, *38*(8), 1363–1371. https://doi.org/10.1007/s10072-017-2975-9.

Iber, C. (2007). The AASM manual for the scoring of sleep and associated events: Rules. terminology and technical specifications.

Ibrahim, A., Cesari, M., Heidbreder, A., Defrancesco, M., Brandauer, E., Seppi, K., Kiechl, S., Högl, B., & Stefani, A. (2024). Sleep features and long-term incident neurodegeneration: A polysomnographic study. *Sleep*, *47*(3), Article zsad304. https://doi.org/10.1093/sleep/zsad304.

Jaqua, E. E., Hanna, M., Labib, W., Moore, C., & Matossian, V. (2023). Common sleep disorders affecting older adults. *The Permanente Journal*, *27*(1), 122–132. https://doi.org/10.7812/TPP/22.114.

Ju, Y.-E. S., Lucey, B. P., & Holtzman, D. M. (2014). Sleep and Alzheimer disease pathology–a bidirectional relationship. *Nature Reviews. Neurology*, *10*(2), 115–119. https://doi.org/10.1038/nrneurol.2013.269.

Kelley, B. J., & Petersen, R. C. (2007). Alzheimer's disease and mild cognitive impairment. *Neurologic Clinics*, *25*(3), 577–609. https://doi.org/10.1016/j.ncl.2007.03.008.

Kemp, B., Zwinderman, A. H., Tuk, B., Kamphuisen, H. A. C., & Oberye, J. J. L. (2000). Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, *47*(9), 1185–1194. https://doi.org/10.1109/10.867928.

Khayretdinova, M., Shovkun, A., Degtyarev, V., Kiryasov, A., Pshonkovskaya, P., & Zakharov, I. (2022). Predicting age from resting-state scalp EEG signals with deep convolutional neural networks on TD-brain dataset. *Frontiers in Aging Neuroscience*, *14*, Article 1019869. https://doi.org/10.3389/fnagi.2022.1019869.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310.

Lee, C.-H., Kim, H., Yoon, B. C., & Kim, D.-J. (2025). Toward foundational model for sleep analysis using a multimodal hybrid self-supervised learning framework [arXiv]. https://doi.org/10.48550/arXiv.2502.17481.

Lee, Y. J., Lee, J. Y., Cho, J. H., & Choi, J. H. (2022). Interrater reliability of sleep stage scoring: A meta-analysis. *Journal of Clinical Sleep Medicine*, *18*(1), 193–202. https://doi.org/10.5664/jcsm.9538.

Malafeev, A., Laptev, D., Bauer, S., Omlin, X., Wierzbicka, A., Wichniak, A., Jernajczyk, W., Riener, R., Buhmann, J., & Achermann, P. (2018). Automatic human sleep stage scoring using deep neural networks. *Frontiers in Neuroscience*, *12*, Article 781. https://doi.org/10.3389/fnins.2018.00781.

Mander, B. A., Winer, J. R., Jagust, W. J., & Walker, M. P. (2016). Sleep: A novel mechanistic pathway, biomarker, and treatment target in the pathology of Alzheimer's disease? *Trends in Neurosciences*, *39*(8), 552–566. https://doi.org/10.1016/j.tins.2016.05.002.

Mander, B. A., Winer, J. R., & Walker, M. P. (2017). Sleep and human aging. *Neuron*, *94*(1), 19–36. https://doi.org/10.1016/j.neuron.2017.02.004.

Maron, M. E. (1961). Automatic indexing: An experimental inquiry. *Journal of the ACM*, *8*(3), 404–417. https://doi.org/10.1145/321075.321084.

Masters, D., & Luschi, C. (2018). Revisiting small batch training for deep neural networks [arXiv]. https://doi.org/10.48550/arXiv.1804.07612.

Mourtazaev, M. S., Kemp, B., Zwinderman, A. H., & Kamphuisen, H. A. (1995). Age and gender affect different characteristics of slow waves in the sleep EEG. *Sleep*, *18*(7), 557–564. https://doi.org/10.1093/sleep/18.7.557.

Muehlroth, B. E., & Werkle-Bergner, M. (2020). Understanding the interplay of sleep and aging: Methodological challenges. *Psychophysiology*, *57*(3), Article e13523. https://doi.org/10.1111/psyp.13523.

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal cognitive assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, *53*(4), 695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x.

Ngo, H.-V. V., Martinetz, T., Born, J., & Mölle, M. (2013). Auditory closed-loop stimulation of the sleep slow oscillation enhances memory. *Neuron*, *78*(3), 545–553. https://doi.org/10.1016/j.neuron.2013.03.006.

Ogg, M., & Coon, W. G. (2024, May). Self-supervised transformer model training for a sleep-EEG foundation model [bioRxiv]. https://doi.org/10.1101/2024.01.18.576245.

Ohayon, M. M., Carskadon, M. A., Guilleminault, C., & Vitiello, M. V. (2004). Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: Developing normative sleep values across the human lifespan. *Sleep*, *27*(7), 1255–1273. https://doi.org/10.1093/sleep/27.7.1255.

Osorio, R. S., Pirraglia, E., Agüera-Ortiz, L. F., During, E. H., Sacks, H., Ayappa, I., Walsleben, J., Mooney, A., Hussain, A., Glodzik, L., Frangione, B., Martínez-Martín, P., & de Leon, M. J. (2011). Greater risk of Alzheimer's disease in older adults with insomnia. *Journal of the American Geriatrics Society*, *59*(3), 559–562. https://doi.org/10.1111/j.1532-5415.2010.03288.x.

Perslev, M., Darkner, S., Kempfner, L., Nikolic, M., Jennum, P. J., & Igel, C. (2021). U-Sleep: Resilient high-frequency sleep staging. *npj Digital Medicine*, *4*, Article 72. https://doi.org/10.1038/s41746-021-00440-5.

Petit, D., Gagnon, J.-F., Fantini, M. L., Ferini-Strambi, L., & Montplaisir, J. (2004). Sleep and quantitative EEG in neurodegenerative disorders. *Journal of Psychosomatic Research*, *56*(5), 487–496. https://doi.org/10.1016/j.jpsychores.2004.02.001.

Power, A., Burda, Y., Edwards, H., Babuschkin, I., & Misra, V. (2022). Grokking: Generalization beyond overfitting on small algorithmic datasets [arXiv]. https://doi.org/10.48550/arXiv.2201.02177.

