Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Engineering Cybernetics

Karoline Seljevoll Herleiksplass

Enhancing Sleep-Wake Detection Using Deep Learning and Optimal Channel Selection from High-Density EEG

Master's thesis in Department of Engineering Cybernetics
Supervisor: Marta Molinas
Co-supervisor: Luis Alfredo Moctezuma
June 2023

NTNU
Norwegian University of
Science and Technology

Karoline Seljevoll Herleiksplass

Enhancing Sleep-Wake Detection Using Deep Learning and Optimal Channel Selection from High-Density EEG

Master's thesis in Department of Engineering Cybernetics
Supervisor: Marta Molinas
Co-supervisor: Luis Alfredo Moctezuma
June 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Engineering Cybernetics

NTNU
Norwegian University of
Science and Technology

Preface

This master's thesis completes my Master of Technology at the Norwegian University of Science and Technology (NTNU) in the Department of Engineering Cybernetics, written during the spring semester of 2023.

The idea of the project was provided by Marta Molinas, acting as the supervisor for this work. The work was initiated with a specialization project [1] conducted during the fall semester of 2022, titled: Binary Classification for Sleep Stages with Deep Learning, EEGNet. The software utilized is available upon request at karolineherleiksplass@gmail.com and on GitHub: https://github.com/wavesresearch

The anticipated reader of this thesis is assumed to have graduate-level knowledge in Cybernetics and Robotics. While this work may also engage those from diverse engineering backgrounds, a foundational understanding of machine learning is strongly recommended for a comprehensive interpretation of the content herein.

In an attempt to contribute to the wider academic community, the findings from this research have been submitted to the 16Th International Conference on Brain Informatics (BI 2023). The submitted paper is attached in Appendix B, *"Automatic sleep-wake scoring with optimally selected EEG channels from high-density EEG"*.

Trondheim, June 5, 2023

Karoline Seljevoll Herleiksplass

Acknowledgement

The journey through the realm of sleep research has been incredibly fascinating, illuminating areas of science I had yet to explore. The field's intriguing intricacies and the immense potential for the future have turned this topic into a deep-seated interest for me.

I am deeply grateful to my supervisor, Professor Marta Molinas, for providing me with the opportunity to delve into sleep research, a field that we both find captivating. Her guidance, insightful suggestions, and in-depth discussions have significantly enriched my understanding of this area.

Additionally, I wish to extend my appreciation to my co-supervisor, Dr Luis Alfredo Moctezuma. His readiness to assist, provide direction, and collaborate on the BI 2023 submission has been an invaluable part of this journey.

I also wish to express my gratitude to the Human Sleep Laboratory of the International Institute of Integrative Sleep Medicine (University of Tsukuba, Japan) for providing me with such a unique dataset. The ability to work with this resource has greatly contributed to the depth of this research.

My sincere thanks go to NTNU for granting access to the computing resources of the High-Performance Computing system IDUN. This access was important for the possibility of carrying out multi-objective optimization and managing the extensive volume of data involved in this research.

Lastly, I express my gratitude to everyone who has contributed to this research journey. Their thoughtful discussions and insights have been incredibly enlightening and inspiring. I am truly privileged to have had the opportunity to learn from such insightful, intelligent, and reflective individuals.

Abstract

This master thesis explores the challenging yet important task of sleep-wake classification using electroencephalography (EEG) data, which has critical implications in both clinical and at-home settings. Given the rising prevalence of sleep disorders, the development of efficient and accessible sleep-wake classification methodologies can significantly contribute to improving public health and well-being.

EEG is a neuroimaging technique that measures brain activity by detecting electrical signals in the brain. EEG is vital in examining and transforming bioelectric brain signals into commands controlling external devices in a brain-computer interface context.

This thesis presents a comprehensive exploration of EEG-based sleep-wake classification, specifically focusing on developing and evaluating a multi-objective optimization algorithm that maximizes model performance while minimizing the number of EEG channels used for classification. In addition, this optimization provides insights into the importance of specific channels for wake detection and sleep-wake classification.

The dataset used consists of raw, unprocessed high-density EEG data collected at the Human Sleep Laboratory of the International Institute of Integrative Sleep Medicine, University of Tsukuba, Japan. The sleep stages in the dataset are annotated by professional sleep technologists. Such a dataset uniquely embraces the use of raw, unprocessed datasets, emulating real-world data and presenting an authentic classification challenge despite the imbalance in the dataset.

This research adopts machine learning using the EEGNet architecture for deep-learning neural network models in a subject-by-subject approach. A Non-dominated Sorting Genetic Algorithm is used to find a reduced and optimized subset of channels. A key aspect is investigating whether such a subset can achieve comparable performance in sleep-wake classification as high as high-density EEG.

The thesis argues that the proposed methodology can significantly contribute to real-time, automatic sleep stage classification. In addition to utilizing raw, unprocessed, and unbalanced data, the reduced subset of channels achieves comparable results as using high-density data.

This research offers potential pathways for future development in sleep research, specifically in developing efficient, accurate, and real-world applicable automated sleep stage classification systems.

Sammendrag

Denne masteroppgaven utforsker den utfordrende men viktige oppgaven med å klassifisere våkenhet fra søvn ved bruk av elektroencefalografi (EEG) data, som er viktig i både kliniske og hjemmemiljøer. Ettersom søvnforstyrrelser er et stadig voksende problem, kan effektiv og tilgjengelig søvn-våkenhet klassifisering bidra til å bedre folkehelsen og menneskets personlige velvære.

EEG er en nevroavbildningsteknikk som måler hjerneaktivitet ved å oppdage elektriske signaler i hjernen. I et hjerne-datamaskin-grensesnitt er EEG avgjørende for å tolke og transformere bioelektriske hjernesignaler til kommandoer for å kunne kontrollere eksterne enheter.

Denne oppgaven presenterer en omfattende studie av EEG-basert søvn-våkenhetsklassifisering. Fokuset ligger spesielt på å utvikle og evaluere en multi-objektiv optimaliseringsalgoritme som maksimerer modellytelsen samtidig som antall EEG kanaler brukt i klassifiseringen blir minimert. I tillegg gir denne optimaliseringen innsikt i viktigheten av spesifikke kanaler for våkenhetsdeteksjon og søvn-våkenhetsklassifisering.

Datasettet som brukes består av rå, ubehandlet EEG-data med høy tetthet samlet inn ved Human Sleep Laboratoriet ved International Institute of Integrative Sleep Medicine, Univeristy of Tsukuba, Japan. Søvnstadiene i datasettet er klassifisert av profesjonelle søvnteknologier. Dette arbeidet omfavner en unik bruk av rå, ubehandlet datasett som representerer virkelig data og presenterer en autentisk klassifiseringsutfordring til tross for den iboende ubalansen i datasettet.

Denne forskningen tar i bruk maskinlæring ved å bruke EEGNet-arkitekturen for dyp læring av nevrale nettverksmodeller i en pasient-for-pasient-metode. En ikke-dominerende sorteringsgenetisk algoritme brukes for å finne et redusert og optimalisert del-sett av kanaler. Et sentralt aspekt er å undersøke om slike del-sett kan oppnå tilsvarende ytelse i søvn-våkenhetsklassifisering slik som EEG data med høy tetthet.

Masteroppgaven argumenterer at den foreslåtte metodikken i betydelig grad kan bidra mot automatisk klassifisering av søvnstadier i sanntid. I tillegg til å bruke rå, ubehandlet og ubalansert data, oppnår de reduserte del-settene av kanaler sammenlignbare resultater slik som ved bruk av høy-tetthets data.

Denne forskningen tilbyr mulige veier for fremtidig utvikling innen søvnforskning, spesielt mot å utvikle effektive, nøyaktige og virkelighetsaktuelle automatiserte klassifiseringssystemer for søvnstadier.

Contents

List of Figures

List of Tables

List of Abbreviations

WPI-IIIS International Institute of Integrative Sleep Medicine

AASM American Academy of Sleep Medicine

EEG Electroencephalogram

EMG Electromyogram

EOG Electrooculograms

ECG Electrocardiogram

CNS Central Nervous System

SCN Suprachiasmatic Nucleus

NREM Non-Rapid Eye Movement

REM Rapid Eye Movement

PSG Polysomnography

WASO Wake After Sleep Onset

SOL Sleep Onset Latency

SWS Slow Wave Sleep

LAMF Low-Amplitude Mixed Frequency

OSA Obstructive Sleep Apnea

ASD Autism Spectrum Disorder

CRSD Circadian Rhythm Sleep Disorder

BCI Brain–Computer Interface

ASSC Automatic Sleep Stage Classification

ASEEGA Automatic Sleep Scoring Software

AUROC Area Under the Receiver Operating Characteristic Curve

CNN Convolutional Neural Network

RNN Recurrent Neural Network

SVM State Vector Machine

ANN Artificial Neural Network

ML Machine Learning

AI Artificial Intelligence

NSGA Non-Dominated Sorting Genetic Algorithm

GA Genetic Algorithm

MOGA Multi-Objective Genetic Algorithm

BiLSTM Bidirectional Long Short-Term Memory

DWT Discrete Wavelet Transform

SCSP Sparse Common Spatial Pattern

PVDF Thin Polyvinylidene Fluoride

ACT Actigraphy

Chapter 1

Introduction

This chapter presents the motivation driving the research undertaken, followed by the objectives and posed research questions. Subsequently, the anticipated contributions of this thesis to the research field of Automatic Sleep Stage Classification (ASSC) are highlighted. Lastly, an organized overview of the subsequent chapters within this thesis is provided.

1.1 Motivation

In today's rapidly evolving, technologically driven world, there is a crucial yet frequently neglected element of our daily lives that requires greater attention: sleep [2]. Far from being just a period of inactivity or merely a dormant state, sleep serves as a fundamental physiological process. Sleep has an indispensable role in influencing both physical health and mental well-being.

Approximately one-third of the human lifetime is spent asleep, and ensuring high-quality sleep is critical for human functioning. Contemporary society poses unique challenges to maintaining healthy sleep patterns with its continuous distractions, interruptions, and demands. The consequent increase in sleep disorders and disturbances has profound implications for human health, given the association with a broad spectrum of physical and psychological health issues [3, 4, 5].

The increasing prevalence of sleep-related health problems and sleep diseases represents a considerable public health concern. It emphasizes the need for a deeper understanding of the complexity of sleep physiology and the development of effective strategies for diagnosing and managing sleep disorders. Recognizing sleep as a crucial factor influencing human health can possibly pave the way for enhanced public health improvements [3, 4, 5].

Sleep interruptions can be caused by various factors and result in nighttime awakenings. Frequent occurrences of nighttime awakenings might signify a risk of developing sleep disorders. Insomnia is the most known sleep disorder where the patient suffers from difficulties sleeping through the night and has trouble getting back to sleep after unwanted awakenings. It affects nearly one-third of the population, but less than twenty percent of patients with insomnia receive an accurate diagnosis and appropriate treatment [5, 6, 7].

Insufficient sleep can have a significant impact on human health, leading to an increased risk of

chronic conditions such as obesity, diabetes, and cancer. Anomalies in sleep patterns have also been associated with neurological and psychological conditions like autism and dementia. If carefully monitored and examined, variations in sleep patterns could serve as a potential early warning system, signaling the onset of these conditions. Tracking sleep patterns and identifying any irregularities are vital for both the prevention and early detection of a range of health conditions. This highlights the necessity of accurate identification and measurement of wake periods within sleep [8, 9, 10, 11].

Identification and measurement of wake periods throughout sleep are essential for understanding a patient's quality of sleep and being able to treat and diagnose sleep disorders. Key metrics to offer significant insight into the overall sleep patterns include identifying but also measuring the length and frequency of wakeful periods throughout a night of sleep [12, 13, 14].

Comprehending the related symptoms, etiology[1], and pathophysiology[2] is crucial in determining appropriate treatment for the underlying sleep disorder. By monitoring and classifying sleep stages, distinctive sleep patterns can be identified, and the quality of sleep can be assessed. Accurate and reliable methods for detecting wake periods are essential for ensuring effective sleep monitoring and management. The employment of these techniques gives valuable insight into the nature of sleep disorders and aids the development of more targeted treatment approaches [11].

Overnight Polysomnography (PSG) requires the patient to sleep at the hospital overnight and is the most common approach to sleep data collection. This comprehensive approach records a multitude of physiological signals during sleep, including brain activity, eye movements, muscle tone, and breathing patterns. However, this method is considered expensive, time-consuming, prone to human error, and requires significant storage space due to large amounts of data. These limitations have led to an increased demand for developing ASSC [9, 15].

The progress in sensor technology has significantly transformed the process of sleep data collection. This evolution leans towards the adoption of more simplified devices and methodologies, making use of a variety of biosignals. Concurrently, the emergence and development of Artificial Intelligence (AI) and Machine Learning (ML) technologies have further accelerated the transformation. Specifically, ML models have played a crucial role in the advancement of Brain–Computer Interface (BCI) technologies. BCIs function by detecting and interpreting the brain's bioelectrical signals and translating these into commands that can control external devices. ML algorithms enhance these systems by enabling them to understand and learn from the unique patterns in these bioelectrical signals, thus improving their performance over time [3, 4].

For sleep-wake monitoring devices, the adoption of ML has been instrumental in enabling a more precise and comprehensive analysis of sleep stages. Sleep, a complex physiological process, presents the challenge of accurately classifying its various stages. As conventional PSG approaches are time-consuming and complicated due to the necessity of manual sleep stage scoring, ML algorithms offer a compelling solution by automating this process [3, 4, 9].

Even with a reduced quantity of data compared to full PSG recordings, ML algorithms have the capacity to learn from datasets of sleep EEG recordings and identify crucial sleep patterns and trends.

[1] Etiology refers to the study of the causes or origins of diseases, disorders, or conditions.

[2] Pathophysiology focuses on understanding the abnormal functioning of the body in diseases or conditions.

However, one challenge with respect to sleep-wake classification is the unbalanced nature of the data. More specifically, sleep data typically contain a significantly larger quantity of sleep periods compared to wake periods, which can result in biased ML models that are less effective in identifying wake stages accurately. Despite this, advancements in ML algorithms continue to improve their performance in handling such imbalances, enhancing the feasibility of at-home sleep monitoring devices [3, 4, 9].

This highlights the importance of advancing the technology in the field of sleep research. With comprehensive methods and increased use of home sleep monitoring devices, better health monitoring and sleep disorder screening can be possible. This also emphasized the importance of further progress in ASSC for real-time sleep stage classification and wake detection [16].

1.2 Research Questions

The increase in sleep disorders and disturbances among human beings is a matter of significant concern. Sleep disturbances often involve periods of wakefulness during the night. By detecting and accurately classifying instances of wakefulness throughout the night, a more comprehensive understanding of a patient's sleep patterns and quality can be obtained. The scope of this thesis is based on answering the following research questions.

Research Question 1: **How do the use of raw sleep data and an unbalanced dataset with a small segment size affect the performance of sleep-wake classification with deep learning methods?**

Modern ML models are often based on the presumption of well-balanced and preprocessed datasets. However, this is often not reflective of real-world scenarios. In the context of real-time classification, minimal overhead, limited storage, and fast processing is of importance, and numerous processing steps between input and output might be limiting and challenging. Utilizing raw and unbalanced data eliminates these processing steps, paving the way for the possibility of real-time sleep monitoring and classification.

Research Question 2: **Can a reduced set of EEG channels achieve comparable performance in sleep-wake classification as high as high-density EEG channels?**

Given access to high-density EEG data from 128 channels derived from real patients, it is of interest to explore potential combinations of these channels that could yield optimal performance. This investigation aims to identify a possibly superior subset of channels that hold particular importance in the context of sleep-wake classification. This process has the potential to reveal and identify key channels that could contribute to a deeper understanding of the specific factors that significantly affect sleep-wake classification performance. Additionally, a careful examination of the sensitivity of wake is essential, given its high correlation with sleep disorders such as insomnia. Precise identification and accurate classification of wake periods are critical in these cases, underscoring the potential benefits of this approach.

1.3 Contributions

This thesis makes several contributions to the research field of ASSC. In making these contributions, the thesis answers the defined research questions and provides a foundation for future research.

An important contribution to this thesis is the investigation based on sleep data labeled by sleep experts from 128 channels of high-density EEG data sourced from The Human Sleep Laboratory of the International Institute of Integrative Sleep Medicine (WPI-IIIS), University of Tsukuba, Japan. This practical approach augments the reliability and relevance of the findings to real-world scenarios, further fortifying the value of the research within the field.

Furthermore, the thesis is set to reassess the traditional practice of employing balanced and preprocessed datasets for sleep classification in ML. In response to the first research question, it advocates for incorporating raw and unbalanced data, which offer a more authentic representation of real-world scenarios. This shift is anticipated to enhance the practical applicability and effectiveness of ML in sleep classification.

In response to the second research question, the thesis aims to ascertain if a subset selection of channels can match the performance of a full set of 128 high-density channels. This investigation offers both theoretical interest and practical value. The insights gained from this study could potentially enhance the efficiency of sleep monitoring and its accessibility. This forms a key contribution of this thesis to the realm of sleep research.

The outcomes and findings of this research, which contribute significantly to the field of brain informatics, have been submitted to the 16Th International Conference on Brain Informatics (BI 2023) with the title *"Automatic sleep-wake scoring with optimally selected EEG channels from high-density EEG"*. The full version of the submitted paper is available in Appendix B.

1.4 Outline of Thesis

The structure of the thesis is organized as follows:

Chapter 2: Theory initiates with an in-depth overview of relevant theoretical foundations, aiming to equip the reader with a comprehensive understanding for further analysis. It begins with a concise yet detailed introduction to brain anatomy, followed by an explanation of Electroencephalogram (EEG) signals. Subsequently, an extensive explanation of sleep is provided, with insight into the different sleep stages, common patterns and characteristics, sleep disturbances, and sleep monitoring. Following this, the thesis delves into ML theory, including neural networks, important performance metrics, and stating the class imbalance problem. The chapter ends with relevant theory considering genetic algorithms for channel selection.

Chapter 3: Literature Review provides a review of the current state-of-the-art in sleep-wake classification relevant and channel selection approaches. The review focuses on the areas of research that align with the proposed methodology of this research.

Chapter 4: Methodology delves deeper into the specifics of the research process. It details the char-

acteristics of the dataset used, as well as the methods and models employed, providing explanations and arguments for each choice made during the research process. The methodology is explained to facilitate a comprehensive understanding of the research process.

Chapter 5: Results presents all the results conducted from the research. This chapter provides comprehensive graphical representations that clearly illustrate the results obtained aligned with the set objectives of the research.

Chapter 6: Discussion, Conclusion, and Further Work concludes the thesis, where the results are discussed, conclusions are drawn, and recommendations for further work are provided.

Chapter 2

Theoretical Background

This chapter provides a theoretical grounding for the subsequent research, starting with an exploration of brain anatomy, particularly focusing on structures and processes related to sleep. This leads to an examination of electroencephalogram (EEG) signals, a crucial diagnostic tool in the study of sleep. This section presents the nature of the signals and the valuable information the signals contain regarding the functioning and status of the human brain during sleep.

The first segment delves into sleep theory, presenting a thorough overview of the different stages of sleep, along with the typical patterns and characteristics associated with each stage. The chapter presents various sleep disturbances and their impact on human health. This section aims to highlight the importance of accurate wake detection and sleep-wake classification for improving sleep quality. A brief overview of sleep monitoring methods is presented as well.

Moving forward, the chapter presents the relevant background for ML, concentrating on neural networks as a tool for pattern recognition and prediction. It elaborates on key performance metrics and important factors when training a neural network, with a special focus on the class imbalance problem.

Subsequently, the concept of channel selection using genetic algorithms is presented. This section frames the methodological approach employed in this research, offering insights into the chosen strategy for channel selection.