Purcell, S. M., Manoach, D. S., Demanuele, C., Cade, B. E., Mariani, S., Cox, R., Panagiotaropoulou, G., Saxena, R., Pan, J. Q., Smoller, J. W., Redline, S., & Stickgold, R. (2017). Characterizing sleep spindles in 11,630 individuals from the National Sleep Research Resource. *Nature Communications*, *8*(1), Article 15930. https://doi.org/10.1038/ncomms15930.

Rando, T. A., & Chang, H. Y. (2012). Aging, rejuvenation, and epigenetic reprogramming: Resetting the aging clock. *Cell*, *148*(1), 46–57. https://doi.org/10.1016/j.cell.2012.01.003.

Rani, V., Nabi, S. T., Kumar, M., Mittal, A., & Kumar, K. (2023). Self-supervised learning: A succinct review. *Archives of Computational Methods in Engineering*, *30*(4), 2761–2775. https://doi.org/10.1007/s11831-023-09884-2.

Rasch, B., & Born, J. (2013). About sleep's role in memory. *Physiological Reviews*, *93*(2), 681–766. https://doi.org/10.1152/physrev.00032.2012.

Rechtschaffen, A., & Kales, A. (1968). *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*. US Government Printing Office.

Reppermund, S., Brodaty, H., Crawford, J. D., Kochan, N. A., Draper, B., Slavin, M. J., Trollor, J. N., & Sachdev, P. S. (2013). Impairment in instrumental activities of daily living with high cognitive demand is an early marker of mild cognitive impairment: The Sydney memory and ageing study. *Psychological Medicine*, *43*(11), 2437–2445. https://doi.org/10.1017/S003329171200308X.

Rosa, R. R., & Bonnet, M. H. (2000). Reported chronic insomnia is independent of poor sleep as measured by electroencephalography. *Psychosomatic Medicine*, *62*(4), 474–482. https://doi.org/10.1097/00006842-200007000-00004.

Russell, T., & Duntley, S. (2011). Sleep disordered breathing in the elderly. *The American Journal of Medicine*, *124*(12), 1123–1126. https://doi.org/10.1016/j.amjmed.2011.04.017.

Shahab, S., Mulsant, B. H., Levesque, M. L., Calarco, N., Nazeri, A., Wheeler, A. L., Foussias, G., Rajji, T. K., & Voineskos, A. N. (2019). Brain structure, cognition, and brain age in schizophrenia, bipolar disorder, and healthy controls. *Neuropsychopharmacology*, *44*(5), 898–906. https://doi.org/10.1038/s41386-018-0298-z.

Silber, M. H., Ancoli-Israel, S., Bonnet, M. H., Chokroverty, S., Grigg-Damberger, M. M., Hirshkowitz, M., Kapen, S., Keenan, S. A., Kryger, M. H., Penzel, T., Pressman, M. R., & Iber, C. (2007). The visual scoring of sleep in adults. *Journal of Clinical Sleep Medicine*, *3*(2), 121–131. https://doi.org/10.5664/jcsm.26814.

Stephansen, J., Olesen, A., Olsen, M., Ambati, A., Leary, E., Moore, H., Carrillo, O., Lin, L., Han, F., Yan, H., Sun, Y., Dauvilliers, Y., Scholz, S., Barateau, L., Hogl, B., Stefani, A., Hong, S., Kim, T., Pizza, F., . . . Mignot, E. (2018). Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nature Communications*, *9*(1), Article 5229. https://doi.org/10.1038/s41467-018-07229-3.

Sun, H., Ganglberger, W., Westover, M. B., & Thomas, R. J. (2023). Artificial Intelligence in Sleep Medicine. In R. J. Thomas, S. Bhat, & S. Chokroverty (Eds.), *Atlas of Sleep Medicine* (pp. 355–369). Springer. https://doi.org/10.1007/978-3-031-34625-5_21.

Sun, H., Paixao, L., Oliva, J. T., Goparaju, B., Carvalho, D. Z., van Leeuwen, K. G., Akeju, O., Thomas, R. J., Cash, S. S., Bianchi, M. T., & Westover, M. B. (2019). Brain age from the electroencephalogram of sleep. *Neurobiology of Aging*, *74*, 112–120. https://doi.org/10.1016/j.neurobiolaging.2018.10.016.

Taillard, J., Sagaspe, P., Berthomier, C., Brandewinder, M., Amieva, H., Dartigues, J.-F., Rainfray, M., Harston, S., Micoulaud-Franchi, J.-A., & Philip, P. (2019). Non-REM sleep characteristics predict early cognitive impairment in an aging population. *Frontiers in Neurology*, *10*, Article 197. https://doi.org/10.3389/fneur.2019.00197.

Thapa, R., Kjær, M. R., He, B., Covert, I., Moore, H., Hanif, U., Ganjoo, G., Westover, M. B., Jennum, P., Brink-Kjær, A., Mignot, E., & Zou, J. (2025). A multimodal sleep foundation model developed with 500K hours of sleep recordings for disease predictions [medRxiv]. https://doi.org/10.1101/2025.02.04.25321675.

Tubbs, A. S., Dollish, H. K., Fernandez, F., & Grandner, M. A. (2019). The basics of sleep physiology and behavior. In M. A. Grandner (Ed.), *Sleep and Health* (pp. 3–10). Elsevier. https://doi.org/10.1016/B978-0-12-815373-4.00001-0.

Vanni, S., Colini Baldeschi, A., Zattoni, M., & Legname, G. (2020). Brain aging: A Ianus -faced player between health and neurodegeneration. *Journal of Neuroscience Research*, *98*(2), 299–311. https://doi.org/10.1002/jnr.24379.

Voumvourakis, K. I., Sideri, E., Papadimitropoulos, G. N., Tsantzali, I., Hewlett, P., Kitsos, D., Stefanou, M., Bonakis, A., Giannopoulos, S., Tsivgoulis, G., & Paraskevas, G. P. (2023). The dynamic relationship between the glymphatic

system, aging, memory, and sleep. *Biomedicines*, *11*(8), Article 2092. https://doi.org/10.3390/biomedicines11082092.

Wennberg, A. M. V., Wu, M. N., Rosenberg, P. B., & Spira, A. P. (2017). Sleep disturbance, cognitive decline, and dementia: A review. *Seminars in Neurology*, *37*(4), 395–406. https://doi.org/10.1055/s-0037-1604351.

Westerberg, C. E., Mander, B. A., Florczak, S. M., Weintraub, S., Mesulam, M.-M., Zee, P. C., & Paller, K. A. (2012). Concurrent impairments in sleep and memory in amnestic mild cognitive impairment. *Journal of the International Neuropsychological Society*, *18*(3), 490–500. https://doi.org/10.1017/S135561771200001X.

World Health Organization. (2015). *World report on ageing and health* (tech. rep.). World Health Organization. Geneva. Retrieved July 2, 2025, from https://www.who.int/publications/i/item/9789241565042.

Wunderlin, M., Zeller, C. J., Wicki, K., Nissen, C., & Züst, M. A. (2024). Acoustic stimulation during slow wave sleep shows delayed effects on memory performance in older adults. *Frontiers in Sleep*, *2*, Article 1294957. https://doi.org/10.3389/frsle.2023.1294957.

Wunderlin, M., Zeller, C. J., Senti, S. R., Fehér, K. D., Suppiger, D., Wyss, P., Koenig, T., Teunissen, C. E., Nissen, C., Klöppel, S., & Züst, M. A. (2023). Acoustic stimulation during sleep predicts long-lasting increases in memory performance and beneficial amyloid response in older adults. *Age and Ageing*, *52*(12), Article afad228. https://doi.org/10.1093/ageing/afad228.

Wunderlin, M., Züst, M. A., Hertenstein, E., Fehér, K. D., Schneider, C. L., Klöppel, S., & Nissen, C. (2021). Modulating overnight memory consolidation by acoustic stimulation during slow-wave sleep: A systematic review and meta-analysis. *Sleep*, *44*(7), Article zsaa296. https://doi.org/10.1093/sleep/zsaa296.

Wunderlin, M., Züst, M. A., Fehér, K. D., Klöppel, S., & Nissen, C. (2020). The role of slow wave sleep in the development of dementia and its potential for preventative interventions. *Psychiatry Research: Neuroimaging*, *306*, 111178. https://doi.org/10.1016/j.pscychresns.2020.111178.

Yaffe, K., Laffan, A. M., Harrison, S. L., Redline, S., Spira, A. P., Ensrud, K. E., Ancoli-Israel, S., & Stone, K. L. (2011). Sleep-disordered breathing, hypoxia, and risk of mild cognitive impairment and dementia in older women. *JAMA*, *306*(6), 613–619. https://doi.org/10.1001/jama.2011.1115.

Yaremchuk, K. (2018). Sleep disorders in the elderly. *Clinics in Geriatric Medicine*, *34*(2), 205–216. https://doi.org/10.1016/j.cger.2018.01.008.

Ye, E., Sun, H., Krishnamurthy, P. V., Adra, N., Ganglberger, W., Thomas, R. J., Lam, A. D., & Westover, M. B. (2023). Dementia detection from brain activity during sleep. *Sleep*, *46*(3), Article zsac286. https://doi.org/10.1093/sleep/zsac286.

Ye, E., Sun, H., Leone, M. J., Paixao, L., Thomas, R. J., Lam, A. D., & Westover, M. B. (2020). Association of sleep electroencephalography-based brain age index with dementia. *JAMA Network Open*, *3*(9), Article e2017357. https://doi.org/10.1001/jamanetworkopen.2020.17357.

Yook, S., Park, H. R., Park, C., Park, G., Lim, D. C., Kim, J., Joo, E. Y., & Kim, H. (2022). Novel neuroelectrophysiological age index associated with imaging features of brain aging and sleep disorders. *NeuroImage*, *264*, Article 119753. https://doi.org/10.1016/j.neuroimage.2022.119753.

Zeller, C. J., Wunderlin, M., Wicki, K., Teunissen, C. E., Nissen, C., Züst, M. A., & Klöppel, S. (2024). Multi-night acoustic stimulation is associated with better sleep, amyloid dynamics, and memory in older adults with cognitive impairment. *GeroScience*, *46*(6), 6157–6172. https://doi.org/10.1007/s11357-024-01195-z.

Zhang, D., She, Y., Sun, J., Cui, Y., Yang, X., Zeng, X., & Qin, W. (2024). Brain age estimation from overnight sleep electroencephalography with multi-flow sequence learning. *Nature and Science of Sleep*, *16*, 879–896. https://doi.org/10.2147/NSS.S463495.

Zhang, X., Xu, M., Li, Y., Su, M., Xu, Z., Wang, C., Kang, D., Li, H., Mu, X., Ding, X., Xu, W., Wang, X., & Han, D. (2020). Automated multi-model deep neural network for sleep stage scoring with unfiltered clinical data. *Sleep and Breathing*, *24*(2), 581–590. https://doi.org/10.1007/s11325-019-02008-w.

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., Peng, H., Li, J., Wu, J., Liu, Z., Xie, P., Xiong, C., Pei, J., Yu, P. S., & Sun, L. (2024). A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT. *International Journal of Machine Learning and Cybernetics*. https://doi.org/10.1007/s13042-024-02443-6.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, *109*, 43–76. https://doi.org/10.1109/JPROC.2020.3004555.

Appendix

A Software and Code Availability

All analyses were performed using Python 3.6.9 in Visual Studio Code. The most important Python libraries used in this thesis include:

- **Keras 2.6.0** (https://keras.io)
- **Matplotlib 3.3.4** (https://matplotlib.org)
- **MNE 0.23.4** (https://mne.tools)
- **NumPy 1.19.5** (https://numpy.org)
- **Pandas 1.1.5** (https://pandas.pydata.org)
- **scikit-learn 0.24.2** (https://scikit-learn.org)
- **SciPy 1.5.4** (https://scipy.org)
- **seaborn 0.11.2** (https://seaborn.pydata.org)
- **TensorFlow 2.6.2** (https://www.tensorflow.org)

The complete list of dependencies and versions is available in the `requirements.txt` file in the project repository.