In essence, this chapter provides a comprehensive theoretical presentation of the research's relevant aspects. It offers a foundation for the complexities of sleep and the tools and methods used to analyze and interpret the obtained results concerning ASSC.

2.1 Brain Anatomy

The human brain is the most complex organ of the human body and governs all aspects of human behavior and cognition. It is the largest part of the Central Nervous System (CNS) and consists of billions of neurons for communication. The cerebral cortex, the brain's outermost layer, is responsible for higher-level cognitive functions. As the key region recordings of electrical signals for studying human brain activity, the cerebral cortex showcases diverse patterns of activity during sleep. This cortex is divided into two hemispheres, each further split into four lobes: frontal lobe, parietal lobe, temporal lobe, and occipital lobe, as illustrated in Figure 2.1 [17, 18].

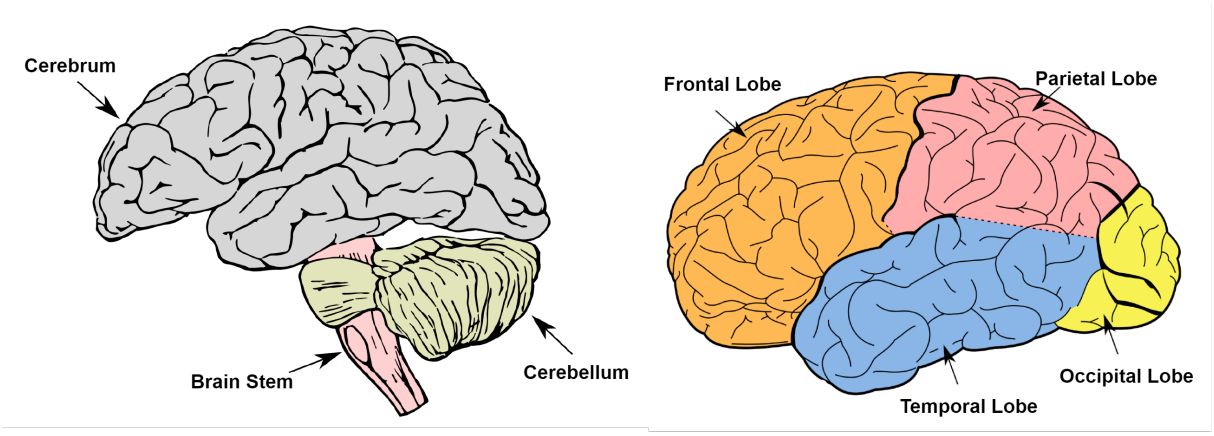


Figure 2.1: Anatomy of the human brain. The left part illustrates the three core parts of the brain; cerebrum, cerebellum, and brain stem. The right part highlights the four lobes of the brain, the frontal lobe, the parietal lobe, the temporal lobe, and the occipital lobe. Adapted from [17, 18]

The frontal lobe, located at the front and occupying about a quarter of the entire cerebral cortex, is the largest lobe. The frontal lobe is responsible for planning, problem-solving, decision-making, and regulating conscious thought and behavior. Within this lobe, the prefrontal cortex is concerned with personality expression, complex cognitive behaviors, motor planning, and execution. Significantly, it also contributes to maintaining wakefulness and alertness [17, 18].

Adjacent to the frontal lobe is the temporal lobe, consisting of the auditory cortex, which processes auditory information that might impact sleep quality. Besides its auditory role, it contributes to the processing and interpretation of other sensory experiences, memory formation, and emotional responses. Other key structures within this lobe include the hippocampus, vital for long-term memory storage, and the amygdala, which mediates emotional responses [17, 18].

The parietal lobe is responsible for sensory information processing, particularly spatial positioning, navigation, and body awareness, such as pressure, touch, and pain. This lobe is associated with mathematical reasoning, visuospatial processing, and important pathways for the visual system. The occipital lobe is the smallest lobe, responsible for visual function and processing, such as color, light, and movement coming from the eyes. The visual system is notably highly active during vivid dreaming [17, 18].

The thalamus is a small and oval-shaped structure located above the brainstem, deep in the center of the brain. The central location makes it possible for connections of the nerve fibers to reach all areas of the cerebral cortex. The thalamocortical system is the network of connections between the

cortex and the thalamus. All information has to pass through the thalamus before being directed to the cerebral cortex for further processing [17, 18].

The brain consists of specialized cells called neurons, which transmit information through chemical and electrical signals in the brain. The neuron generates electrical impulses when receiving a signal, releasing neurotransmitters at the synapse. The neurotransmitter is a chemical messenger that transmits signals between nerve cells in the brain and throughout the body. Synapses are the space between the end of an axon of one neuron and the dendrites of another neuron. The synapse is where the electrical signal is converted to a chemical signal with information, releasing the neurotransmitter to another neuron. Figure 2.2 illustrates the anatomy of the neuron and synapse [17, 18].

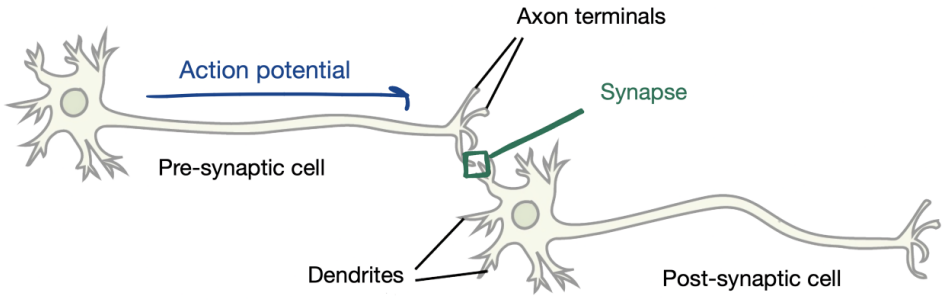


Figure 2.2: Specialized neuron cells with their axon terminals and dendrites. The synapse and action potential are shown, enabling the release of neurotransmitters to transfer information in the brain.

The neurons have a difference in charge between the outside and inside of the cell, referred to as the neuron's resting potential. When a neuron is stimulated, it generates electrical signals referred to as action potentials. When an action potential arrives in the synapse between two neurons, it triggers a transmission to a postsynaptic cell and generates an electrical signal called postsynaptic potential. The charge changes when another neuron receives a signal, creating a membrane potential. If electrical potential across the membrane exceeds a certain threshold, the neuron stimulates other connecting neurons, and the neuron fires, resulting in electrical signals [17, 18, 19].

2.2 Electroencephalogram (EEG) signals

Electroencephalogram EEG is a non-invasive[1] medical imaging technique used to measure the electrical activity in the brain. By placing electrodes on the scalp, EEG detects the electrical activity of large populations of neurons that fires at the same time. EEG mostly records signals from small areas of the brain surrounding each electrode. The electrical signals that are detected through EEG are amplified before further processing [19, 20, 21].

When several neurons fire synchronously, the electrical signals can be detected as wave patterns from different frequency bands. The brain patterns are commonly sinusoidal wave shapes, known as rhythmic activity. The different brain waves are categorized and illustrated in Table 2.1 and Figure 2.3.

[1] Non-invasive medical approaches avoid penetrating the skin or using invasive instruments, ensuring patient comfort and reducing potential complications.

Table 2.1: EEG frequency bands observed during sleep [20].

Waves	**Frequency**
Slow wave activity	0.5 Hz - 2 Hz
Delta (δ) waves	0.5 - 4 Hz
Theta (θ) waves	4 - 8 Hz
Alpha (α) waves	8 - 13 Hz
Beta (β) waves	13 - 30 Hz

Delta waves are the slowest waves with the highest amplitude. Slow wave activity is a subgroup of the delta waves, commonly referred to in association with deep sleep. The delta waves range from 0.5 to 4 Hz [19].

Theta waves represent consciousness towards drowsiness and light sleep, associated with relaxation. Theta waves are also associated with creative inspiration and are often referred to as creative brain rhythm. Suppression of theta waves can lead to anxiety, stress, and depression. Theta waves range from 4 to 8 Hz [19].

Alpha waves are slow rhythmic waves mainly established in the occipital lobe. Alpha waves are typically present when a person is awake but relaxed, typically with eyes closed, but can be attenuated by focused attention and mental concentration. Alpha waves range from 8 to 13 Hz [22].

Beta waves are associated with active thinking, focus, logical thinking, alertness, and concentration. Beta waves range from 13 to 30 Hz, most commonly in the frontal and central regions of the brain. The presence of beta waves is related to mental activities and increased energy, but can also refer to high levels of stress and anxiety [23].

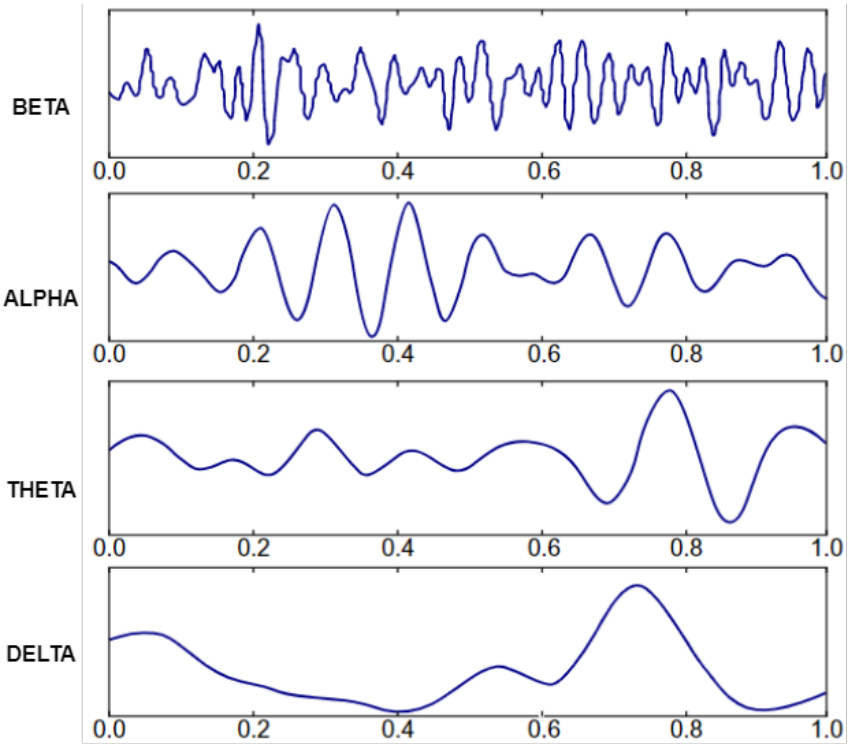


Figure 2.3: EEG brain wave patterns, adapted from [19]

2.2.1 Electrode Positioning

High-density EEG, which utilizes between 64 and 256 electrodes, offers a comprehensive coverage of brain regions, thereby providing a valuable tool in a clinical context. This surpasses the coverage of standard EEG, which often utilizes up to 28 channels. The arrangement of the electrodes on the scalp and how the signals are displayed and processed is referred to as a montage. Among montages, the bipolar and referential montages are the most commonly employed. It is crucial to acknowledge that the channel distribution in these montages is usually intended for general use. It may not be ideal for specific neuroscientific purposes, potentially overlooking the specific brain regions or patterns essential for certain applications, such as sleep-wake classification. [24].

The bipolar montage is characterized by each channel representing the difference in voltage between two adjacent electrodes, forming a chain of electrodes. This montage is useful for observing local differences in brain activity and detecting the electrical activity between two closely located points on the scalp. The term 'bipolar pair' refers to the two electrodes used for each channel [19].

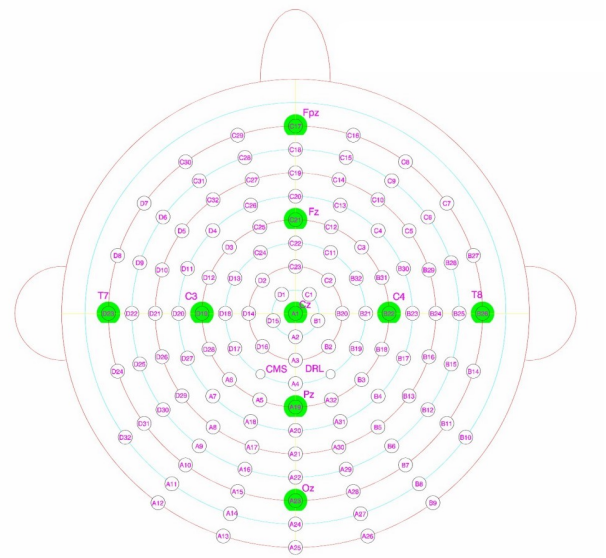
Conversely, the referential montage involves each electrode referenced to a single, common electrode. The voltage from each channel represents the difference between the electrode and the reference electrode. This montage is useful for providing a more global view of brain activity, where each channel represents electrical activity at a specific point relative to the common reference point [19].

Using a head cap facilitates the implementation of montages in the EEG recordings. Various cap configurations exist for high-density EEG electrode positioning. Figure 2.4b illustrates the international 10-10 head cap configuration, while Figure 2.4a illustrates the BioSemi 128 configuration with referential montage [25]. These configurations are similar in layout but use a different naming convention.

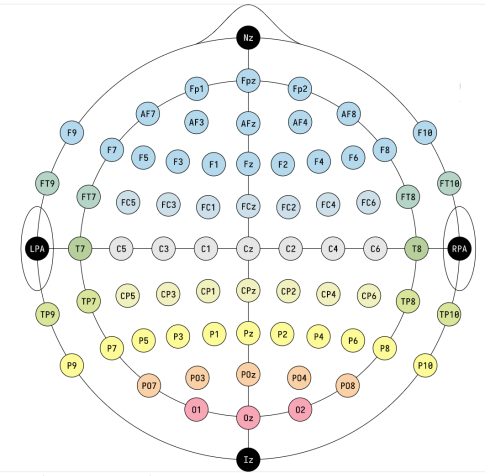
The international, standardized 10-10 system splits the skull into increments of 10% to place the electrodes, making each electrode relatively positioned to the others. This enables consistent positioning with respect to individual head shape and size. Each electrode has a letter corresponding to the region of the brain, together with a number corresponding to the side of the brain. The letters represent each lobe, F for the frontal region, P for the parietal, T for the temporal, and O for the occipital. Even numbers represent the right side of the brain, odd numbers represent the left side of the brain, and the central electrodes are clarified with the letter z instead of a number [26].

2.3 Sleep Physiology

This section presents an exploration of sleep physiology which is fundamental to understanding the structure of sleep, including its stages, patterns, and unique characteristics. Knowledge of these elements is crucial to understand the importance of sleep and for the identification and treatment of sleep disorders.



(a) The BioSemi 128 configuration for EEG channel positioning, adapted from [25].



(b) The international 10-10 head cap configuration EEG channel positioning. Adapted from [27]

Figure 2.4: Headcap configurations for high-density EEG recordings

2.3.1 Basic Sleep Elements

The process of transitioning from sleep to wakefulness is referred to as sleep arousal. Numerous internal or external causes, such as noise, movement, or adjustment in body posture, can cause arousal. Aurosals cause the patient's brain activity to change, together with increased blood pressure and heart rate. Irregularity of sleep arousals might indicate the presence of sleep disorders, such as insomnia [20, 28, 29, 30].

Sleep spindles are burst of rapid and oscillatory brain activity with a frequency between 11-16 Hz and duration between 0.5-2 seconds. Sleep spindles help the patient maintain sleep by protecting against external distractions and suppressing arousal. Higher density of sleep spindles correlates with improved memory performance, causing integration into long-term memory. K-complexes consist of specific patterns of brain activity characterized by a sharp negative deflection followed by a slower positive deflection. The K-complex has an approximate duration of 0.5 to 1 second. Both sleep spindles and K-complexes can be used as sleep depth markers and contribute important characteristics for sleep stage classification. Figure 2.5 illustrates the phenomenons [20, 28, 29].

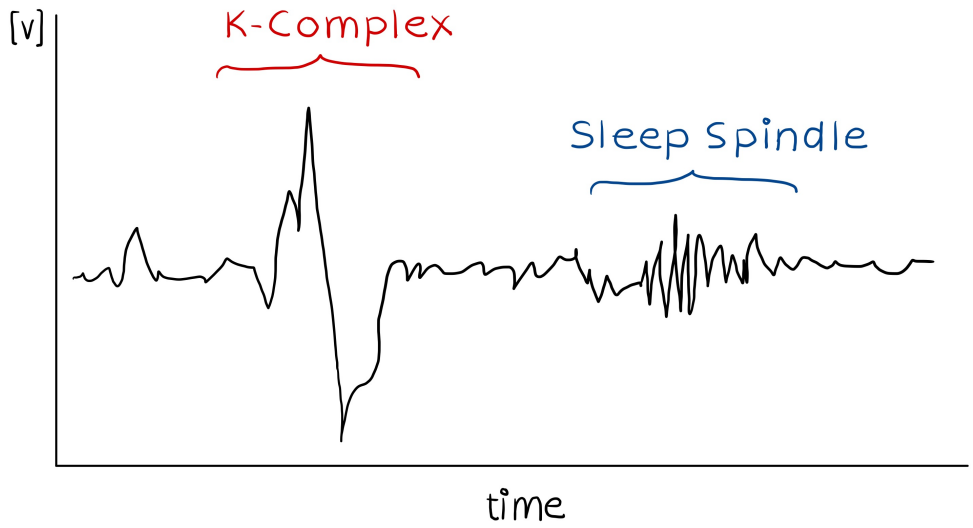


Figure 2.5: Illustration of sleep spindle and K-complex [1].

The circadian rhythm is another essential internal process that regulates the sleep-wake cycle of every individual, repeating every 24 hours. This process is controlled by the Suprachiasmatic Nucleus (SCN), located in the hypothalamus and the central biological clock of the body. The circadian rhythm operates based on environmental cues, mainly light and darkness, to align the body's internal clock with the solar day [9].

The circadian rhythm is regulated from the hypothalamus region at the base of the brain; see Figure 2.1. This region controls the production and release of the neurotransmitter orexin. During wakefulness, orexin is highly active and stimulates the arousal system, which promotes alertness and wakefulness. It projects to brain regions such as the cerebral cortex, thalamus, and brainstem. Orexin is less active during NREM sleep and silent during REM sleep (NREM and REM sleep are explained in Section 2.3.3). This allows orexin to control both the transition between the different sleep stages and the regulation timing of periods of wakefulness and sleep [9].

Melatonin is the neurohormone mediating the circadian rhythm and acts as the biological signal for nightfall, known as the hormone of darkness or the sleep hormone. The SCN instructs the regulation of melatonin, regulating the timing of when sleep occurs by systematically signaling darkness. The release of melatonin is stimulated by darkness and suppressed by light. The absence of circulating melatonin informs the body to prepare for a return to wakefulness. Melatonin itself does not directly induce sleep but rather facilitates the conditions for sleep to occur by communicating the time-of-day information to the relevant body parts helping to align and synchronize the body's numerous circadian rhythms [9].

Together with the circadian rhythm, the regulation between wake and sleep is also affected by sleep pressure. Increasing values of the neurotransmitter adenosine are correlated with an increased desire to sleep. Adenosine slows down neuron activity and promotes sleep and relaxation. The level of adenosine gradually increases in the brain throughout waking hours, and the progressive accumulation forms sleep pressure. When asleep, the levels of adenosine in the brain decrease gradually, reducing the sleep pressure and preparing for wakefulness. To suppress sleep pressure, caffeine

has the ability to block the adenosine receptors, delaying sleep pressure while maintaining alertness [31].