All code used for model development and analyses in this thesis including descriptions can be found in the following repository:
https://github.com/portmannh/brain-age-from-sleep.git

B Pretext Task Model Architecture

Table B.1: Overview of the sleep staging model's layers and hyperparameters, including output dimensions during training. The input data consists of 17.5-minute recordings sampled at 100 Hz, resulting in a time dimensionality of 105,000 (17.5 x 60 x 100).

#	Layer Type	Output Dimensions (training)	Kernel	Filters	Activation
0	Input (no actual layer)	105000 x 4	-	-	-
1	Encoder Block	52500 x 6	9	6	elu
2	Encoder Block	26250 x 9	9	9	elu
3	Encoder Block	13126 x 11	9	11	elu
4	Encoder Block	6564 x 15	9	15	elu
5	Encoder Block	3282 x 20	9	20	elu
6	Encoder Block	1642 x 28	9	28	elu
7	Encoder Block	822 x 40	9	40	elu
8	Encoder Block	412 x 55	9	55	elu
9	Encoder Block	206 x 77	9	77	elu
10	Encoder Block	104 x 108	9	108	elu
11	Encoder Block	52 x 152	9	152	elu
12	Encoder Block	26 x 214	9	214	elu
13	Bottle Neck	26 x 306	9	306	elu
14	Decoder Block	52 x 428	9	214	elu
15	Decoder Block	104 x 304	9	152	elu
16	Decoder Block	206 x 216	9	108	elu
17	Decoder Block	412 x 154	9	77	elu
18	Decoder Block	822 x 110	9	55	elu
19	Decoder Block	1642 x 80	9	40	elu
20	Decoder Block	3282 x 56	9	28	elu
21	Decoder Block	6564 x 40	9	20	elu
22	Decoder Block	13126 x 30	9	15	elu
23	Decoder Block	26250 x 22	9	11	elu
24	Decoder Block	52500 x 18	9	9	elu
25	Decoder Block	105000 x 12	9	6	elu
26	Convolution	105000 x 6	1	6	tanh
27	Convolution	105000 x 5	1	5	elu
28	Convolution	105000 x 5	1	5	linear

C Substudies during Pretext Task Development

During the development of the pretext task, the sleep staging network, three substudies were conducted using the SC dataset.

C.1 Number of Epochs

To determine whether training the model for 500 epochs was sufficient, the model was once trained for 10000 epochs. This resulted in the following curves of loss and accuracy (Figure C.1).

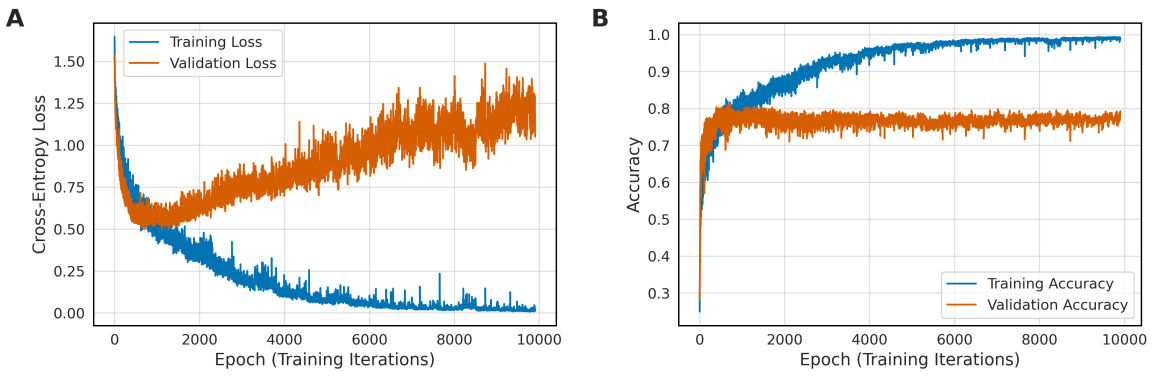

Figure C.1: Training and validation loss (A) and accuracy (B) plotted over 10000 training epochs.

It can be observed that the validation loss stops decreasing and even starts increasing again on average after approximately 700 epochs. Subsequently, only the training loss continues to decrease. The opposite can be observed in the plot showing accuracy. While the training accuracy continues to increase slightly, the validation accuracy remains constant after approximately 700 epochs.

The explanation for why the validation loss increases again while the validation accuracy remains the same can be found in the way that loss and accuracy are computed. The categorical cross-entropy loss function considers the predicted probabilities of all labels, whereas accuracy is based solely on the label with the highest probability. This leads to the conclusion that the model, on average, still predicts similar labels, maintaining the same level of accuracy. However, the model becomes overconfident, and when it predicts the wrong class, the probability of that class is much higher than at earlier epochs, causing the loss to increase again.

Power et al. (2022) described a phenomenon they called grokking, where generalization of a model occurs long after overfitting to the training set. This phenomenon could not be observed when training our model for 10000 epochs. However, the possibility that it could occur when training for even longer cannot be ruled out. This was not done at this point due to the limited timeframe of the Master thesis.

These results indicate that training for 1000 epochs or more is not necessary and may lead to the model overfitting to the training data. Therefore, the performance after 500 epochs of training is expected to be good and was subsequently maintained.

C.2 Batch Size

Choosing a suitable batch size can significantly impact the training of a model. Batch size 4 was chosen due to memory storage reasons, as initially, samples longer than 17.5 minutes were used. However, with the final training sample duration of 17.5 minutes, larger batch sizes would also be possible.

Both larger and smaller batch sizes have demonstrated their advantages and disadvantages in the past, resulting in a trade-off in the choice of batch size. Large batch sizes can make efficient use of parallel computing and speed up training, while smaller batch sizes can lead to more reliable and stable results with higher generalization performance and faster optimization convergence (Masters & Luschi, 2018). As training time was reasonable, a smaller batch size might, therefore, actually lead to better results in this case.

However, it was still investigated what would happen if the batch size were set to 64, which is the batch size used in the original U-Sleep (Perslev et al., 2021). The loss and accuracy curves are shown in Figure C.2. It can be observed that the final training and validation accuracies were between 0.7 and 0.8, similar to or even a bit lower than with batch size 4. The loss with batch size 64 was higher after 500 epochs than with batch size 4. In conclusion, it was decided to keep the batch size at 4.