The regulation of the sleep-wake cycle is associated with the fluctuating levels of adenosine and melatonin in the body, operating in synchronization with the circadian rhythm. The accumulation of sleep pressure, steered by adenosine accumulation, in combination with the influence of the circadian rhythm and melatonin levels, creates a regulation that stimulates the desire to sleep. This relationship is illustrated in Figure 2.6. Process S refers to the sleep pressure, while process C is the circadian rhythm. When the gap between the lines is small, the patient has a strong urge to be awake. The greater the gap, the stronger urge to sleep. When the patient falls asleep, around 11 pm in the figure, the sleep pressure decreases together with the concentration of adenosine.

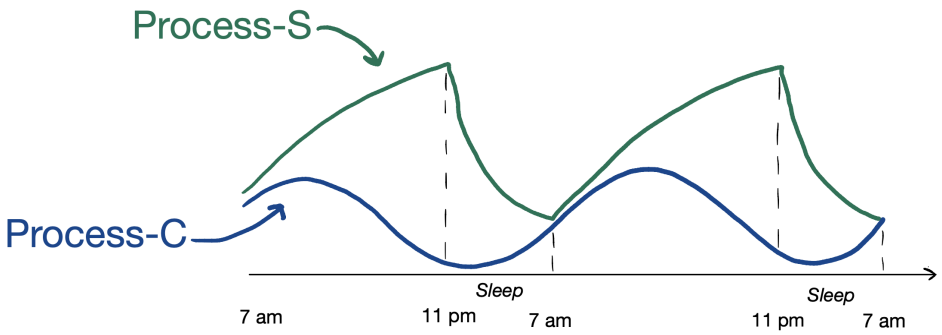


Figure 2.6: Illustration of the correlation between sleep pressure and circadian rhythm. Process S refers to the sleep pressure, while process C refers to the circadian rhythm.

2.3.2 Sleep Monitoring

Polysomnography (PSG) has been the gold standard method for sleep stage evaluation since the 1950s. PSG requires the patient to sleep overnight in a sleep laboratory at the hospital to record the required biophysiological signals. PSG uses multiple sensors to monitor the following signals; brain activity trough EEG, muscle activity through Electromyogram (EMG), hearth rhythm with Electrocardiogram (ECG) and eye movement with Electrooculograms (EOG). Body position is monitored with a camera, and breathing functions are monitored through oxygen saturation together with respiratory airflow and effect. PSG requires a controlled hospital environment and may discomfort the sleeping patient because of the number of sensors attached to the face and body [9, 15].

The recorded PSG data are split into 30-second epochs of data, and each epoch is manually classified into a sleep stage by a certified sleep expert. Manual scoring of the sleep stages and the sleep-related events are considered complex and time-consuming. Sleep monitoring based on PSG is limited to specialized sleep laboratories. The challenges of PSG has motivated research towards using ML models for ASSC and developing alternative devices to perform sleep monitoring outside of the clinical setting [4, 32].

Another important technology is the Brain-Computer Interface (BCI). The BCI is a communication system between a subject and an electronic device based on the brain activity of the user. BCI aims to translate brain activity into action, such as using classification algorithms. The algorithm is used to detect patterns in brain activity, making the BCI system a pattern recognition system. The informa-

tion from the EEG signals is analyzed and classified by ML algorithms, effectively translating neural patterns into stages of sleep. Thus, BCIs not only provide the platform for capturing vital sleep data but also facilitate the integration of this data with sophisticated classification systems, paving the way for a deeper understanding and monitoring of sleep health [33, 34].

When employing both PSG and BCI techniques, it is essential to score the acquired data. The scoring of the epochs from the sleep data is performed following the rules and recommendations from the manual of American Academy of Sleep Medicine (AASM). The AASM-manual consists of recommended rules for the routine of sleep scoring of both clinical PSG and at-home sleep monitoring. AASM also provides a recommendation for the use of at least three EEG channels from the international 10-10 configuration, specifically A28, B22, and C4 [20].

2.3.3 Sleep Stages

The body cycles through all the different sleep stages four to six times during a night of sleep. Each cycle lasts for about 90-110 minutes. Time spent in each stage of sleep changes both throughout the night and with age. The sequence of sleep stages throughout a night of sleep is visualized through a hypnogram. Sleep hypnograms are a graphical representation of the different sleep stages experienced during a block of sleep. The X-axis is represented by time, and the different sleep stages occur on the Y-axis. The hypnogram provides valuable information regarding sleep patterns, time spent in each stage of sleep, sleep disruptions or awakenings, and how long it takes for the subject to fall asleep [9, 28, 29, 35].

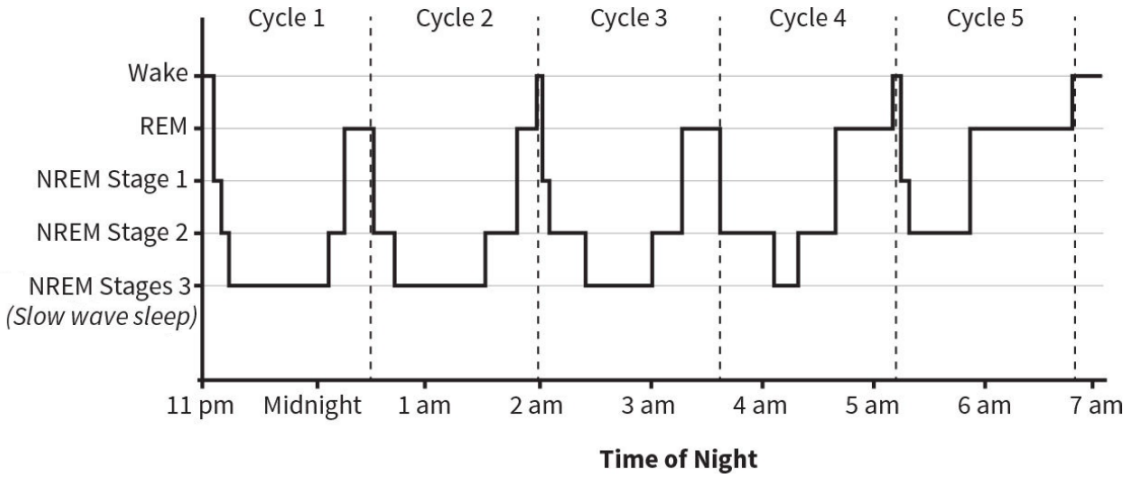


Figure 2.7: Hypnogram and normal distribution of sleep stages throughout a night of sleep. Adapted from [15].

Sleep is divided into three primary categories: Wakefulness, Non-Rapid Eye Movement (NREM), and Rapid Eye Movement (REM) sleep. NREM sleep is further subdivided into three distinct stages: N1, N2, and N3, each representing various depths of sleep in the absence of rapid eye movements. Contrarily, the presence of rapid eye movements identifies REM sleep and is predominantly associated with vivid dreaming. As presented in Figure 2.7, the hypnogram shows that the distribution of each sleep stage throughout a sleep cycle varies throughout the night. In the first half of the night, the majority of the ninety-minute cycle consists of deep NREM sleep. The second half, the majority of

the ninety-minute cycle, consists of REM sleep. The following provides a detailed characterization of these distinct sleep stages and their defining features [20, 28, 29].

Stage W (Wakefulness) is considered when the patient is awake and in the early stages of drowsiness. One of the first indications of drowsiness is a lack of eye blinks. Wake is characterized by reading eye movement, eye blinking, and rapid eye movement, together with a high level of chin muscle activity. This stage primarily consists of beta waves, and when the patient becomes more drowsy, the predominant pattern includes alpha waves. The chin EMG is usually higher than during sleep stages. The EOG may observe rapid eye blinks at a frequency between 0.5 - 2Hz. According to the AASM rules [20], the epoch is to be scored as stage W if more than 50% of the epoch either contains a) alpha waves over the occipital region or b) findings consistent with stage W are present. This includes eye blinking, reading eye movement, and rapid eye movement (normal or high chin muscle tone) [20, 28, 29].

Stage N1 (NREM 1) is the stage of drowsiness and the first stage of NREM sleep, a transitioning stage from wakefulness to sleep. During this stage, the brain produces high-amplitude theta waves, specifically slow brain waves. N1 makes up about 5-10% of total sleep time. Being the lightest stage of sleep, it is characterized by a slowing heartbeat, breathing, and eye movements, as well as muscle relaxation. According to the AASM rules [20], the epoch is to be scored as stage N1 if the alpha rhythms are decreasing and replaced by Low-Amplitude Mixed Frequency (LAMF) waves for more than 50% of the epoch. If arousals occur in stage N2, the successive part of the epoch is to be scored as N1 if LAMF is present and no K-complexes or sleep spindles occur. If arousals occur in stage R with subsequent occurrence of LAMF and slow eye movement is present, the eye movement part is to be scored as N1 until evidence of another sleep stage occurs [20, 28, 29].

Stage N2 (NREM 2) is the stage of light sleep, where the body temperature starts decreasing, and the heart rate slows further. This stage is characterized by the presence of K-complexes, sleep spindles, or both. The duration of N2 is about 25 minutes in the first cycle, increasing with each successive cycle. N2 makes up approximately 45% of total sleep time. During N2, slow eye movement may be present, but the EOG usually shows no eye movement activity. According to the AASM rules [20], an epoch is to be scored as N2 if one or both of a and b occur during the last half of the previous epoch or the first half of the present epoch. A) one or more K-complexes occur with unassociated arousals, or b) one or more sleep spindles occur. The sleep should be scored as N2 if the criteria for stage N2 is present for the majority of the epoch. If an epoch occurring after stage N3 does not meet the criteria for stage N3 anymore, it is to be scored as N2 if the epoch fails to meet the criteria for stage R or stage W and no intervening arousals occur [20, 28, 29].

Stage N3 (NREM 3) is the stage of deep sleep, referred to as Slow Wave Sleep (SWS) or delta sleep. This stage was previously divided into stages N3 and N4 but was combined in 2007. This stage consists of delta waves with higher amplitudes but lower frequencies. Stable breathing, decreased blood pressure, and heart rate are characteristics of this stage. This is the deepest stage of NREM sleep and makes up approximately 25% of the total sleep time. According to the AASM rules [20], if more than 20% of the epoch contains slow wave activity, the epoch should be scored as stage N3. Eye movement is unusual in stage N3, while sleep spindles may persist [20, 28, 29].

Stage R (REM) is the phase of sleep where vivid dreaming occurs, and the brain processes emotional information and consolidates memories. This stage is characterized by rapid eye movements, in-

creased respiration rate, and increased brain activity. REM makes up approximately 20 - 25 % of the total sleep time. The name REM encapsulate the periodic bursts of rapid eye movement that occur during this stage. According to the AASM rules [20], the epoch is defined as stage R if LAMF is present without K-complexes or sleep spindles, low chin EMG tone for most of the epoch concurrent with rapid eye movement, and rapid eye movement is present at any position within the epoch [9, 20, 28, 29].

2.3.4 Measurement of Wakefulness

The transition from wakefulness to sleep, known as sleep onset, is identified by a decrease in alpha and beta brainwaves and an increase in theta waves. Wake After Sleep Onset (WASO) and Sleep Onset Latency (SOL) are both significant indicators used in diagnosing sleep disorders like insomnia. Therefore, accurately detecting these periods of wakefulness is crucial for the proper diagnosis and treatment of such conditions [13].

WASO is defined as a measure of sleep fragmentation measured in minutes or percentage of total sleep time. It refers to the total amount of time the patient spends awake after initially falling asleep. WASO is used as an indicator of sleep quality and sleep continuity. A high level of WASO can be an indicator of the inability or difficulty of maintaining sleep. This is often associated with various sleep disorders, such as sleep apnea or insomnia. WASO of more than 30 minutes per night is used as a criterion for diagnosing insomnia [13, 36].

SOL measures the amount of time a patient uses to fall asleep, usually denoted by minutes. The duration is defined from the moment the patient attempts to fall asleep until the sleep onset occurs. The SOL can vary and be affected by factors such as circadian rhythm, age, or environmental factors such as light exposure and noise [12, 13].

The analysis of WASO and SOL underscores the importance of assessing sleep quality and the diagnosis of sleep disorders. As pivotal indicators of sleep fragmentation and latency to sleep, accurate detection and assessment are crucial in diagnosing sleep disorders effectively. With variables such as circadian rhythm, age, and environmental factors influencing these measures, WASO and SOL provide a comprehensive lens through which sleep can be examined, paving the way for personalized sleep treatments and interventions.

2.3.5 Sleep Patterns

As presented in Figure 2.7, sleep is typically divided into cycles of sleep, each consisting of NREM and REM sleep. The distribution naturally changes throughout the night, with more deep sleep occurring in the early part of the night and more REM sleep occurring towards the morning.

The most common sleeping pattern is a monophasic pattern, consisting of a continuous block of sleep time, usually seven to nine hours at night. Polyphasic sleep is a sleep pattern with multiple blocks of shorter sleeping periods throughout the day rather than a continuous block of sleep at night, similar to the sleeping pattern of young children. The continuous block of monophasic sleep is very commonly interrupted due to various factors, such as noise, temperature, or movement [37].

The architecture and structure of sleep patterns significantly transform with age, impacting both the duration and quality of sleep. As age advances, total sleep time linearly decreases, with an approximated reduction of ten minutes for each passing decade. Age-related changes in sleep are often manifested as earlier bedtimes and waking times, prolonged sleep-onset latency, reduced overall sleep duration, increased sleep fragmentation, and increased fragility of sleep. Additionally, there is a notable decline in deeper NREM sleep, increased time spent in N1 and N2 sleep, shortened and fewer sleep cycles, and increased wakefulness throughout the night. Figure 2.8 illustrates how the relationship between NREM, REM, and wake stages changes over time. How these changes affect the sleeping patterns is highlighted in the hypnograms in Figure 2.9, comparing the sleep patterns of a younger and an older adult [38, 39, 40].

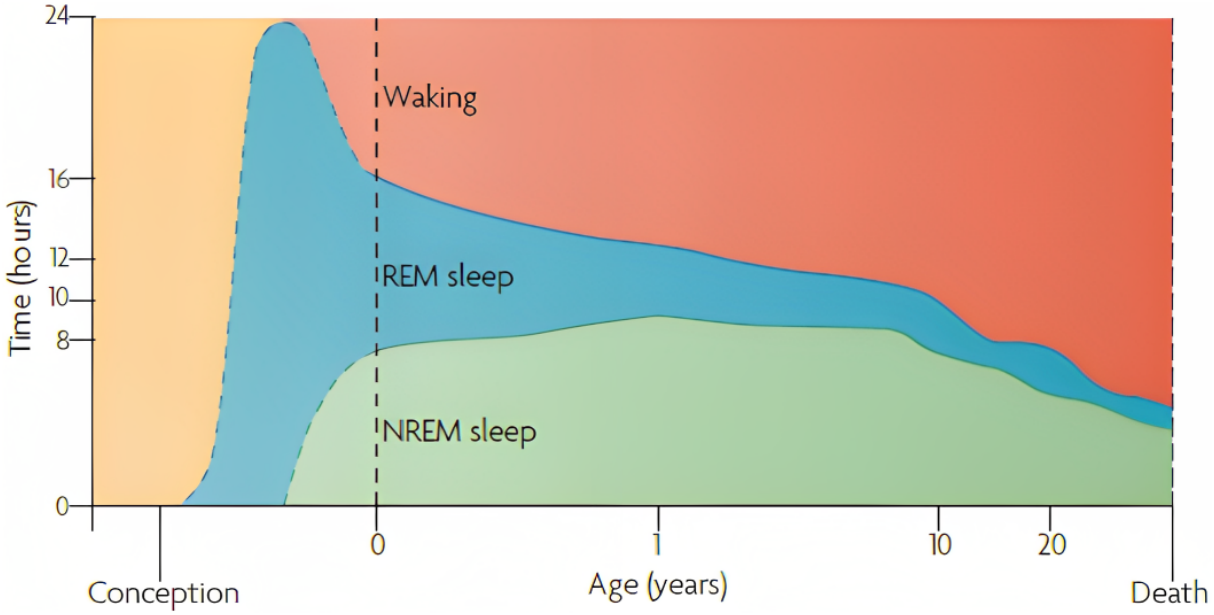


Figure 2.8: Evolution of sleep through age with decreasing amount of REM and NREM sleep. Adapted from [41].

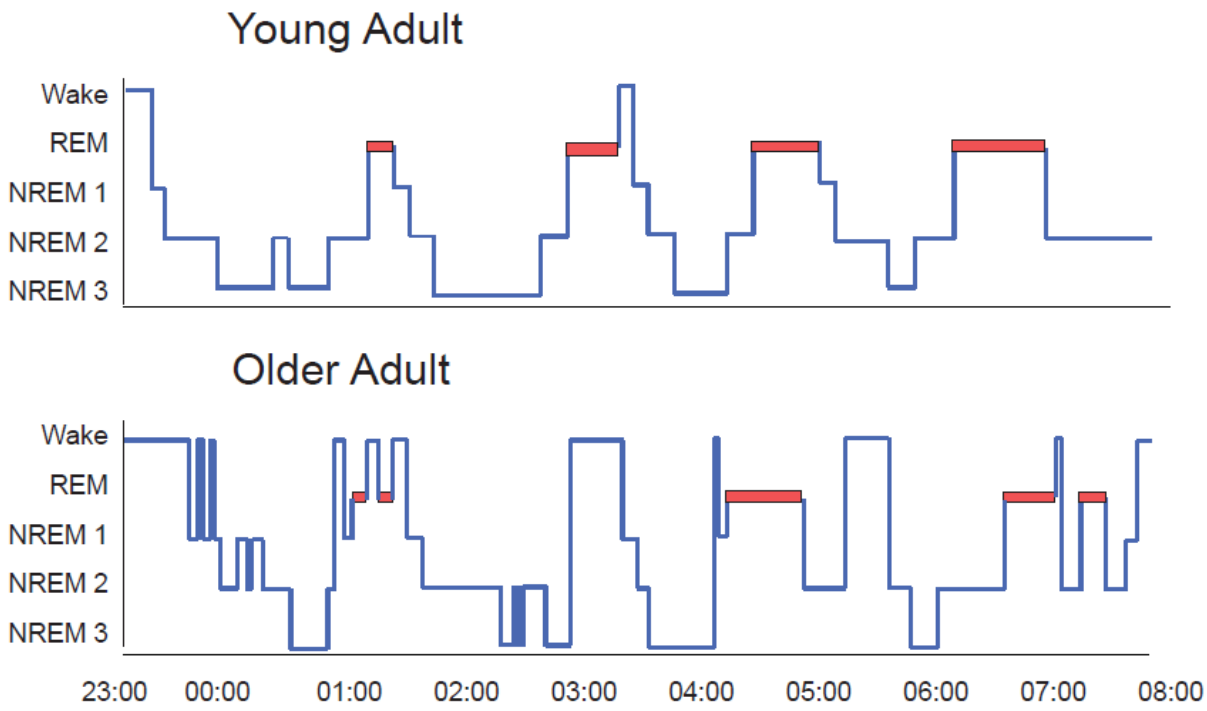


Figure 2.9: Example of change in hypnogram between young and older adult. Adapted from [41].

The sleep pressure observed in older adults compared with younger adults is of lower value, shown in Figure 2.10, contributing to the increasing difficulty of falling asleep. This reduced sleep pressure is intimately related to the significant decrease in SWS, a stage during which sleep pressure is predominantly relieved [38, 39, 40].

Another age-associated change is found in the dynamics of adenosine. As individuals age, adenosine accumulation levels tend to increase during wakefulness while the number of adenosine receptors decreases. This can potentially lead to adenosine insensitivity over time. Despite higher levels of adenosine in older adults, the reduced number of receptors coupled with a weakened adenosine signaling in the aging brain contributes to a reduced sleep pressure signal [38, 39, 40].