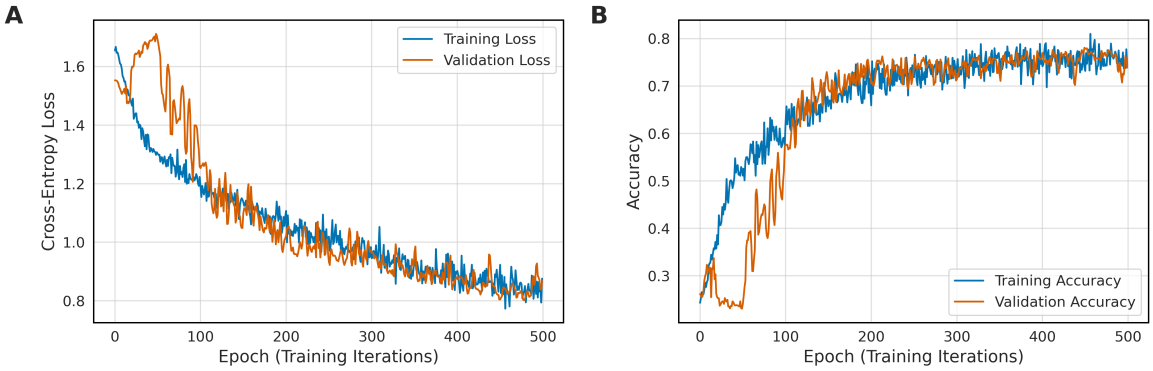

Figure C.2: Training and validation loss (A) and accuracy (B) curves during training with a batch size of 64.

C.3 Validation vs. Training Set Performance

In the original learning curves (Figure E.1), loss and accuracy seem to be slightly better on average for the validation set than the training set at most time points. To investigate why this is the case, the training process was repeated with random sampling of the training set instead of the balanced sampling described in the Methods section and without setting sample weights.

These were the processing steps initially applied to the training set but not to the validation set. This could have been a possible explanation for the observed phenomenon, because the validation set would be expected to perform better if it contained more of the overrepresented N2 that the model is good at predicting correctly, whereas in the training set, more of the classes that cannot be predicted as well, such as N1, would be present. This training session produced the learning curves shown in Figure C.3. It can be observed that the performance of the validation set still seems to be slightly better than that of the training set. The phenomenon, therefore, cannot simply be explained by balanced sampling and sample weights, and the reason for this remains unclear. It cannot be excluded that a bug is responsible for this. However, this could not explain why this phenomenon is not observed in the substudy C.2 when the batch size is increased to 64.

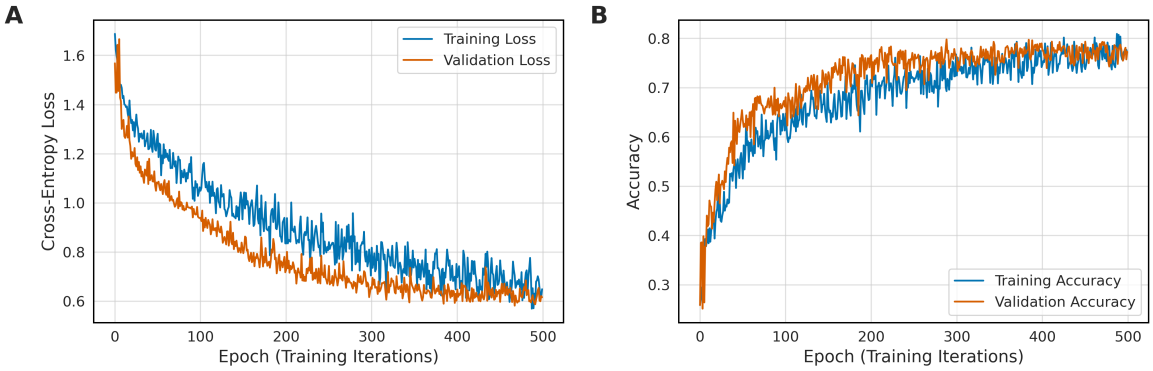

Figure C.3: Training and validation loss (A) and accuracy (B) curves when both sets were prepared using the validation preprocessing pipeline.