With respect to sleep spindles, the spectral power within their frequency range, as well as their occurrence, decreases with age, limiting the ability to suppress external disturbances. Furthermore, as individuals age, there are significant increases in sleep latency, WASO, and duration spent in N1 and N2 sleep stages. Conversely, the duration spent in REM sleep decreases with age, affecting processes such as the processing of long-term memories and increasing the risk of dementia [38, 39, 40, 42].

2.3.6 Sleep Disruption and Impacts

Sleep disturbances cause significant disruptions in sleep cycles. These aberrations in sleep patterns are not only unwanted but can also lead to a multitude of health-related concerns. Sleep disorders are conditions that affect the ability to obtain good sleep on a regular basis and have a significant impact on daily functioning and quality of life. The prevalence of sleep disorders increases with age, and lifestyle factors such as jet lag, shift work, poor sleep habits, or excessive screen time before bed can increase the risk of developing sleep disorders.

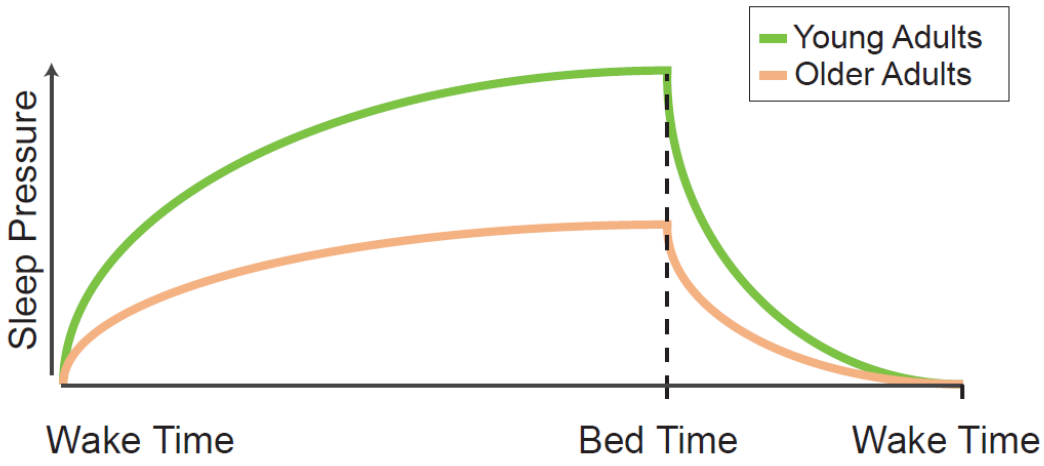


Figure 2.10: Reduction of sleep pressure influenced by age, adapted from [39].

Insomnia is defined as persistent complaints of dissatisfaction with sleep duration or sleep quality accompanied by difficulties falling asleep and staying asleep. Patients with insomnia often suffer from frequent and prolonged awakenings during the night, with difficulties falling back to sleep. Insomnia can result in a reduced amount of time spent in deeper sleep stages. Consequently, those with insomnia may spend more time in the lighter stages of sleep or awake, leading to a fragmented sleep pattern. Patients with insomnia have been noted with increased levels of beta waves during the night and often experience hyperactivity in the prefrontal cortex [7, 43, 44].

Obstructive Sleep Apnea (OSA) is characterized by repeated episodes where the patient stops and restarts breathing during sleep due to partial or complete blockage of the upper airway. The condition also consists of altered cardiopulmonary function together with daytime sleepiness. The patient can be prevented from getting enough oxygen. The condition is estimated to affect two to four percent of middle-aged adults but is often not diagnosed. Sleep apnea can disrupt the typical progression of sleep stages, often causing an individual to revert back to a lighter stage or awake state after each episode.

Circadian Rhythm Sleep Disorder (CRSD) is a group of sleep disorders characterized by misalignment between the timing of the individual's circadian rhythm and the external environment. This misalignment can lead to various types of sleep disturbances and complaints, most commonly insomnia and excessive daytime sleepiness. This can lead to sleeping and waking at unusual times, which can disrupt the typical sequence and duration of sleep stages. Such a disease is often treated with light therapy, behavioral strategies, or melatonin supplement [9].

Disturbances in sleep have been broadly associated with an array of diseases, spanning from physical conditions such as cardiovascular disease and diabetes to mental health disorders like schizophrenia, anxiety, and depression [45, 46].

For instance, between 40-90% of schizophrenia patients suffer from sleep disruptions, characterized by sleep fragmentation, extended SOLs, reduced sleep efficiency, and sleep duration. Similarly, sleep disruptions are common in children with Autism Spectrum Disorder (ASD), affecting approximately

40-80% of patients. Research suggests that autistic children experience a substantial 30-50% deficit in REM sleep compared to children without autism disorders [45, 47].

In the elderly population, specifically those over the age of sixty-five, one in ten adults is affected by Alzheimer's disease, a form of dementia. Identifying the specific type of dementia is crucial in order to implement the most effective treatment strategy. Moreover, sleep patterns have been increasingly recognized for their potential to aid in disease identification [42].

Additionally, research has linked irregular sleep duration with an increased risk of obesity and type 2 diabetes. Similarly, sleep disorders, including CRSD and insomnia, elevate the risk for various types of cancer [8, 9, 10, 11].

These findings underscore the vital role of addressing sleep disturbances and promoting healthy sleep patterns in disease prevention and overall health maintenance. Given the diversity and complexity of sleep disorders and their impacts, there's a pressing need for an accurate and reliable method of sleep stage classification and monitoring.

2.4 Machine Learning and Neural Networks

Machine Learning (ML), a sub-field of artificial intelligence, is concerned with developing algorithms that enable machines to learn and improve from their experience without being explicitly programmed. One of the main tasks in ML is classification, where the objective is to categorize data into predefined classes. When the classification task involves two classes, it is referred to as binary classification, and when it involves more than two classes, it is referred to as multi-class classification [48].

Neural networks are a type of ML model created with inspiration from the structure and function of the human brain. Neural networks consist of a network of neurons, which process and transmit data, similar to the brain's anatomy. There exist multiple types of neural networks. The neural networks are structured into layers: one input layer, one or more hidden layers, and an output layer, illustrated in Figure 2.11. A neural network is categorized as deep if the network consists of an input layer, an output layer, and two or more hidden layers. Otherwise, the network is classified as a shallow network. The complexity of the network increases with the number of layers. The networks are made of simple non-linear units, and the most abstract representation of the data is found in the higher or deeper layers of the network [49].

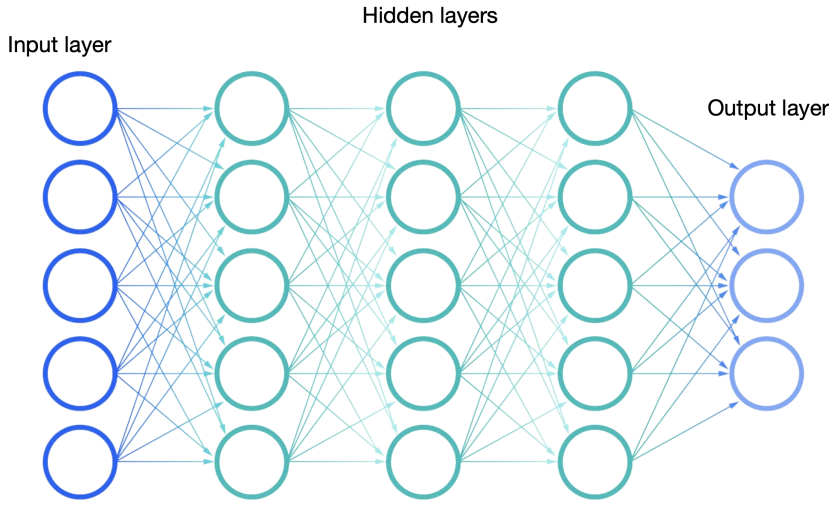


Figure 2.11: The setup of a neural network, shows the input layer, hidden layers, and output layer. The neural network is represented by interconnected circles, symbolizing neurons, with lines indicating the connections between them.

As stated above, deep learning is a subset of ML which uses artificial neural networks with multiple layers to analyze and learn from the data. The objective of a deep neural network is to detect and recognize complex patterns that are represented in the data. These patterns can be more abstract than those learned from more traditional models in ML. Deep learning methods have revolutionized many fields of computer science, including computer vision, speech recognition, and natural language processing. Deep learning models able to learn such complex patterns can be used for ML tasks such as classification [48].

The data begins in the input layer and progresses through the hidden layers. Each node in these layers applies a weighted sum operation on the inputs, followed by an activation function, thereby

successively transforming the data. The final transformed data is emitted at the output layer, resulting in the network's prediction. A neural network in which information solely flows from the input layer, through one or more hidden layers, and ultimately to the output layer is referred to as a feedforward neural network. A common algorithm often employed to train deep neural networks to discover structure in the data used for classification tasks is the backpropagation algorithm. In this method, the weights between nodes in the network are adjusted based on the discrepancy between the predicted output and the true output after a given epoch [50].

Convolutional Neural Network (CNN) is a form of deep learning algorithm consisting of numerous layers of nodes. CNN uses backpropagation to automatically and adaptively learn patterns and structures of the data. The different layers address a variety of feature detection and identification, allowing CNN to extract significant features from raw data without requiring hand-crafted feature extraction. With these qualities, CNNs have found widespread use in domains such as computer vision, natural language processing, and medical imaging [48, 49].

Recurrent Neural Network (RNN) is a type of Artificial Neural Network (ANN) which consists of recurrent connections. A recurrent connection allows information to flow in a cycle; this means that the output of a neuron can be fed back into the network as input. RNN are designed and used to analyze sequential data and time-series data, such as speech or text [49, 51].

2.4.1 Neural Network Training

ML models learn and acquire knowledge from datasets, which are typically split into training, validation, and test sets. It is essential to ensure that the data is representative and that random sampling is employed to avoid biases. Normal splits between the datasets are commonly set at 50%-25%-25%, 60%-20%-20%, or 70%-15%-15%. The training subset is used to train the model and should be sufficiently large to capture the underlying patterns in the data. The test set is used to evaluate the performance of the model on unseen data. It is crucial to keep the different subsets of the dataset completely separated from each other to prevent data leakage, which can lead to excessively optimistic, unrealistic, and misleading results for unseen datasets [48].

Training of a neural network is carried out in batches and epochs. A batch refers to the number of samples from the training set that is included in a subset that the optimizer propagates through the network during a single iteration. An epoch refers to the number of times the entire dataset passes through the network [48].

The validation set is utilized to tune the hyperparameters of the model. Hyperparameters refer to the model parameters set before training, such as the number of epochs, percentage of dropout, or learning rate. The hyperparameters can be adjusted to better generalize to new data by evaluating the model's performance on the validation set. Optimization and fine-tuning of the hyperparameters are often challenging due to issues like high dimensionality, expensive computation costs, non-convex optimization landscapes, and numerous interacting effects [48].

During the training of a neural network, the cost function is used to update the weights in the network to optimize the performance and reduce the loss. The generalization error estimates the model's performance for unseen data. The validation set is used to generate an estimate of the generalization

error of the model [48].

Overfitting is a critical issue in ML, which occurs when the model fails to generalize well to new data, resulting in high generalization errors. This poses a challenge for real-world applications, as the model may perform poorly on unseen data. In essence, overfitting arises when the model memorizes properties in the training data that are not present in the test set, such as noise. It typically happens when the model is too complex for the quality and amount of available training data or if the model is trained for too many epochs [48].

Regularization techniques refer to any modifications made to the model that aims to reduce the generalization error of the model and are used to prevent overfitting. Early stopping is a regularization technique that monitors the performance of the model throughout the training. When the performance of the validation set stops increasing, the training terminates. Dropout is another regularization technique that randomly drops some of the neurons during training in the hidden layers, removing their contribution to the next layer [48].

Batch normalization is a frequently used approach for improving the performance and stability of a neural network. It normalizes the inputs to a layer inside the network by subtracting the mean and dividing by the standard deviation of the inputs. This stabilizes the input distribution, improving the convergence and accuracy of the network, resulting in faster and more accurate training [48].

To evaluate the robustness and reliability of the model, it is recommended to run the model multiple times with different random partitioning of the data and report the average performance value and the standard deviation among them. The model is considered to perform consistently if the standard deviation is low [48].

2.4.2 Performance Evaluation

When designing ML algorithms, it is essential to evaluate the ability to perform well on new and unseen data. Classification algorithms are also known as predictors, as they are used to predict the class to which the incoming data belongs. The performance of these predictors is indicative of their real-world applicability and is particularly crucial for bioinformatics predictors. The ability to generalize well to unseen data, also referred to as generalization, is highly desirable. The use of quantitative performance measures makes it possible to effectively compare different models [52].

Performance measures are defined using the following components: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). These components form the confusion matrix, as illustrated in Figure 2.12. The confusion matrix provides a visual representation of the performance of the predictor. Perfect performance is obtained with nonzero values on the leading diagonal (highlighted in yellow) and zeros on all non-diagonal elements [52].

Type I error corresponds to a false positive, which occurs when a patient is incorrectly diagnosed with a disease despite being healthy. On the other hand, Type II error refers to a false negative, which occurs when a patient is classified as healthy despite actually having a disease. In the medical field, Type II error is generally considered more critical [52].

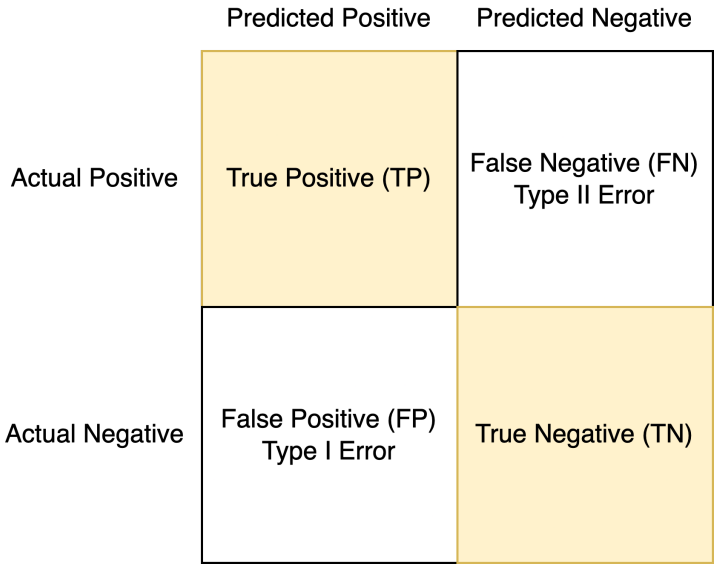


Figure 2.12: Definition of confusion matrix in binary classification. Actual values vertically and predicted values horizontally. Correctly classified samples are on the yellow diagonal.

Accuracy is the proportion of correctly predicted samples with respect to all samples. In a highly unbalanced dataset, however, accuracy can be misleading because it does not distinguish between the classes. Accuracy is defined in Eq. (2.1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{1}{n}(TN + TP) \tag{2.1}$$

Another performance metric in ML is recall, also known as sensitivity. It refers to the proportion of correctly classified true positives, defined in Eq. (2.2). In the context of sleep-wake classification, sensitivity represents the proportion of the total wake samples that were correctly identified as wake by the model. Sensitivity is useful when the cost of false negatives is high. In contrast, specificity is defined in (2.3) as the proportion of correctly classified true negatives. In the context of sleep-wake classification, specificity represents the proportion of total sleep samples correctly identified as sleep [52].

$$Recall = Sensitivity = \frac{TP}{TP + FN} \tag{2.2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2.3}$$

Precision measures the proportion of true positives out of all predicted positives. Precision is useful when the cost of false positives is high as the higher precision corresponds to fewer false-positive errors. Precision is defined in Eq. (2.4).

$$Precision = \frac{TP}{TP + FP} \tag{2.4}$$

The F1-score is the harmonic mean of precision and recall in one metric. It ranges from 0 to 1, where 1 represents perfect performance. The F1-score is especially useful for imbalanced datasets, as it punishes extreme values. It is defined in Eq. (2.5).

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.5}$$

Cohen's Kappa coefficient, referred to as the kappa score, measures the level of agreement between the predicted and true class while taking into account the possibility of agreement occurring by chance. The kappa score ranges from -1 to 1. Negative values indicate poor agreement, 0 indicates agreement by chance, and positive values indicate good agreement. The kappa score is defined in Eq. (2.6).

$$Kappa = \frac{p_o - p_e}{1 - p_e} \tag{2.6}$$

Where p_o is the observed agreement, and p_e is the expected agreement.

Area Under the Receiver Operating Characteristic Curve (AUROC) ranges from 0 to 1 and is a metric for how well the classifier distinguishes between the two binary classes. 1 indicates a perfect performance, while 0.5 is equal to random guessing. It represents the likelihood that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. AUROC is not affected by the proportion of the classes in an unbalanced dataset [52].

Metrics are essential tools for assessing the effectiveness of ML models in classification tasks. While accuracy is a useful metric, other metrics can give a more in-depth picture of the model's performance, especially for unbalanced datasets or when it's necessary to balance the trade-off between false positives and false negatives. Selecting appropriate performance metrics is an important aspect of the design process for a ML model.

2.4.3 The Class Imbalance Problem

The problem of class imbalance is of significant concern in ML when dealing with raw data. This issue is particularly relevant in problems involving the detection of potentially critical events, such as medical diagnostics. Class imbalance arises when the distribution of one class significantly outperforms the distribution of another class [53, 54].

When not appropriately addressed, the minority class is often ignored or undiscovered by modern ML algorithms. Common assumptions in ML techniques include roughly balanced datasets and equal cost of misclassification among the classes. However, these assumptions are often not present in real-world datasets. The cost of misclassification can be significantly different. In the context of wake-sleep classification, the problematic and desired part is to detect the presence of the minority class of wake samples compared to the majority of sleep samples [53, 54].

With unbalanced datasets, the choice of evaluation metrics is of great importance. For instance, consider a binary dataset with a 5-95% distribution between the minority and majority classes. The model can obtain 95% accuracy by classifying every sample as the majority class. However, if the objective is to detect the minority class due to the high risk of death, the results will be misleading and critical. In such cases, accuracy is not considered a properly isolated performance metric. Instead,

performance metrics such as precision and recall have to be considered.

Common approaches for handling unbalanced datasets are oversampling and undersampling or a combination of both techniques. When performing oversampling of the dataset, the distribution of the minority class is expanded to match the majority class. However, the use of oversampling increases the risk of overfitting, as the expansion can also propagate noise from the data. When oversampling high-dimensional datasets, the computational cost is significantly increased. This approach is common in image classification, as the minority class can be expanded by flipping, mirroring, or blurring the images. However, these approaches are not possible when oversampling EEG signals. A potential problem with oversampling EEG signals is that the additional copies may not be representative of the true distribution [53, 54, 55].

Undersampling involves using the same amount of data from the majority class as the amount of available data from the minority class. However, if the overall available training data is limited, this approach may work against its purpose. The disadvantage of the undersampling technique is that crucial information and characteristics from the majority class may be undiscovered or ignored [53, 54, 55].

2.5 Multi Objective Optimization with Genetic Algorithms

Genetic Algorithm (GA)s is a class of metaheuristic algorithms used for optimization. Metaheuristic algorithms do not guarantee to return the global optimal solution but provide good solutions in a reasonable time for complex optimization algorithms. GAs are inspired by genetics and the principles of natural selection [56].

In a GA a possible solution to the optimization problem is defined as a chromosome. A chromosome is constructed by a set of genes, each representing a variable that can be changed in the optimization problem. Several chromosomes are then grouped together in a population. Each chromosome is then evaluated by a fitness function with respect to one or more chosen objectives of the optimization problem [56].

The GAs generates an initial population consisting of N number of chromosomes. All N chromosomes in the population are evaluated by the fitness function, resulting in a subset of chromosomes with the best fitness. This subset is used as parents to produce a new population of N new chromosomes for the second generation in the algorithm [56].