D Scatterplots of Correlational Analyses

D.1 SC Dataset

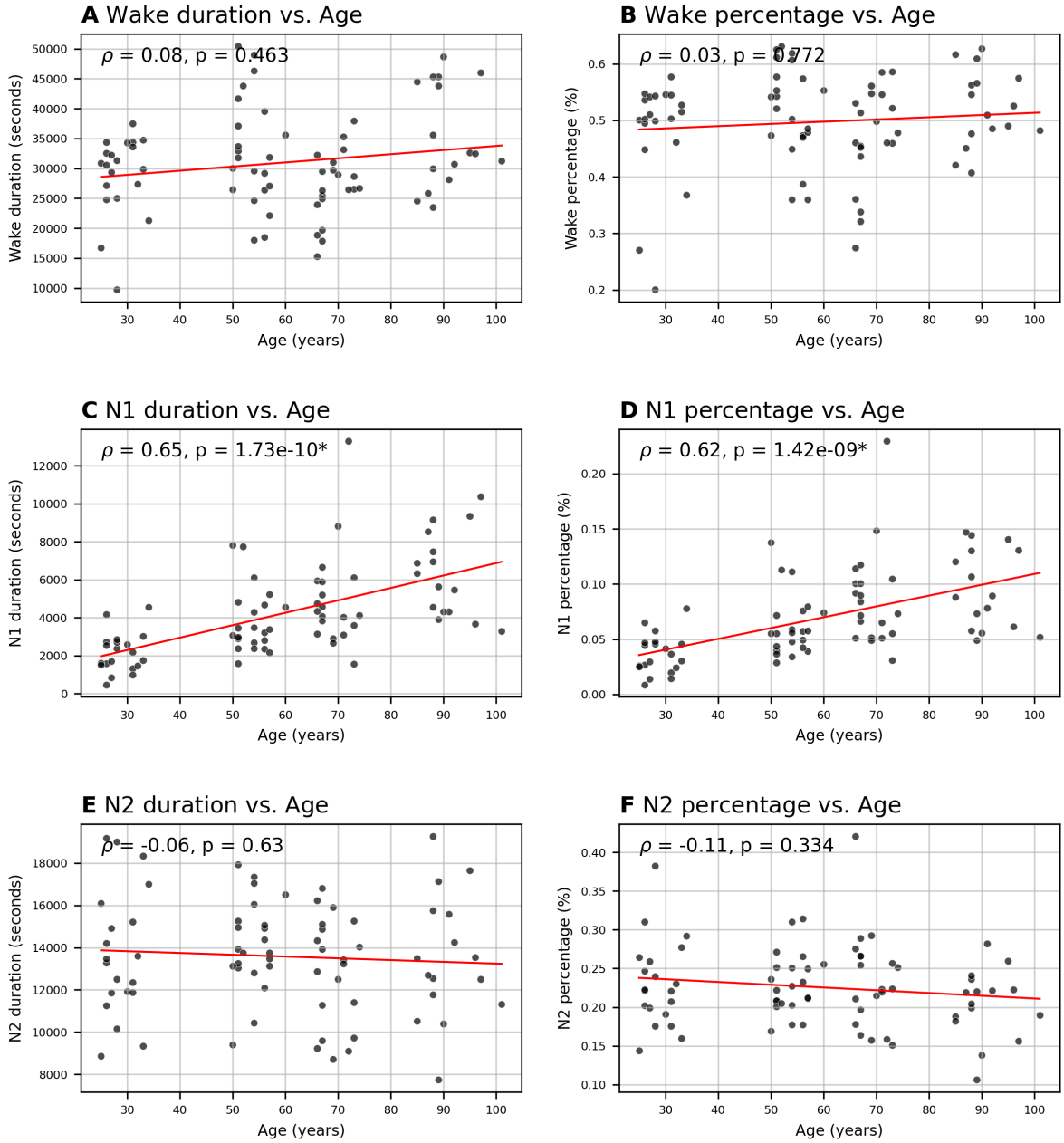

Figure D.1: Scatterplots showing the correlations between sleep stage durations (in minutes) and percentages (%) with age in the SC dataset (Part 1). Spearmans' rank correlation coefficient (ρ) and corresponding p-values are reported. An asterisk (*) indicates significance after Bonferroni correction. A linear regression line is shown for visualization.

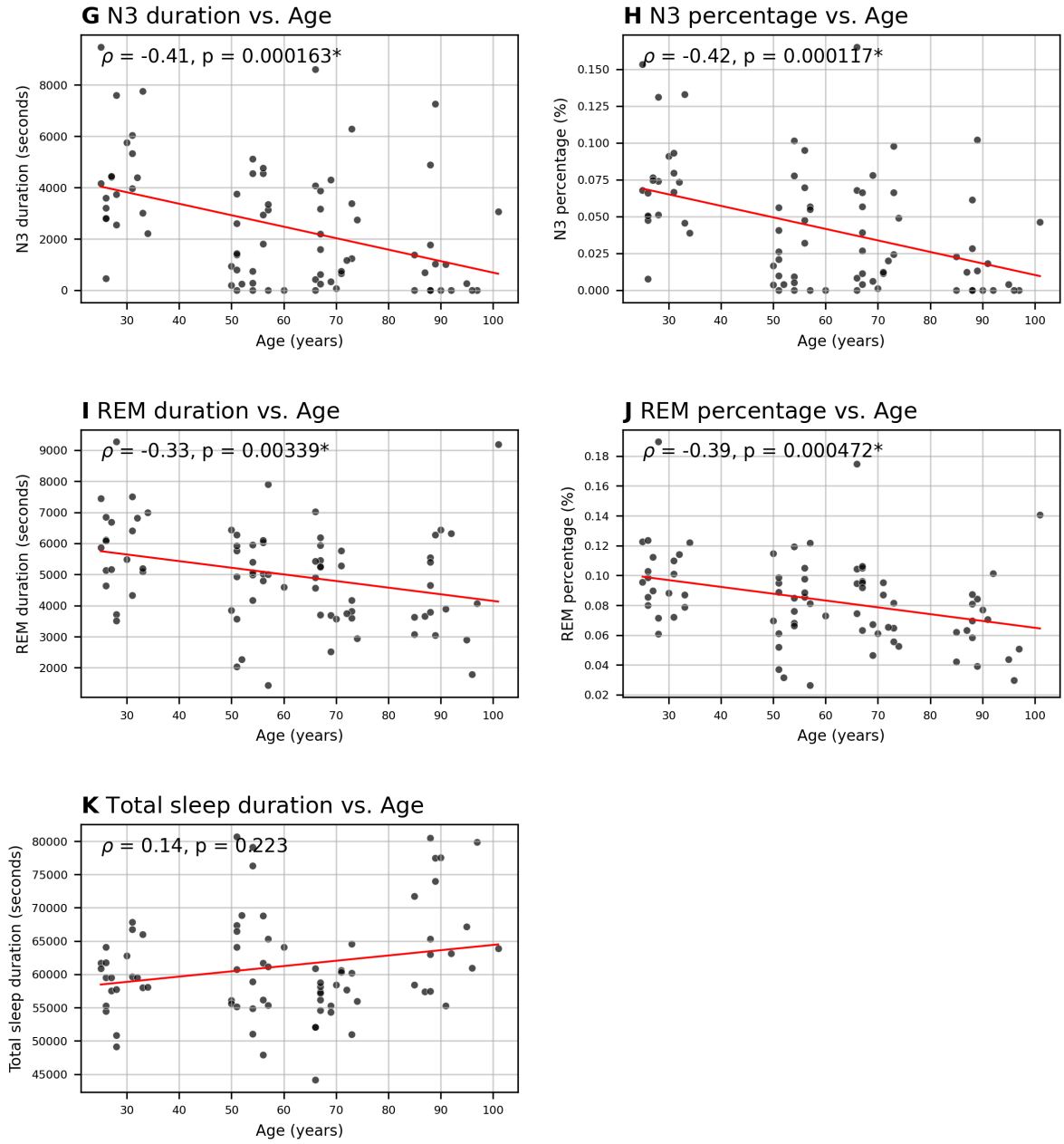

Figure D.2: Scatterplots showing the correlations between sleep stage durations (in minutes) and percentages (%) with age in the SC dataset (Part 2). Spearmans' rank correlation coefficient (ρ) and corresponding p-values are reported. An asterisk (*) indicates significance after Bonferroni correction. A linear regression line is shown for visualization.