The next generation of chromosomes is created using genetic operators such as crossover and mutation. The crossover operation takes two parent chromosomes and returns a new chromosome child with genes randomly combined from the parent chromosomes. Crossover is used to produce N new child chromosomes for the next generations. After all N children are generated, the mutation is performed. The mutation operation randomly changes the value of some genes in a chromosome. The number of chromosomes in each population and the number of generations of populations are considered hyperparameters of the optimization problem [56].

If the optimization problem has of interest to optimize for multiple objectives, Multi-Objective Genetic Algorithm (MOGA)s can be used. The result of a MOGA is a population of non-dominated candidate solutions. A set of non-dominated solutions cannot be improved in one objective without sacrificing another objective, representing a trade-off between multiple conflicting objectives. This is also known as a set of Pareto-optimal solutions that are used to create a Pareto Front [56].

2.5.1 Non-dominated Sorting Genetic Algorithm (NSGA)

Non-Dominated Sorting Genetic Algorithm (NSGA) is a multi-objective optimization algorithm used to find a set of non-dominated solutions to a problem. NSGA sorts the solutions into different non-dominated levels and then uses genetic operations such as selection, crossover, and mutations to generate new solutions. The algorithm starts with an initial population of solutions and iteratively generates new solutions in each generation until a termination criterion is met. Figure 2.13 illustrates the workflow of the algorithm [56].

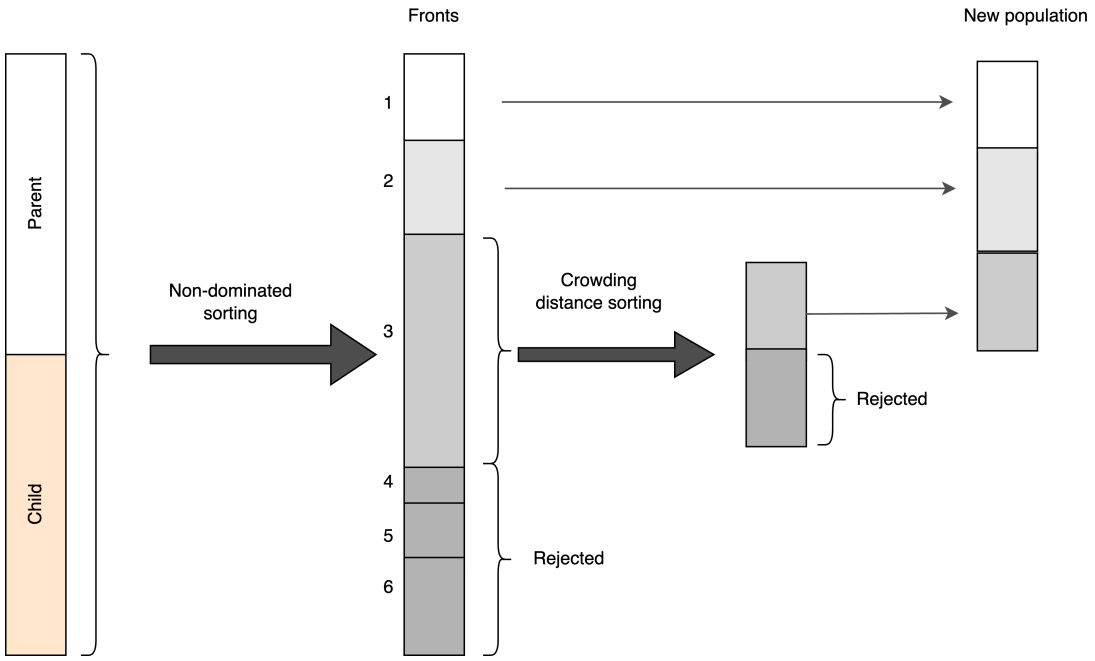


Figure 2.13: Illustration of NSGA procedure. The figure shows how a set of parent and child chromosomes is used to form a new population with non-dominated sorting and crowding distance sorting. Adapted from [56].

NSGA is not guaranteed to find the true Pareto optimal solutions as it only guarantees to find a set of non-dominated solutions that approximate the Pareto front. The quality of the solutions depends on the choice of the initial population, the genetic operations, and the termination criterion [56, 57].

NSGA-II is an extension of NSGA with modifications to the survival of the chromosomes, which is selected frontwise. If the front has to be divided, the surviving individuals are based on a crowding distance metric calculated as the Manhattan Distance. Additionally, the mating selection process is modified to compare individuals by rank and then by crowding distance [56].

NSGA-III has extended modifications and relies on initialized reference directions to space the reference points for optimization. If the front has to be divided, the surviving individuals are selected by filling up the most underrepresented reference direction. The solution with the smallest perpendicular distance in the normalized objective space survives. NSGA-III converges with each reference line, seeking to identify a non-dominated solution for each line [56].

Chapter 3

Literature Review

This chapter provides a comprehensive overview and examination of the previous research relevant for insight and comparison with the research performed in this thesis. The objective of this chapter is to present an overview of the status of the research in the field of sleep stage classification using EEG signals, proposed ML techniques, and channel selection algorithms.

3.1 Review of State-of-the-Art

As stated, overnight PSG clinic recordings are complex and time-consuming. Simple and cost-effective methods are preferred when gathering data for ASSC. As highlighted in Chapter 2.3, sleep influences many biological factors in the body. Several of these factors and biosignals can be measured and used in the classification of the sleep stage.

Monitoring body movement has been used for two-class sleep-wake classification. The detection of body movement was in the research of Hwang et al. [32] estimated based on the mechanical load of the bed using Thin Polyvinylidene Fluoride (PVDF) film. The obtained accuracy of estimated sleep-wake classification was 89.2%. By monitoring activity levels with near-infrared video, Liao et al. [58] obtained a sleep-wake estimation accuracy of 92.1% [32, 58].

The technology of Actigraphy (ACT) measures body movement by collecting data using wrist-worn devices and has had increasing usage in both clinical settings and for sleep medicine research. By comparing four wearable devices, Mantua et al. [59] found a correlation higher than 84% with PSG recordings. Despite EEG being more accurate, Mullaney et al. [60] found ACT methods to be more favourable with respect to the cost-benefit analysis. A limitation of EEG signals usage in sleep monitoring is the required time and expensive costs. ACT has the drawback of overestimating sleep by mixing the quiet wakefulness of sleep with a lack of detected movement [59, 60].

Such wrist-worn devices have been used to bring at-home sleep monitoring closer to clinical usage. Radha et al. [16] combined a deep RNN with transfer learning to perform four-class sleep stage classification. Transfer learning is a machine learning technique where a pre-trained model developed for a certain task is reused as the starting point for a model on another related task. The data were collected from a wrist-worn wearable device and obtained an accuracy of 76.36% and a kappa score of 0.65 [16].

Essential aspects of at-home sleep monitoring devices are computational power, memory usage, and battery time. The research of Van et al. [61] analyzed how utilizing a reduced sampling rate affected the precision of wake, NREM, and REM sleep measurement. The data was collected from barnacle geese with data loggers recording EEG signals, EMG signals, and head movement. The sleep-wake correlation was 0.9 when collecting 1 minute of data every 5 minutes. The correlation gradually decreased when recording down to 1 minute per 60 minutes, reduced to a sleep-wake correlation of 0.5 [61].

Striving to deliver affordable, user-friendly, and comfortable EEG sleep recording experiences at home, the DREEM Headband presents an appealing alternative to PSG. By enabling recording in the comfort of the own home of the subject, the DREEM Headband broadens access to sleep monitoring. The DREEM Headband is equipped with five EEG channels and a 3D accelerometer to measure movement, head position, respiratory trace, and rate. It adheres to the 10-10 international convention using the following EEG channels: F7, F8, Fp1, O1, and O2. The DREEM Headband is a reliable and easy-to-use tool for detailed at-home sleep assessments [62, 63].

The performance of the DREEM Headband has been compared with the staging of five sleep experts. When considering five classes, the device yielded an overall average accuracy of 83.5%, closely following the 86.4% accuracy rate achieved by the sleep experts. Moreover, it recorded an overall average kappa score of 74.8 compared to the average of 79.8 by the sleep experts. Such results underscore the potential of using simpler recording methods to gather high-quality biosignals for ASSC beyond clinical settings [62, 63].

When developing such BCI devices for sleep monitoring and staging, ML and different aspects of ANNs are used to enable ASSC. Such models are often based on publicly available datasets, usually preprocessed and balanced. Commonly used datasets are the Sleep-EDFX dataset [64], the PhysioNet Sleep-Expanded dataset [65], the DREAMS dataset [66] and the MASS dataset [67].

The most used dataset is from the Sleep-EDF database. Hassan et al. used the bipolar single-channel pair Pz-Oz from this database and proposed a model based on ANNs and extraction of statistical and spectral features, where spectral features refer to specific characteristics or patterns observed in the frequency domain of a signal. The distribution of the dataset consists of 54% wake samples and 46% sleep samples, balanced in favor of wake samples. The researchers obtained accuracy for two, three, four, five, and six sleep stages of 95.05%, 89.77%, 87.49%, 86.53%, and 85.57%, respectively [68]. Hassan et al. performed an additional experiment one year later and obtained accuracy for two, three, four, and five sleep stages as 99.48%, 94.10%, 92.14%, and 90.69%, respectively [68, 69].

Similar sleep-wake distributions are present in the two subgroups of data extracted from the ISRUC-Sleep dataset, 53% of the samples representing wake and 47% of the samples representing sleep. The data used was EEG signals from the C3-A2 channel together with EOG and EMG signals. With a nine-layer deep one-dimensional CNN, Satapathy et al. performed four experiments, utilizing one biosignal at a time and then as a combination of all of them. For the ISR-SG-I dataset, the obtained accuracies for sleep-wake classification were 98.44% for the single EEG, 98.91% for single EOG, 98.93 for single EMG and 98.93% for the combined signals. The obtained accuracies for the ISR-SG-III dataset were 98.51% for the single EEG channel, 99.24% for the single EOG channel, 99.14% for the single EMG channel and 99.21% for the combination of the signals for the ISR-SG-III dataset [70].

With respect to data distribution, Lee et al. [71] explored the data imbalance problem together with data augmentation. The model proposed uses EEGNet together with a sequence processing model of Bidirectional Long Short-Term Memory (BiLSTM). The EEGNet-BiLSTM was used by Lee et al. to compare the performance with or without data augmentation. The data utilized was Fpz-Cz bipolar EEG single-channel pair from the Sleep-EDFX Expanded dataset. The results obtained for five-stage sleep classification without data augmentation yielded an accuracy of approximately 82%, together with a kappa score of 0.70. The proposed method with data augmentation yielded an accuracy of 87% and a kappa score of 0.73 [71].

The greatest sleep physiology database available is made by data collected from Massachusetts General Hospital Sleep Laboratory. The algorithm is trained on PSG recordings from more than 10 000 patients, sampled at 200Hz from the following EEG-channels; F3, F4, C3, C4, O1, and O2. With this dataset, Biswal et al. [72] developed the specialized clinical decision support tool for sleep scoring called SLEEPNET utilizing deep RNN technology. SLEEPNET obtained an average accuracy of 85.76% and a kappa value of 79.46 for the five-stage classification [72].

Using several datasets, Fu et al. [64] developed a BiLSTM neural network (AT-BiLSTM) and obtained an accuracy of five-stage sleep classification with an accuracy of 83.78% and kappa of 0.766 on the PhysioNet Sleep-EDF Expanded dataset [65] and 81.72% and kappa of 0.751 on the DREAMS Subjects dataset [66]. The single EEG channel pair used from the DREAMS dataset were Fp1-A1 and Fpz-Cz from the Sleep-EDFX dataset [64].

Another BiLSTM architecture was proposed by Supratak et al. [73], DeepSleepNet. This deep learning model was designed with a CNN in combination with BiLSTM. Using the Sleep-EDF dataset, the model obtained an overall accuracy of five-stage sleep classification of 82.0% with the Fpz-Oz channel-pair. With the F4-EOG (Left) channel from the MASS dataset, the model obtained an accuracy of 86.2% for five-stage sleep classification [67, 73].

Bipolar single-channel pairs were also used by Berthomier et al. [74], proposing the Automatic Sleep Scoring Software (ASEEGA) software. The data was based on their own data and data from the PhysioNet database [65], using the two bipolar single-channel pairs C4-O2 and Cz-Pz. The obtained accuracy for two, three, four, and five sleep stage classification was 96.0%, 92.1%, 84.9%, and 82.9%, respectively, together with corresponding kappa values: 0.82, 0.81, 0.75, and 0.72 [74].

Pz-Oz is another commonly used bipolar single-channel pair, as used by Zhu et al. [75]. Balanced data from the Sleep-EDF database was used with a State Vector Machine (SVM) model with output from different visibility graphs. SVM is a ML algorithm that separates data points into distinct classes by finding an optimal hyperplane in a high-dimensional feature space. For sleep-wake classification, the proposed method obtained an accuracy of 97.9% and a kappa score of 0.96 [75].

Another research utilizing the Pz-Oz pair from the Sleep-EDF dataset was Al et al. [76]. Six-stage sleep classification obtained an overall average accuracy of 97.4% and a kappa of 0.87. The data was split into 30-second epochs, each epoch divided into 60 sub-segments before applying Discrete Wavelet Transform (DWT), a signal processing technique that decomposes a signal into different frequency bands, revealing both high and low-frequency components, and providing a time-frequency representation of the signal. The DWT was followed by clustering, a data analysis technique that groups

similar objects together based on their inherent characteristics, aiming to discover underlying patterns or structures in the data. 30 features were extracted to be used for the least squares SVM classifier [76].

3.2 Channel Selection, State-of-the-Art

The use of multichannel EEG is a common practice in BCIs. A key aspect of this methodology is the selection of EEG channels. The channel selection aims to improve the performance of the BCI by removing irrelevant or noisy channels. Decreasing the number of channels utilized reduces the amount of data to be processed, thereby saving computational resources and reducing processing time, particularly critical for real-time applications. The utilization of fewer channels can reduce the risk of overfitting the ML model due to the utilization of unnecessary channels. Channel selection might also provide information and identification of what brain areas provide class-event activity [77].

In the context of sleep stage classification, Stenwig et al. [78] used the NSGA-II algorithm for feature and channel optimization. The proposed method restrictively searches for 5, 4, 3, 2, and 1 channel to obtain the highest accuracy and F1 score [78].

EEG channel selection for sleep state classification in neonatal was proposed by Piryatinska et al. [79]. The research obtained an optimal selection of EEG channels and characteristics considered most suitable for separating EEG sleep stages. The proposed method used a multivariate analysis approach, adopting filtering with a search strategy for subset channel selection. The approach first scored the sleep stages based on each possible EEG channel combination before selecting the optimal channel combination. The method was applied to EEG signals from 14 channels collected from 36 neonates [79].

Optimization of channel selection and classification accuracy in EEG-based BCIs were investigated by Arvaneh et al. in [34]. The research utilizes a Sparse Common Spatial Pattern (SCSP) algorithm for EEG channel selection. A SCSP identifies and extracts relevant spatial patterns from EEG signals while promoting sparsity. The objective of the optimization problem was to select the least number of channels without compromising classification accuracy. The method was tested with two motor imagery datasets. The proposed method obtained an average improvement of the classification accuracy of 10% [34].

In the case of epileptic seizure classification, Moctezuma et al. [80] presented a multi-objective optimization method for EEG channel selection based on the NSGA-III algorithm. The method was tested on 24 patients from a public dataset. The objective of the optimization was to maximize classification accuracy while reducing the number of EEG channels. The optimization was able to obtain better results utilizing two channels selected by the NSGA-III analysis compared to using all 22 channels. [80].

3.2.1 Summary and Chosen Approach

The current state-of-the-art methods demonstrate high performance, with accuracy rates reaching up to 97.9% and kappa values as high as 0.96. These approaches span a variety of methodologies,

from SVM to deep ANN. The data deployed in these studies are both EEG signals derived from PSG recordings, but also supplemented with other biosignals. A common factor across these studies is the use of pre-processed, publicly available datasets characterized by balanced class representation. Several of the studies perform feature extraction, a complex and time-consuming process.

Distinctively, the methodology proposed in this study will use raw data without any form of pre-processing. This implies using a highly imbalanced dataset, with few samples of the class that are desired to detect the class of wakefulness. It also ensures that the data aligns closely with real-world raw data. Even though a highly imbalanced dataset makes the classification task harder, it poses a more realistic classification challenge with respect to real-time classification.

Given the unique opportunity to utilize 128 high-density EEG channels, the proposed method performs a multi-objective optimization process aimed at identifying optimal subsets of channels. The goal is to achieve comparable performance levels but with a significantly reduced number of channels.

From the literature on channel selection algorithms, the NSGA algorithm has provided promising results for several BCI paradigms. Utilizing a MOGA such as NSGA provides efficiency and versatility in handling multi-objective optimization problems.

The selected ML method is deep learning with deep neural networks. The combination of raw data together with the deep learning ML method obviates the need for prior knowledge of sleep, as compared to several of the methods with feature extraction from the literature. These choices are strategically aligned towards advancing real-time, ASSC, specifically, wake detection.

Chapter 4

Methodology

This chapter outlines the methodology applied throughout this research. The methodology consists of various key components. First, an overview of the employed dataset, its configuration, and the distribution of distinct sleep stages within the data is presented. This is followed by an explanation of the ML models and their set-up, along with the corresponding hyperparameters and selected evaluation metrics.

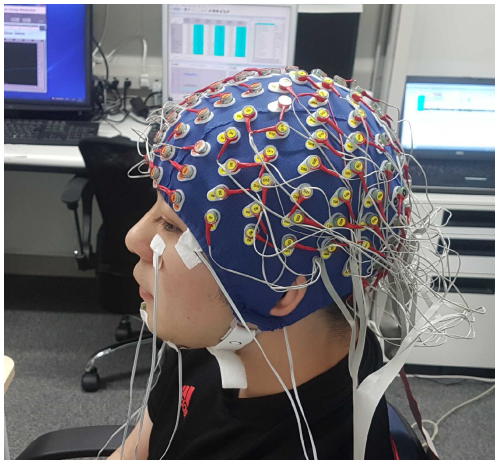
Next, a detailed explanation of the conducted channel selection experiment is presented. This includes the choice of the NSGA optimization algorithm, how the problem is formulated, and how the optimal solutions are selected and validated. The channel selection experiment also provides insight into the importance of each channel, which is then compared to important channels from the literature. The computational resources utilized are presented at the end of the chapter.

4.1 Human Sleep Laboratory Dataset

The Human Sleep Laboratory of the International Institute of Integrative Sleep Medicine (WPI-IIIS)[1], University of Tsukuba, Japan, collected and provided the dataset for this research. There is no public access to the dataset, and the University's ethics committee has approved the data collection. The dataset has not been preprocessed and gives access to raw sleep data. The dataset consists of fourteen subjects: five females and nine males, aged 22.5±0.9 years. Due to data interruption during the night of sleep, one subject was excluded from the current research, resulting in a final sample size of thirteen subjects. The recordings are performed in standardized sleep chambers, illustrated in Figure 4.1b.

The data is collected through PSG recordings using a total of 136 channels and a sampling frequency of 1024 Hz. This includes two mastoid channels, three EOG channels detecting eye movement, three EMG channels detecting muscle activity, and 128 EEG channels for measuring brain activity. The necessary equipment and setup for the PSG recordings are illustrated in Figure 4.1. The data for each subject consists of approximately eight hours of recorded sleep data with a sampling frequency of 1024 Hz. The data is divided into 30-second epochs, and a sleep expert labels each epoch into its corresponding sleep stage according to the AASM rules. Figure 4.2 presents a snippet of the data.

[1] https://wpi-iiis.tsukuba.ac.jp/



(a) PSG recording placement

(b) Patient sleeping in sleep chamber

Figure 4.1: Photos from the Human Sleep Laboratory for PSG recordings of patients.

The PSG recordings used for the WPI-IIIS dataset are performed using the BioSemi head cap, as illustrated in Figure 4.1a [25]. Access to such high-density real-world data is a unique asset in the conducted research. The positioning of the EEG channels follows the BioSemi 128 configuration, as illustrated in Figure 2.4a in Section 2.2.1.

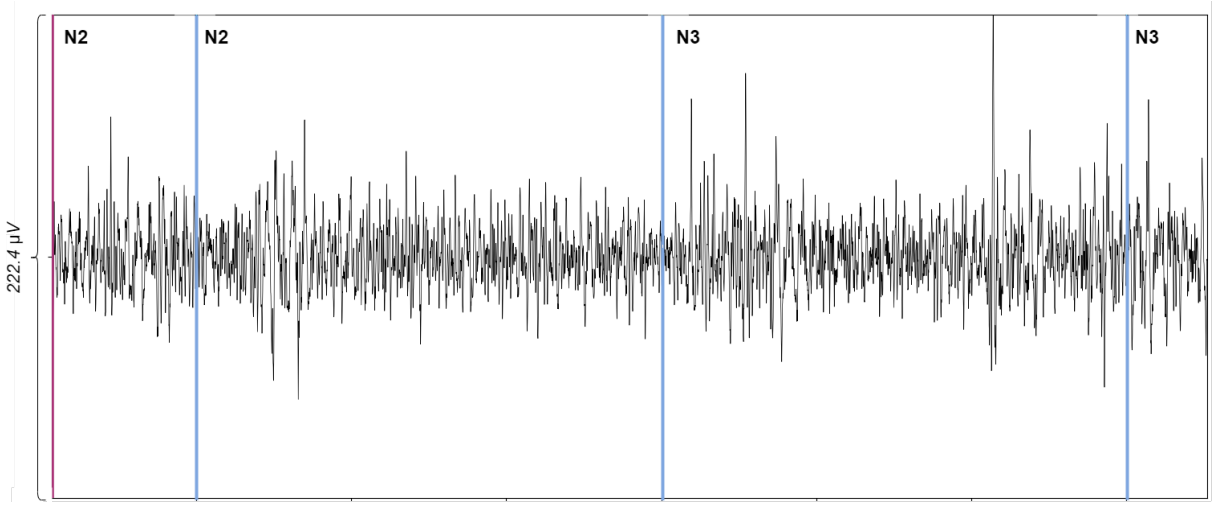


Figure 4.2: Extract of EEG signal from channel A28 used in the ML models. Resolution: 222.4 μV. Each blue line represents a new epoch, each epoch corresponding to 30 seconds of sleep data.

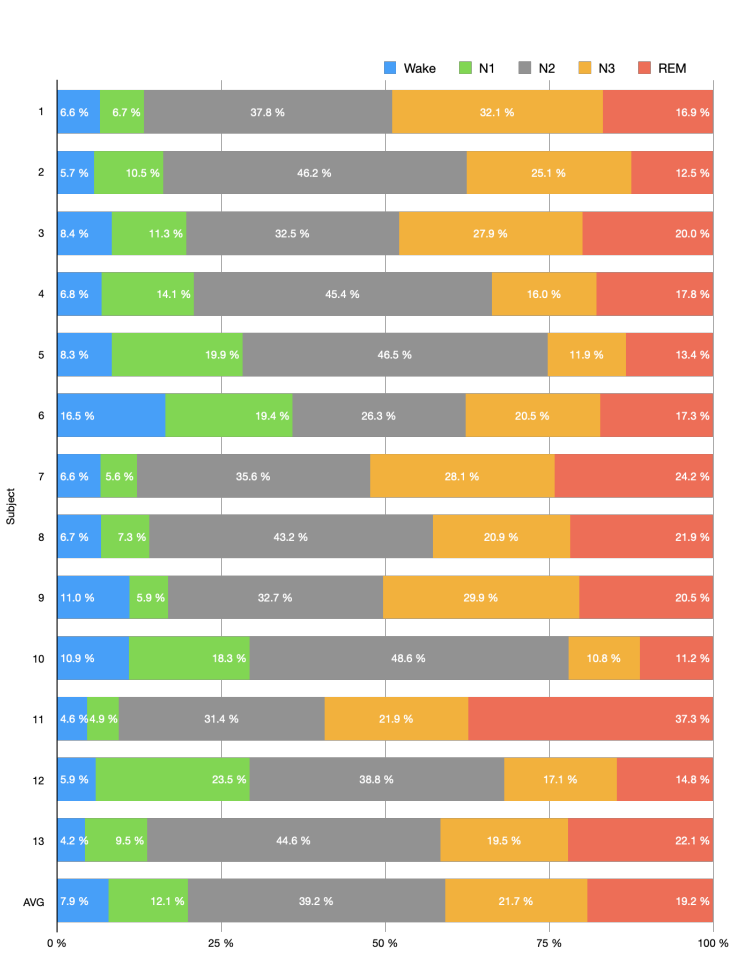
4.1.1 Data Preprocessing

To be able to mimic a real-world scenario where raw EEG data would be fed to the classification model, no preprocessing steps were performed on the dataset. By utilizing raw data without any preprocessing, there is a possibility of noise or other non-neural signals being present, which can pose additional challenges for the machine learning model in accomplishing the given task. The use of raw data collected from a human sleep laboratory makes this research unique compared to other studies performed on publicly available and preprocessed datasets that typically take these disturbances into account.

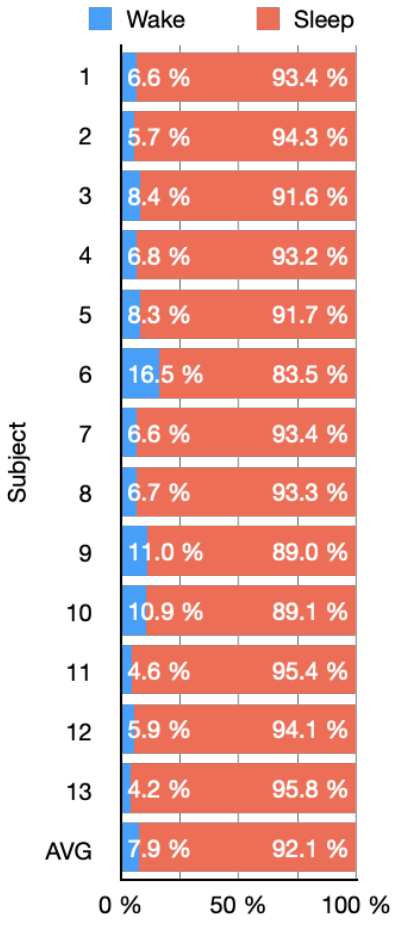
The original sampling frequency of the dataset is 1028 Hz. Analysis considering different sampling frequencies was performed by Herleiksplass in [1]. The analysis evaluated the model's performance with a sampling frequency of 1024 Hz, 512 Hz, 256 Hz, and 128 Hz. The analysis reported that downsampling the sampling frequency with a power of 8, utilizing 128 Hz, yielded the overall best results. Based on this conclusion, the dataset is downsampled to a sampling frequency of 128 Hz throughout this research.

4.1.2 Distribution of Sleep Stages

The distribution between the five different sleep stages for all subjects is illustrated in Figure 4.3a. This shows that the least represented sleep stage is wake, while the most represented sleep stage is N2. The main objective of this research is the detection of wake samples. Table 4.3b presents the distribution between wake and sleep samples, given in percentage. The lowest distribution is 4.2% of wake samples in Subject 13, and the highest distribution of 16.5% of wake samples is from Subject 6. With its high distribution of wake samples, Subject 6 is the most balanced subject.



(a) The distribution of wake and sleep stages for all subjects in the WPI-IIIS dataset, given in percentage.



(b) The distribution of wake and sleep stages for all subjects in the WPI-IIIS dataset, given in percentage.

Figure 4.3: Sleep stage distribution of the WPI-IIIS dataset given in percentage.

4.1.3 Sleep Hypnograms

Sleep hypnograms provide insight into the structure and sequence of the different sleep stages throughout a night of sleep. These characteristics can significantly influence the model's performance, particularly in the detection of wake samples. Hence, these observations can contribute to a deeper understanding of the model's behavior. In conjunction with the data distribution, hypnograms can reveal whether wake samples are predominantly located at the beginning of the recording, suggesting difficulty in falling asleep, or if extended periods of wakefulness occur in the middle of the night, potentially indicative of sleep disorders such as insomnia.

4.2 Machine Learning Model

The ANN model used for this research is the CNN EEGNet, developed for EEG-based BCI by Lawhern et al. in 2018 [81]. The architecture of the model aims to be generalizable across different BCI paradigms. Considering the input data is in its raw format, deep learning algorithms such as EEGNet have the ability to extract and learn relevant patterns from the data automatically.

The architecture consists of two convolutional blocks followed by a classification layer. The first block is designed to learn temporal frequency filters through temporal convolution. The second block extracts frequency-specific spatial filters through depth-wise convolution.

The combination of spatial and temporal filtering techniques produces the neurophysiologically interpretable feature, which can be more easily understood in the context of the underlying neurophysiology of sleep. This might be useful in understanding the classification decision and identifying potential biases in the model. The design encapsulates the known features of EEG signals, including optimal spatial filtering and frequency information, while reducing the number of trainable parameters.

With few trainable parameters, the EEGNet model is less likely to overfit and more efficient to train compared to other ANN architectures. EEGNet has been shown to be robust to varying sampling rates and robust to limited training data. These qualifications are promising for use in sleep-wake classification together with the WPI-IIIS dataset.

The recommended and default values are used for the hyperparameters of the EEGNet model. This includes a batch size of 16 and a dropout rate of 0.5 for subject-by-subject analysis. The ADAM optimizer is used. The loss function for binary sleep-wake classification is binary cross-entropy, and the loss function for multiclass sleep-stage classification is categorical cross-entropy. The models are trained for 100 epochs but implemented with early stopping to prevent the model from overfitting.

The training of the model is performed subject by subject. To evaluate the reliability and robustness of the model, the dataset is randomly split into training, validation, and test sets five times, resulting in different splits each time. The presented performance metrics are obtained through average performance among the five executions. The flow of the model execution is shown in Figure 4.4.

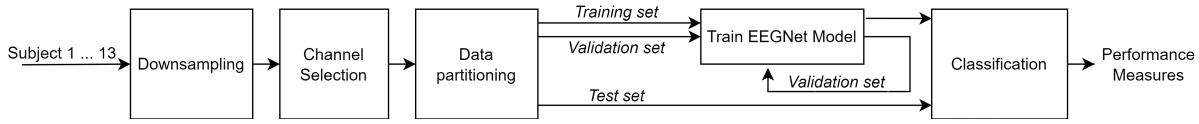


Figure 4.4: Workflow of ML system setup.

This research focuses on the analysis of wake detection within the framework of a binary sleep-wake classification system. To gain a more in-depth understanding of the classifier's performance, an experiment encompassing all five sleep stage classes is executed. By exploring the confusion matrices, this approach provides an opportunity to gain insight into the sleep stages where wake samples are commonly misclassified, particularly considering the similarity of wave patterns between wakefulness and REM stage. Extracting such insights can provide valuable guidance in identifying the factors and concerns that should be taken into account when implementing sleep-wake classification systems.

4.2.1 Hyperparameter Tuning

When training and evaluating the performance of the classification model, the data is split into training, test, and validation subsets. The dataset was chosen to be consistently split into 50% training data, 25% validation data, and 25% test data. The test set is kept completely separated from the training and is used to evaluate the model's performance on unseen data.

A comparison analysis is performed to decide what segment size to use to process the dataset. Utilizing a small segment size makes it possible to increase the number of instances of the minority wake class, which may contribute to better learning of patterns from the minority class. In addition, this allows for swift detection of the transition between the different sleep stages. Similar comparisons have been conducted in the research by Moctezuma et al. [82]. Hence, an analysis of different segment sizes is conducted.

4.2.2 Model Evaluation

Several performance measures are used to cover different aspects of the model's performance. A single performance metric may be misleading, especially due to the imbalanced dataset. The most presented performance metric reported in the literature is accuracy. It is a simple and intuitive metric commonly used in different ML paradigms. Dealing with an imbalanced dataset, accuracy may be misleading. Accuracy measures the proportion of correct predictions out of all predictions, but does not distinguish between the numbers of correctly classified samples for different classes.

The kappa score is considered more robust to an imbalanced dataset than accuracy, as it considers the possibility of the prediction being correct by chance. For binary classification, the trade-off between specificity and sensitivity is measured by the AUROC. This metric is not sensitive to imbalanced datasets. Accuracy, kappa, and AUROC are considered single performance metrics encapsulating the performance of the whole model. Given the imbalanced nature of the dataset, it is crucial to monitor performance measures for individual classes as well.

As the primary objective of this research is to detect occurrences of wake, the correct classification of wake is of high importance. Due to the trade-off between the importance of sensitivity versus specificity, sensitivity is considered more important. Sensitivity measures the proportion of actual positives correctly identified by the model. It is more important to actually detect a wake sample, risking false positives by classifying a sleep sample as wake. Since the evaluation considers medical diagnosis, it is considered more crucial to leave a patient undiagnosed with a disease compared to misdiagnosing a patient, which would lead to a further diagnostic analysis of the patient. In this context, higher sensitivity is considered more critical than high specificity.

Since the models are executed five times, and the presented results are average among the models, it is also favorable for the model to obtain a low standard deviation. This would imply that the model is consistent, more generalizable, less overfitted and robust to different splittings of the dataset.

Evaluating and examining this set of evaluation metrics provide a comprehensive view of the model's performance. This ensures that the model performs well across all the relevant dimensions, and not just well in the overall accuracy of the model.

4.3 Channel Selection using NSGA-III

Channel selection is a step that involves selecting a subset of the available EEG channels to use in the analysis. It is often performed to reduce the dimensionality of the data while increasing the interpretability of the results [77].

Channel selection is an essential step, as it reduces the computational complexity of the sleep stage classification algorithm by reducing the number of channels used in the analysis. This leads to faster and more efficient processing of the data. Another benefit is the improved interpretability, as it identifies the electrodes most relevant for the particular problem. This can provide valuable insights into the underlying physiology of sleep and how it is affected by different factors. In addition, channel selection can help improve the classification algorithm's performance by removing noisy or irrelevant channels, making the classifier more robust and reliable [77].

The objective is to investigate whether it is possible to achieve comparable results for sleep-wake classification utilizing an optimal subset of fewer channels. Reducing the number of channels utilized in the ML model can lead to a more efficient and less computational-heavy algorithm. This contributes towards making the model more practical and scalable for real-world applications. Multiple conflicting objectives arise since both performance measures and the number of channels are variables in the optimization problem. The multi-objective optimization algorithm NSGA-III was implemented for the EEG channel selection.

The NSGA algorithm has proven efficiency and versatility in optimizing multiple conflicting objectives simultaneously [78, 80, 83]. The algorithm considers both the inclusion of the informative channels and the exclusion of the irrelevant ones. NSGA has demonstrated good performance in handling large search spaces, which is the case when dealing with high-density EEG datasets. The diversity in the populations of NSGA reduces the likelihood of convergence on sub-optimal solutions, which ensures continuous exploration of potential channels even after a local optimum is found.

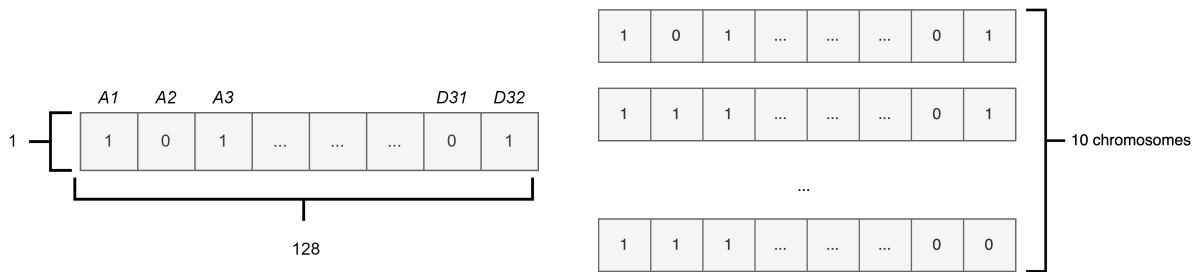
The algorithm assures flexible optimization by returning a set of Pareto-optimal solutions representing different trade-offs between the objectives. This flexibility provides greater insight into the decision-making process for solution selection, allowing one to prioritize the different objectives freely after execution. A solution that is optimal for multiple objectives is likely to perform reasonably well across all objectives, even in the presence of noise or uncertainty. Given these considerations, NSGA-III is considered a powerful tool for the channel selection problem.

4.3.1 Problem Formulation and Implementation

To accurately analyze the sleep data, the optimization problem consisted of four objectives: 1) maximizing model accuracy, to improve the reliability of the classification, 2) maximizing the kappa score, which measures the agreement between predicted and actual classifications, 3) maximizing the AUROC, indicative of the ability to distinguish between different classes, and 4) minimizing the number of EEG channels used during the classification. Each of these objectives is vital to the performance and efficiency of the sleep-wake classification model. However, increasing the number of objectives significantly adds to the computational cost and increases the complexity of the problem. The NSGA analysis is performed in a subject-by-subject matter.

In the experimental phase, it was observed that utilizing as few as eight channels generated encouraging outcomes. Given these findings, it was logical to explore the feasibility of solutions that used a maximum of ten channels. The decision to limit to ten channels was also influenced by considerations such as computational efficiency, simplicity of the model, and practical applicability, particularly with the view towards portable, at-home sleep monitoring devices which have constraints in the number of sensors that can be comfortably attached to the user. Therefore, the four objectives were optimized within the constraint of utilizing ten or fewer channels.

In this optimization setup, the chromosomes of the NSGA are defined as a binary array of size $1x128$, representing each available channel in the dataset. A value of 1 indicates that the channel will be used for the classification process, while 0 indicates that the channel is not to be used. The NSGA implementation was designed to generate ten chromosomes for each generation, iterating the process for a maximum of 150 generations. The algorithm's termination criteria were set based on the maximum number of generations, with a tolerance of 0.001 over a 10-generation span. This allowed for a balance between maintaining diversity in the population and managing computational resources effectively. The configuration of the chromosome and population is illustrated in Figure 4.5.



(a) Definition of one chromosome in the NSGA optimization algorithm.

(b) Definition of one population in the NSGA optimization algorithm.

Figure 4.5: Definition of chromosome and population in the NSGA optimization algorithm.

4.3.2 Selection and Validation of Optimal Solutions

The completed NSGA analysis returns a set of non-dominated solutions to approximate the Pareto front. These solutions are considered equally optimal as improving one objective would compromise another. The set of solutions does not guarantee to include a solution for every possible variation from one to ten channels. Moreover, it may include multiple potential solutions with the same number of channels but featuring different subsets of channels. A manual assessment is made to determine the best solution among the provided potentials. The preferred solution is chosen based on the prioritized order of highest kappa, highest AUROC, and highest accuracy, with varying numbers of channels.

The best-performing subset of channels from the NSGA analysis for each subject is then used to train and evaluate a full-fledged model. This allows for further analysis with extended performance metrics than the ones used in the optimization problem itself. This makes it possible to compare the performance with the other subsets of channels presented in this research.