D.2 UPD Dataset

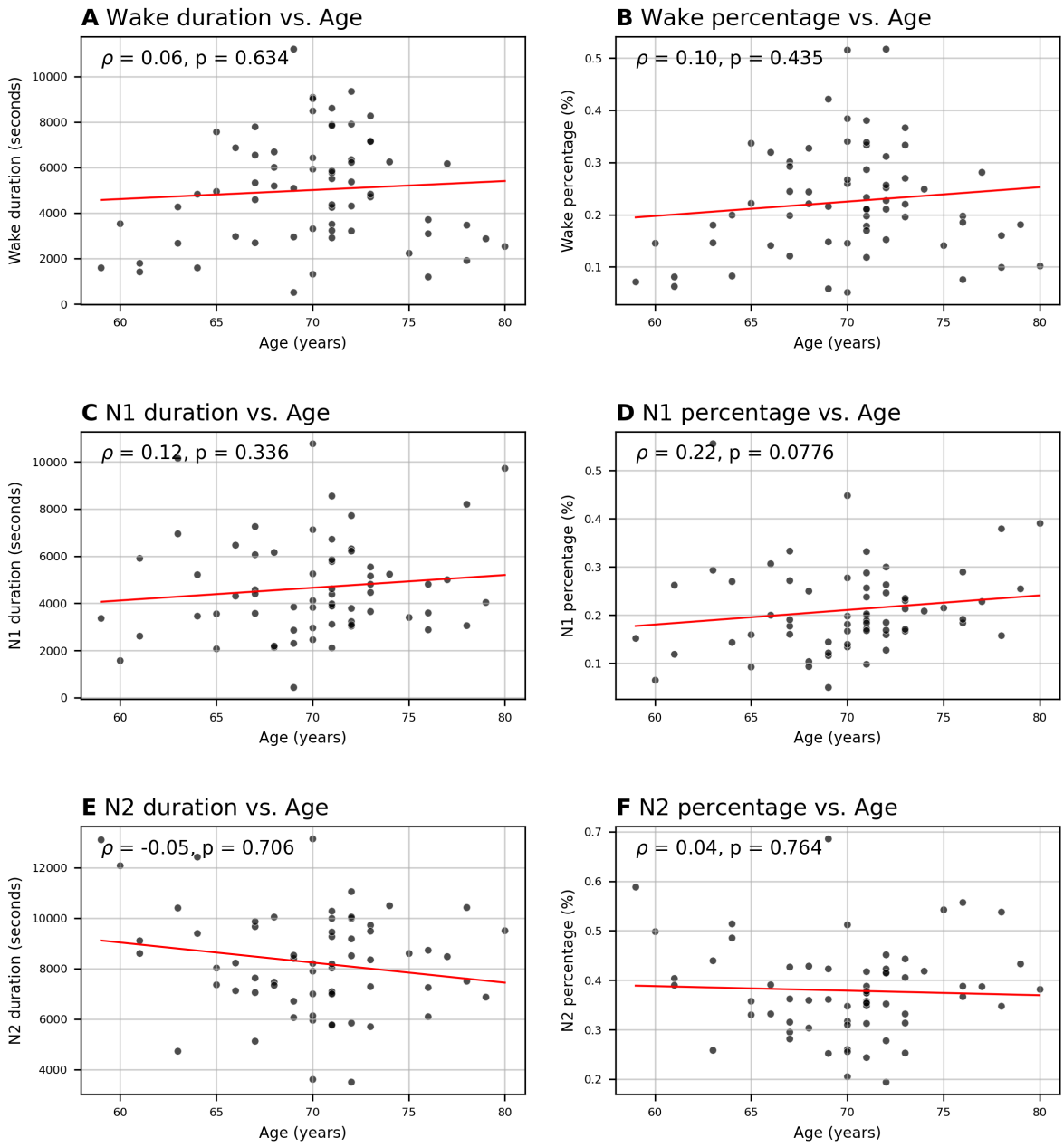

Figure D.3: Scatterplots showing the correlations between sleep stage durations (in minutes) and percentages (%) with age in the UPD dataset (Part 1). Spearmans' rank correlation coefficient (ρ) and corresponding p-values are reported. None of the results were statistically significant after Bonferroni correction. A linear regression line is shown for visualization.