4.3.3 Channel Importance Analysis

Given the results of the NSGA analysis for each individual subject, the frequency of occurrence for the different channel's presence is of interest. The most present channels, regardless of subject, can be considered the most important channels. In order to investigate the frequency of occurrence, a channel importance analysis is conducted. This research can gain insight into the factors contributing the most to sleep-wake classification and, specifically, the detection of wake periods.

Considering each subject's entire set of solutions, the frequency of a given channel's presence is counted and analyzed. This makes it possible to extract the order of importance of all channels and create different subsets based on their importance. These subsets are then used to train and evaluate a full-fledged model, allowing for further analysis with extended performance metrics.

This approach enables the investigation of whether a common subset of selected channels can be extracted and yield good results across different subjects. This enables suggesting the possibility of a generalized set of channels that could be expected to perform similarly on unseen data and might indicate to be more universally applicable for sleep-wake classification.

4.4 Channel Selection and Comparison

The availability of high-density EEG data from 128 channels from a human sleep laboratory is considered a unique asset. The initial experiment of utilizing all 128 available channels allows for the most detailed analysis possible, utilizing the maximum amount of available data for the deep learning model. The underlying objective of research question two is to assess if achieving a similar performance level utilizing a significantly reduced amount of channels is possible. This approach then provides a robust baseline for comparison with the subsequent experiments utilizing fewer amount of channels.

The main objective of this research is wake detection and sleep-wake classification. However, the dataset used provides five-class sleep stage labeling, and utilizing all 128 high-density EEG channels

in a five-class classification can give additional and deeper information on the problem at hand. The multiclass classification is conducted to further investigate the impact of sleep physiological aspects, aiming to give a better understanding of the performance of the sleep-wake classification. Examining the confusion matrices in the multi-class scope gives insight into which classes the instances of wake are most often misclassified as and can be difficult to detect. As five-class sleep-stage classification is not the main objective, it is not included in the optimization part of the research.

Various studies from the literature have used and recommended specific subsets of channels with promising results [20, 63, 68-76]. However, several important aspects have to be considered when comparing the conducted results from this study with the results from the state of the art.

Firstly, the data used in this study is utilized in its raw format, not subject to any preprocessing, in comparison to publicly available datasets, which are preprocessed. The preprocessing process can involve techniques that remove artifacts, clean and improve the data, such as band-pass or notch filters. The data used in this study may be considered more challenging to analyze.

Secondly, the distribution between sleep and wake samples from the datasets used in the literature are balanced. The high level of sleep versus wake imbalance of the WPI-IISI dataset is an additional contribution to why the data used may contribute to a more challenging analysis.

It is of interest to validate the performance of the suggested and reported subsets from the literature with the model and dataset used in this research. If these subsets perform well, it indicates robustness to the presented challenge, while poor performance could highlight the importance of further research in sleep-wake channel optimization. Due to the increased level of challenge for the data analysis, poorer performance for the presented subset might be lower than presented in the literature. This would also indicate that the performance for the presented subset from the NSGA-III analysis might perform better on publicly available datasets. This is an essential insight to investigate concerning the comparison of the model performance.

On behalf of the presented aspects, numerous models were trained and evaluated with the method presented with different subsets from the literature. This includes utilizing the recommended channels from AASM to investigate if the subset obtained from the NSGA-III analysis performs comparably well.

The best-obtained results with few channels from state-of-the-art are conducted with the two bipolar single-channel pairs , Pz-Oz and Fpz-Cz. These channels are utilized for comparison. As this research wants to contribute with analysis that benefits the research towards the goal of real-time ASSC at home, the channels used from the sleep-monitoring device DREEM are used.

This establishes a solid foundation for a thorough assessment and comparison of the chosen subsets of channels, as well as a robust evaluation of the results. These models are trained and tested alongside the resulting subsets of channels from the NSGA-III analysis. To avoid confusion, the corresponding channels between the 10-10 and Biosemi configurations are explicitly laid out in Table 4.1.

Table 4.1: Correspondence between the channels presented in the BioSemi configuration and the international 10-10 configuration. The subset of channels presented is chosen to compare the channels used in this study with the channels used in the literature.

	10-10 configuration	**Biosemi configuration**
AASM	O2	A28
	C4	B22
	Approximately F4	C4
Bi-polar pairs	Pz - Oz	A19 - A23
	Fpz - Cz	C17 - A1
DREEM	F7	D7
	F8	C7
	Fp1	C29
	O1	A15
	O2	A28

4.5 Computational Resources

Given the configuration of the conducted experiments and computational complexity, all executions are performed using GPUs on the NTNU IDUN computing cluster [84]. The cluster is equipped with more than 70 nodes and 90 general-purpose GPUs. Each node is connected to an Infiniband network and comprises two Intel Xeon cores and a minimum of 128 GB of main memory. Half the nodes have two or more Nvidia Tesla P100 or V100 general-purpose GPUs. The storage infrastructure of IDUN consists of two storage arrays and a Lustre parallel distributed file system, which collectively facilitate efficient access and data management.

Chapter 5

Results

This chapter presents the obtained results from the conducted research. Firstly, the analysis of segment size is presented, and the obtained results are utilized for the remaining experiments.

Next, the performance of the ML models is presented. The chapter comprehensively evaluates the models with multiple metrics, including accuracy, kappa, AUROC, and class-specific sensitivity. Models are created with the use of all available 128 high-density EEG channels. The results will be presented both in tabular and graphical formats for better comprehension. These results are used to select a subset of four subjects, which can be used for a more in-depth analysis of their performance.

To gain additional insight into the classification of sleep stages, a five-class sleep-stage classification is performed, highlighting examples of confusion matrices and sleep hypnograms from the four chosen subjects.

Subsequently, the results obtained from the NSGA-III optimization for channel selection will be presented. This involves overall average results among all subjects for three to ten channels, and results for each subject utilizing the subset of channels from their best NSGA solution. A channel importance analysis is presented with the use of the NSGA results from all subjects. The most frequent channels occurring among the results are presented in a heatmap to enlighten the comparison. Following, the performance of the ML models utilizing the three subsets extracted from the channel importance analysis is presented.

Performance from several models with subsets of channels from the literature is presented for a comprehensive comparison of performance.

5.1 Segment Size Analysis

A segment size analysis was performed on three different segment lengths: 2 seconds, 5 seconds, and 10 seconds to ascertain the optimal segment size. The primary intent behind this examination was to determine whether shorter segment sizes could produce performance comparable to those of longer segments. Smaller segment sizes would potentially facilitate better learning of patterns from the minority class due to the increased number of instances. Furthermore, shorter segments could

lead to more immediate detection of transitions between different sleep stages. As Figure 5.1 shows, the 2-second segment size outperformed the other sizes. Consequently, this segment size was chosen as the standard for all subsequent experiments in this research.

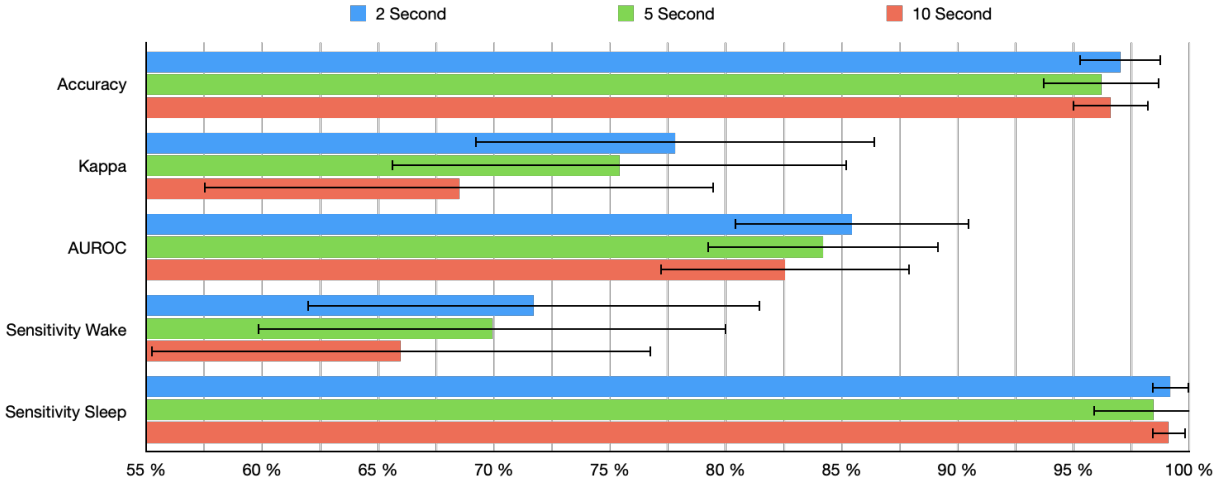


Figure 5.1: Comparison of performance with the use of different segment sizes for each epoch in the dataset. The best-performing segment size is 2 seconds, highlighted in blue.

5.2 Performance of EEGNet utilizing all 128 high-density EEG channels

This section presents the results derived from utilizing all 128 high-density EEG channels. The objective is to investigate if it is possible to obtain comparable results with a smaller but optimally chosen subset of channels. Utilizing the complete dataset of 128 high-density channels, the performance for all subjects in the dataset for sleep-wake classification is presented in Figure 5.2. These metrics represent the average computed from the training and evaluation of five distinct models. Each of these models employed unique divisions of training, validation, and testing subsets.

In line with expectations, the model accuracy does not sufficiently encapsulate the model performance as a single metric, even though it is the most commonly reported performance measure in the literature. As the main objective of the research is to detect and accurately classify instances of wake effectively, the sensitivity of wake is considered the most important metric. Since this is an isolated and class-specific metric, Figure 5.2 shows that the kappa score is the most comprehensive metric to capture the required performance nuances as a single performance metric for the overall model.

When utilizing all 128 channels, the model performance varied among subjects. The lowest performance was observed for Subject 3, followed by Subject 10. Subject 6 obtained the best results, followed by Subject 11. Notably, the model's overall accuracy was consistently high, exceeding 0.93 for all subjects. The isolated metric of wake sensitivity ranges more widely, from 0.55 to above 0.85. The isolated metric sensitivity of sleep is above 0.95 for all subjects. The full performance for all subjects is included in Table A.1 in Appendix A.

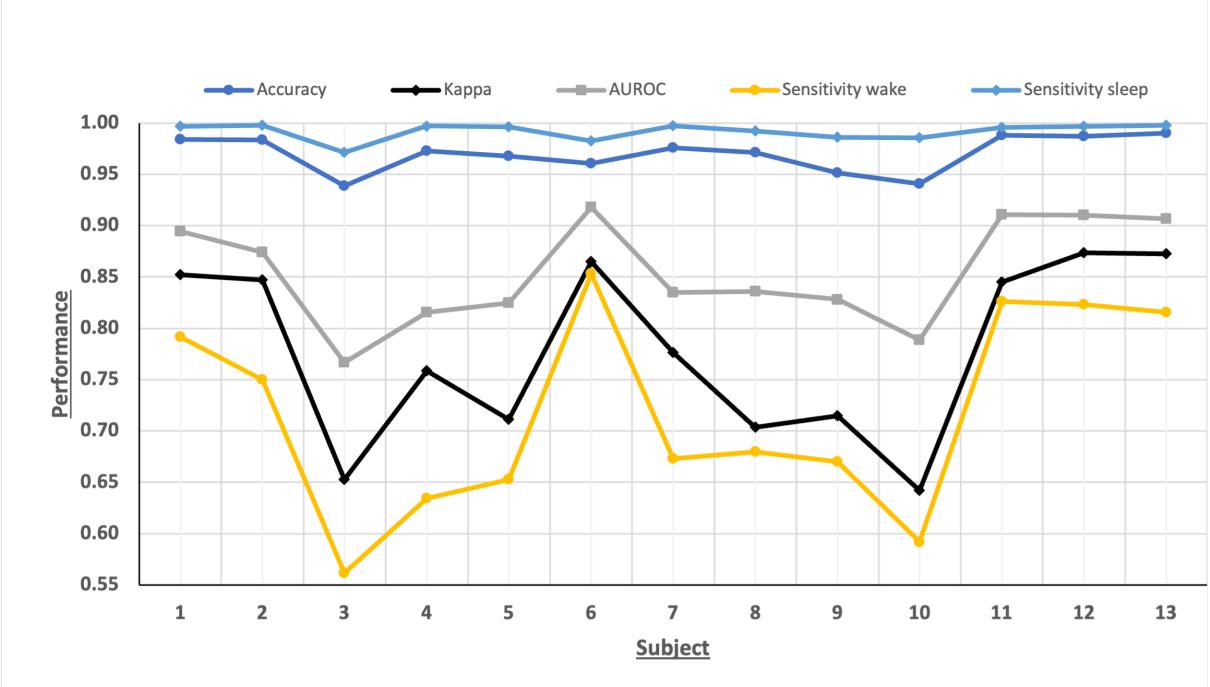


Figure 5.2: Performance of sleep-wake classification for all subjects using 128 high-density EEG channels

5.3 Selection of Subjects for In-Depth Analysis

While the experimental methodology was applied to all subjects, only four selected will be presented for in-depth analysis within this thesis. Investigating specific and isolated cases might reveal patterns or behavior that could be overlooked in the general and average analysis. Analyzing the individual subjects isolated makes the findings more relatable and tangible. Hence more nuanced discussion regarding specific instances is possible. It is important to note that only a subset of subjects has been chosen for this detailed discussion to maintain a coherent narrative and avoid unnecessary repetition. The full results for all subjects are available in Appendix A.

The selected subjects can demonstrate diverse characteristics such as different sleep patterns, different sleep distributions, and varying performance in the classification. To ensure a comprehensive representation of the dataset, the criteria for selecting the subset of subjects is based on the distribution of sleep stages, and the performance using all available 128 high-density EEG channels.

By examining Figure 4.3, it reveals that the highest wake distribution is found in Subject 6, which also obtained the best performance from the high-density experiment in Figure 5.2. In contrast, Subject 3 obtained the worst performance in the high-density experiment. Subject 11 obtained a similar performance for wake sensitivity compared to Subject 6, despite having the second-lowest wake distribution. Subject 9, on the other hand, has a high distribution of wake but shows significantly poorer performance compared to the other subjects. This concludes the following subset of subjects for in-depth analysis; Subject 3, Subject 6, Subject 9, and Subject 11.

5.4 Five-Class Sleep-Stage Classification based on 128 high-density EEG Channels

The dataset provides the labeling of five sleep stage classes. Utilizing all 128 high-density EEG channels, Figure 5.3 presents the overall performance for all subjects with five-class sleep-stage classification. The sleep stage obtaining the worst sensitivity is N1, which is often misclassified as another stage. The sleep stage obtaining the best sensitivity is REM. The full performance of all subjects is available in Table A.2 in Appendix A.

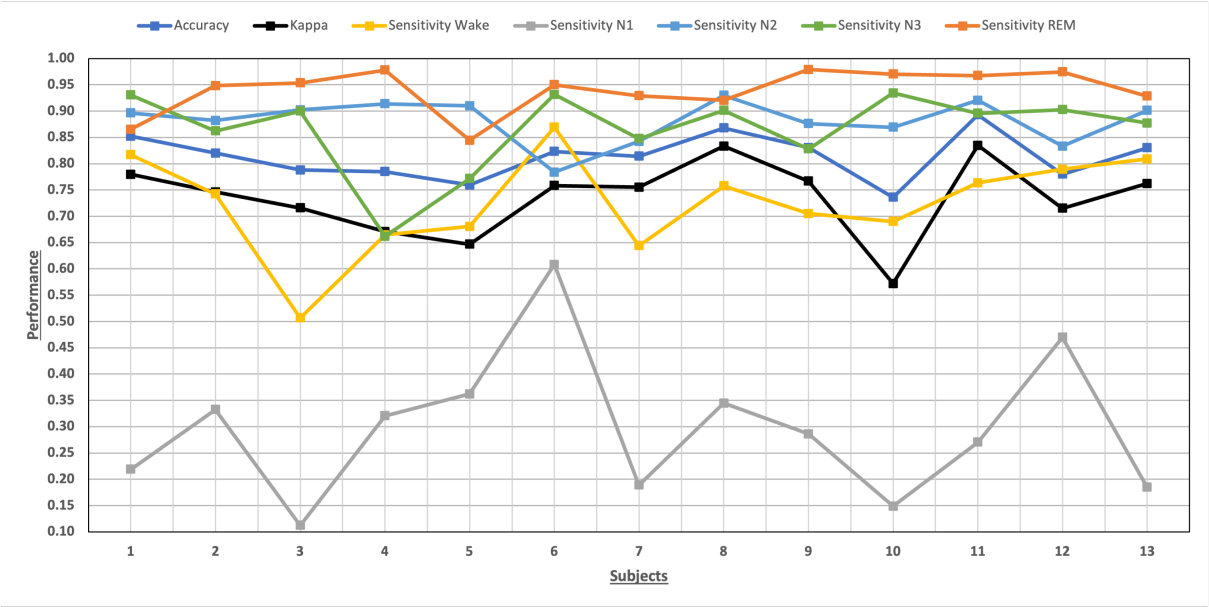


Figure 5.3: Performance of five-class sleep-stage classification for all subjects using all 128 high-density EEG channels

As mentioned, the objective of the conducted research is to detect and classify wake samples. A comparison of the ability to correctly detect wake samples through the sensitivity of wake with two-class sleep-wake classification compared to five-class sleep stage classification for the four chosen subjects is presented in Table 5.1. Subject 3 and Subject 9 obtained the best sensitivity of wake in the case of binary sleep-wake classification. In contrast, Subject 6 and Subject 11 obtained the best sensitivity of wake in the case of the five-class sleep stage classification.

Table 5.1: Sensitivity of wake comparison for binary sleep-wake classification and five-class sleep stage classification, utilizing all 128 high-density EEG channels.

	Sleep-Wake	**Five-Class**
Subject 3	0.562	0.507
Subject 6	0.853	0.869
Subject 9	0.670	0.705
Subject 11	0.826	0.764

5.4.1 Examples of Confusion Matrices for the Four Selected Subjects

Based on the five-class sleep-stage classification in Figure 5.3, a sample of the obtained confusion matrices is presented for each of the chosen subjects. Since the results in Figure 5.3 are based on the average among five executions, the confusion matrices presented are extracted from one of the executions. They may not exactly represent the illustrated value of performance.