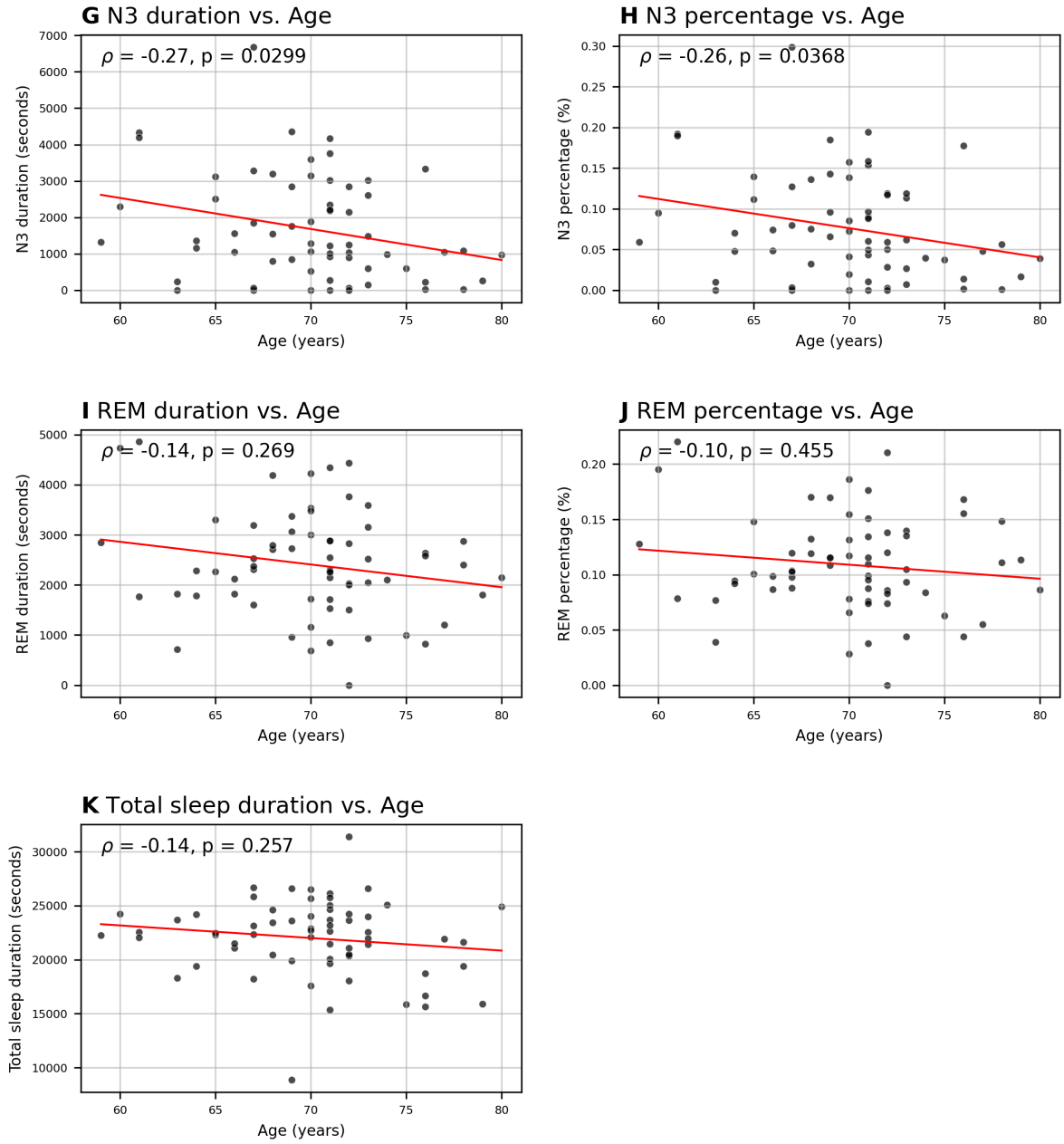

Figure D.4: Scatterplots showing the correlations between sleep stage durations (in minutes) and percentages (%) with age in the UPD dataset (Part 2). Spearmans' rank correlation coefficient (ρ) and corresponding p-values are reported. None of the results were statistically significant after Bonferroni correction. A linear regression line is shown for visualization.

E Training Curves for the Pretext Task

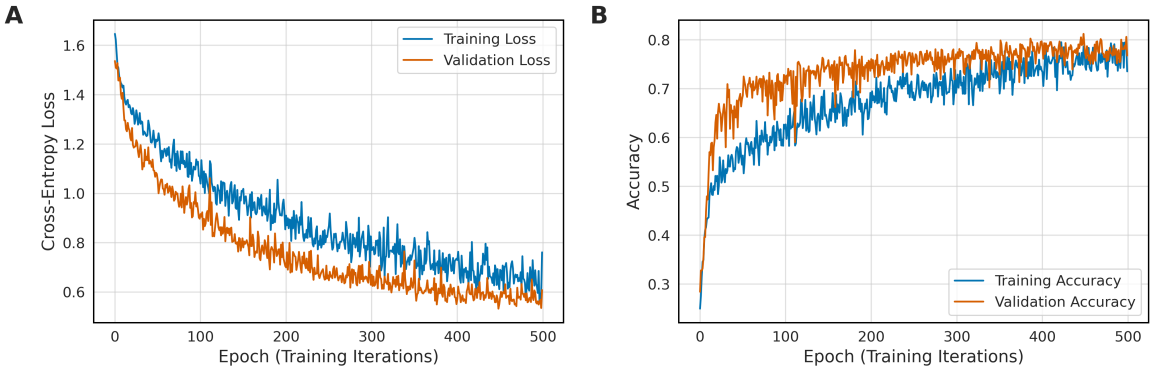

Figure E.1: Training and validation loss (A) and accuracy (B) curves on the SC dataset over 500 epochs of training.

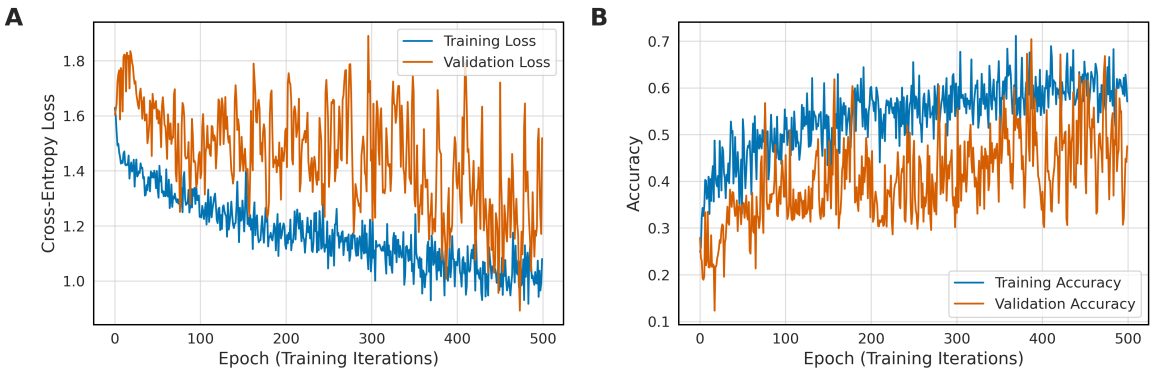

Figure E.2: Training and validation loss (A) and accuracy (B) curves on the UPD dataset over 500 epochs of training.

F AI Utilization

The following points describe the artificial intelligence (AI) tools used during the development of this thesis and their specific purposes. The author takes full responsibility for all content in this thesis.

- **ChatGPT (GPT-4-turbo)**:
  - Finding and correcting bugs or errors in Python scripts.
  - Adapting and implementing segments of code.
  - Enhancing code structure and efficiency.
  - Rephrasing and refining individual sentences in the written thesis.
  - Summarizing selected research papers for relevance assessment.
  - Improving the structure and clarity of specific sections in the written thesis to improve logical flow.
- **GitHub Copilot**:
  - Finding and correcting bugs or errors in Python scripts.
  - Code auto-completion, mainly in the case of repetitions in Python scripts.
- **NotebookLM**:
  - Finding specific information in long or multiple research papers.
- **Grammarly**:
  - Checking grammar and spelling throughout the written thesis.
  - Enhancing clarity and reducing redundancy by accepting suggestions of selected alternative word choices.