P / R	Wake	N1	N2	N3	REM
Wake	161	20	92	1	11
N1	43	115	198	35	30
N2	39	74	929	83	58
N3	0	0	75	976	0
REM	10	0	3	0	730

(a) Confusion Matrix for Subject 3

P / R	Wake	N1	N2	N3	REM
Wake	508	68	14	0	0
N1	60	438	167	14	37
N2	0	85	863	55	13
N3	0	4	42	724	1
REM	0	6	25	0	608

(b) Confusion Matrix for Subject 6

P / R	Wake	N1	N2	N3	REM
Wake	263	26	71	14	1
N1	65	58	86	7	15
N2	15	18	1067	77	23
N3	0	0	204	913	1
REM	2	0	7	0	769

(c) Confusion Matrix for Subject 9

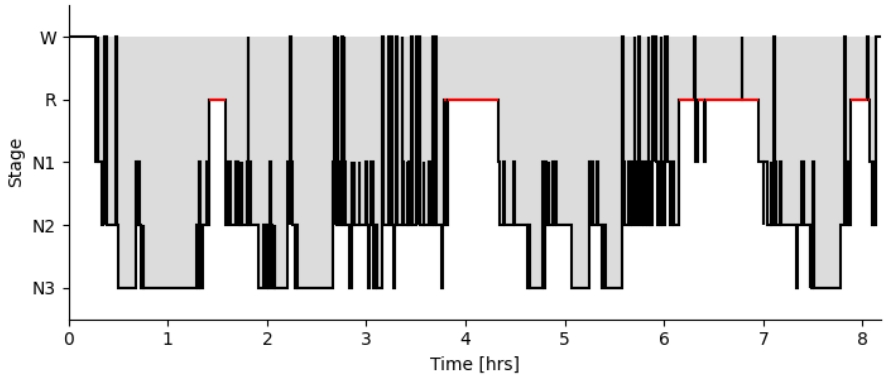
P / R	Wake	N1	N2	N3	REM
Wake	121	12	33	13	4
N1	28	74	58	2	33
N2	0	17	1042	43	58
N3	0	0	114	682	0
REM	1	0	10	0	1360

(d) Confusion Matrix for Subject 11

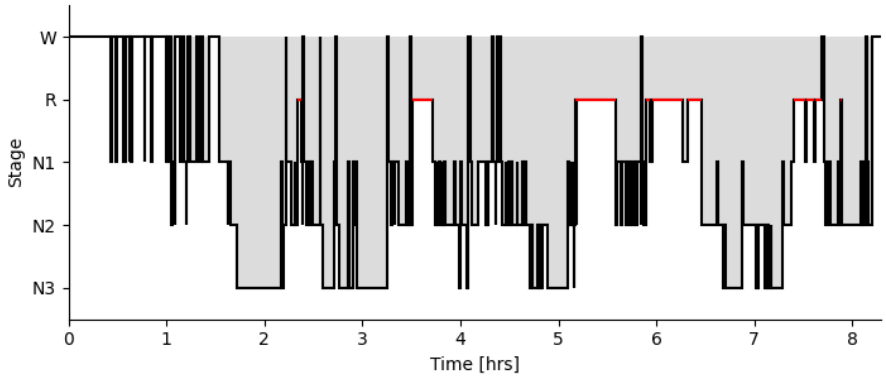
Figure 5.4: Confusion Matrices for the four chosen subjects for five-class sleep stage classification. P: Predicted class, R: Real class

5.5 Sleep Hypnograms for the Selected Subjects

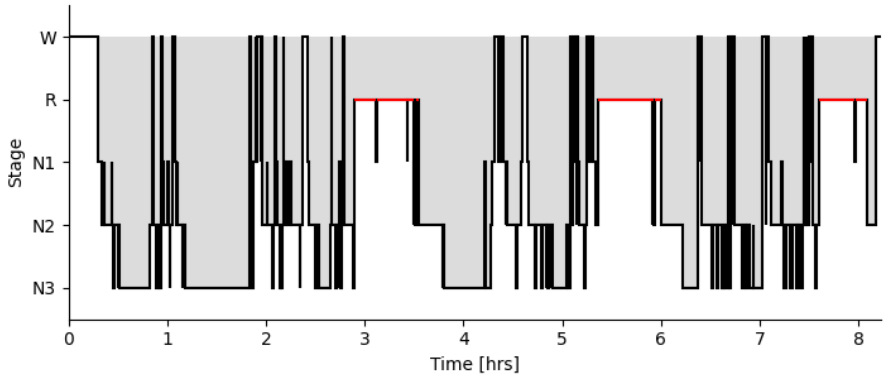
The hypnograms for the four chosen subjects are presented in Figure 5.5, and show varying characteristics. Subject 3 in Figure 5.5a represents the worst-performing subject using all 128 high-density EEG channels. The best-performing subject with the highest distribution of wake samples is Subject 6, presented in Figure 5.5b. Subject 9 has some longer wake periods and is present in Figure 5.5c while the high distribution of REM in Subject 11 is illustrated in Figure 5.5d.



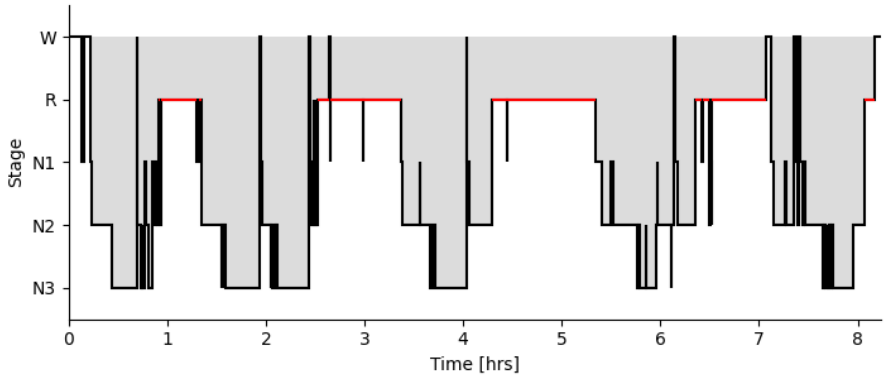
(a) Hypnogram for Subject 3.



(b) Hypnogram for Subject 6.



(c) Hypnogram for Subject 9.



(d) Hypnogram for Subject 11.

Figure 5.5: Hypnograms for the four selected subjects 3, 6, 9, and 11. Illustrating the sequence of sleep stages throughout their night of sleep.

5.6 NSGA-III Channel Selection Results

The channel selection optimization was executed using the NSGA optimization algorithm, an operation carried out individually for each subject. The resultant Pareto fronts offered a spectrum of solutions varying between four to eleven per subject. It should be noted that not every subject yielded NSGA results across the entire channel range from three to ten. Furthermore, a single subject could render multiple optimal solutions corresponding to the same number of channels. The average results among all subjects from the NSGA analysis are presented in Table 5.2. However, it is important to note that the averages reported are based on a varying number of subjects. In cases where a subject had multiple possible solutions for the same amount of channels, the best result was selected for inclusion in the analysis.

Table 5.2: Average NSGA results based on all subjects for different amounts of channels for sleep-wake classification. The average is based on the number of subjects presented in the last column.

Number of channels	**Accuracy**	**Kappa**	**AUROC**	**Based on number of subjects**
3	0.955 ± 0.023	0.678 ± 0.097	0.790 ± 0.075	5
4	0.966 ± 0.019	0.719 ± 0.100	0.811 ± 0.066	9
5	0.965 ± 0.020	0.714 ± 0.129	0.814 ± 0.081	12
6	0.970 ± 0.017	0.757 ± 0.101	0.832 ± 0.065	12
7	0.972 ± 0.014	0.781 ± 0.085	0.849 ± 0.060	12
8	0.973 ± 0.014	0.783 ± 0.075	0.855 ± 0.056	9
9	**0.975 ± 0.015**	**0.795 ± 0.075**	**0.855 ± 0.047**	8
10	0.974 ± 0.013	0.790 ± 0.067	0.853 ± 0.047	8

When using nine channels, all evaluated metrics slightly outperformed the results achieved with ten channels. The results are slightly decreasing from nine channels down to three channels. The difference between the maximum and minimum values for accuracy, kappa, and AUROC were 0.2, 0.117, and 0.065, respectively. The best results are obtained by the use of nine channels, with Accuracy, kappa, and AUROC of 0.975, 0.795, and 0.855, respectively. However, it is important to note that the averages based on 9 and 10 channels do not consider all subjects. The nine-channel average excludes subjects 2, 3, 6, 8, and 12, while the ten-channel average omits subjects 2, 4, 6, 9, and 13.

5.6.1 Channel Importance Across All Subjects

The NSGA analysis produced a set of potential solutions for each subject with different subsets of channels. The heatmap presented in Figure 5.6 highlights the frequency of occurrence for the most common channels from the NSGA analysis. The frequency was extracted from the entire subset of solutions for all subjects derived from the NSGA analysis.

The NSGA analysis identified channel D25 as the most relevant channel for sleep-wake classification, appearing in the set of optimal solutions of 12 out of all 13 subjects. Following, channels A19 and B14 were present in the sets of 10 subjects each, not in Subject 11. Subsequently, channels A14, B29, and D4 were present in 9 subjects, while channels B11 and B31 were present in 8 subjects. Given the frequent presence of these channels, they were considered the most relevant and formed the subsets presented in Table 5.7. Subset 1 includes the most common channel D25 together with the

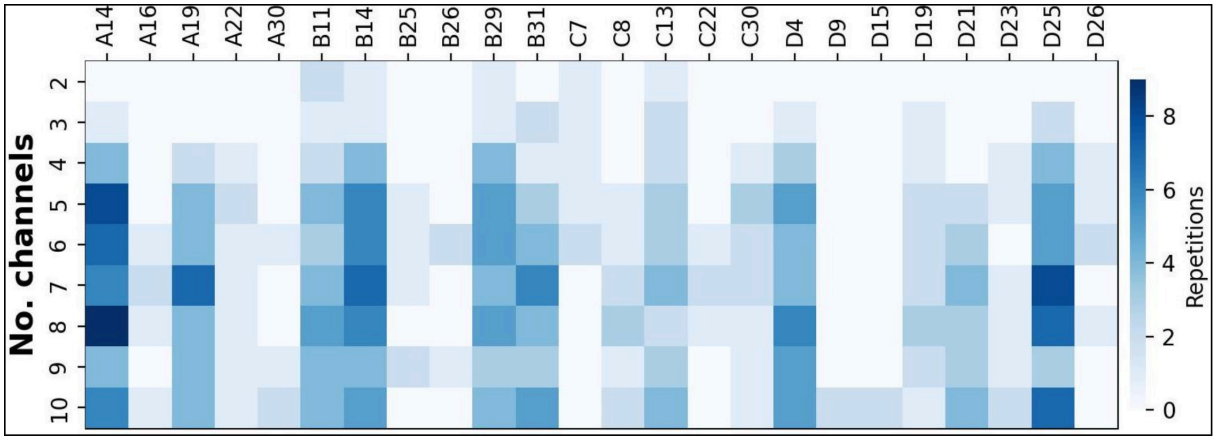
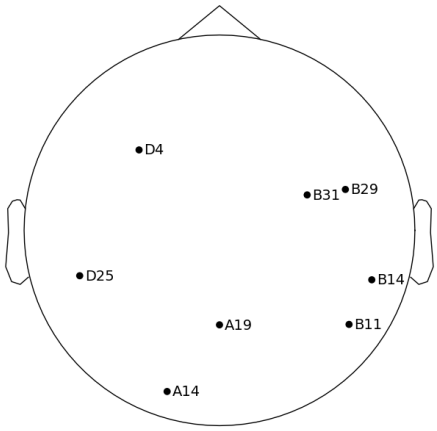


Figure 5.6: Heatmap of channel importance in sleep-wake classification based on the results from the NSGA analysis for all subjects.

two channels present in 10 subjects, A19 and B14. Subset 2 additionally adds the three channels present in nine subjects, channels A14, B29, and D4. Subset 3 additionally adds the two channels present in eight subjects, channels B11 and B31.



	Channels
Subset 1	D25, A19, B14
Subset 2	D25, A19, B14, A14, B29, D4
Subset 3	D25, A19, B14, A14, B29, D4, B11, B31

Figure 5.7: Positioning and subsets of channels obtained from the NSGA analysis.

The corresponding head-positioning of the channels from the NSGA subsets is presented in Figure 5.7. The most frequently occurring channel, D25, is located in the back left region of the brain. Channel B29 is located in the right frontal region together with B31. B11 is located in the right back region together with B14. Channel D4 is located in the left frontal region, and A14 is in the back center-left region. A19 is located in the back center of the head. This arrangement of important channels envelops a broad area of the head, ensuring coverage of most regions.

Frequency of Occurrence, Channels From the Literature

The frequency of occurrence for the channels extracted from the literature is presented in Table 5.3. The corresponding head-positioning of the subset of channels from the literature is visualized in Figure 5.8.

Table 5.3: Frequency of occurrence of channels from the literature extracted from the NSGA analysis

	Channels	Frequency of Occurrence
AASM	A28	0 subjects
	B22	0 subjects
	C4	1 subject
Bi-polar pairs	A19 - A23 (Pz- Oz)	10 and 3 subjects
	C17 - A1 (Fpz - Cz)	2 and 0 subjects
DREEM	D7	1 subject
	C7	2 subjects
	C29	1 subject
	A15	0 subjects
	A28	0 subjects

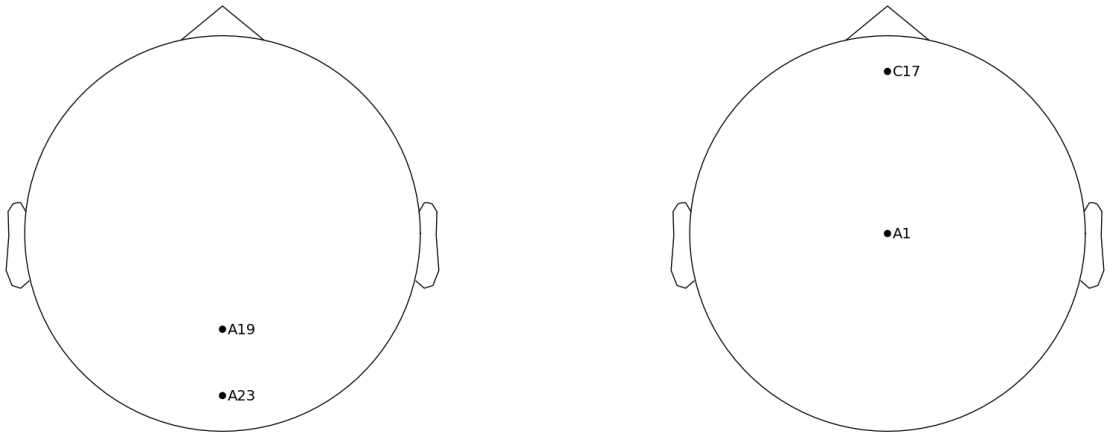
The AASM-recommended channels, A28 and B22, were not included in any optimal subsets, while C4 was present in one subject. For the subset of channels used in the DREEM device, channel C7 was present in two subjects, while channels C29 and D7 were present in one subject. The remaining channels, A15 and A28, were not present in any subjects. For the bipolar single channel pairs, A19-A23 was present in ten and three subjects, while C17-A1 was present in two and zero subjects.

5.7 Performance of Subsets of Channels from the NSGA-III Analysis

The NSGA analysis led to the identification of the three subsets of channels, as presented in Figure 5.7. Each of these subsets was then utilized to evaluate the performance of all subjects. Figure 5.9 presents the performance of all subjects when utilizing Subset 1. Similarly, Figures 5.10 and 5.11 illustrate the performance of the subjects based on Subsets 2 and 3, respectively. The full performance of the three subsets with all subjects is available in Appendix A. Respectivly, Subset 1 in Table A.7, Subset 2 in Table A.8, and Subset 3 in Table A.9.

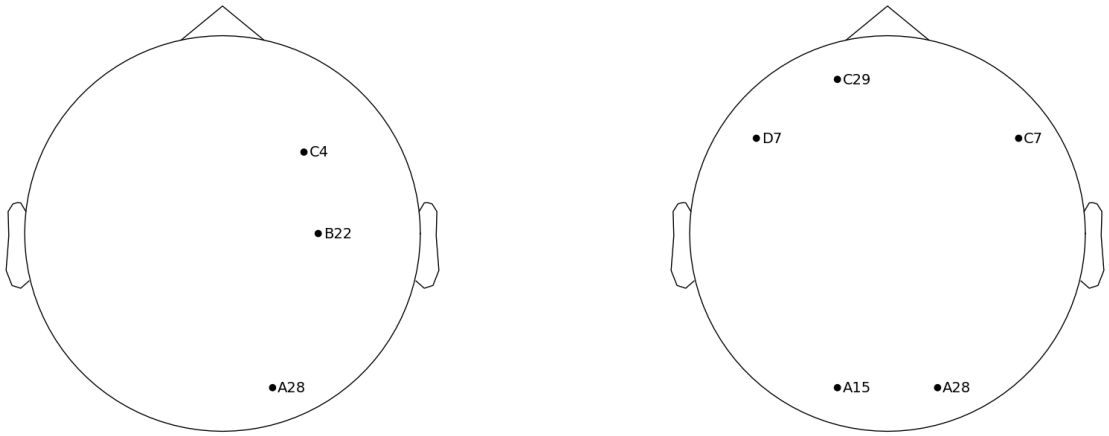
Examination of the results reveals several trends and noteworthy observations. Subject 6 consistently delivered the highest performance across all subsets. In contrast, the performance of other subjects varied considerably. Subject 9, for instance, had the lowest performance when utilizing Subset 1, with a wake sensitivity of just 0.15. However, this subject displayed a marked improvement in performance when evaluated using Subsets 2 and 3, even surpassing the performance of Subjects 3, 7, 8, and 10.

In the case of Subset 2, both Subjects 3 and 10 emerged as the worst performers, achieving wake sensitivities of less than 0.4. Most subjects demonstrated an upward trend in performance from Subset 1 to Subset 2 to Subset 3. An exception was observed for Subjects 3 and 13, with a slight decrease in performance from Subset 2 to Subset 3.



(a) Bipolar pair Pz-Oz

(b) Bipolar pair Fpz-Cz



(c) AASM-recommended channels [20]

(d) DREEM device [62]

Figure 5.8: Positioning of channels from the literature, including the AASM-recommended channels, the Bi-polar pairs Pz-Oz and Fpz-Cz, and the channels utilized in the monitoring device DREEM.

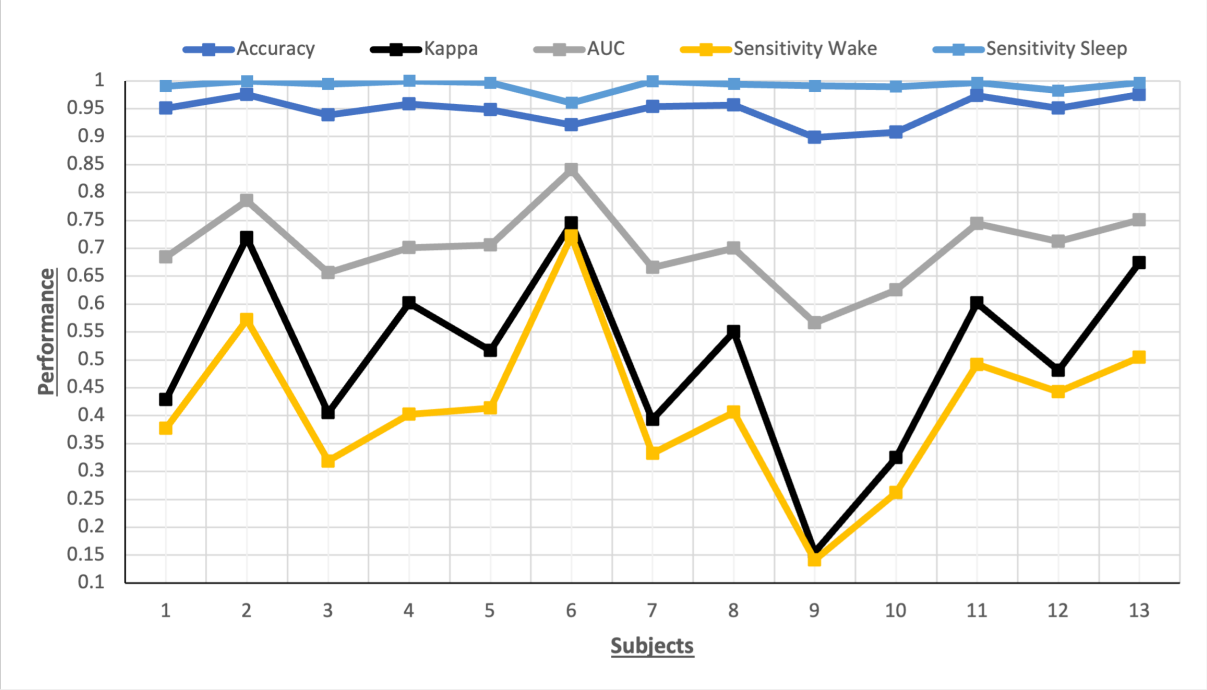


Figure 5.9: Performance of all subjects utilizing NSGA Subset 1.

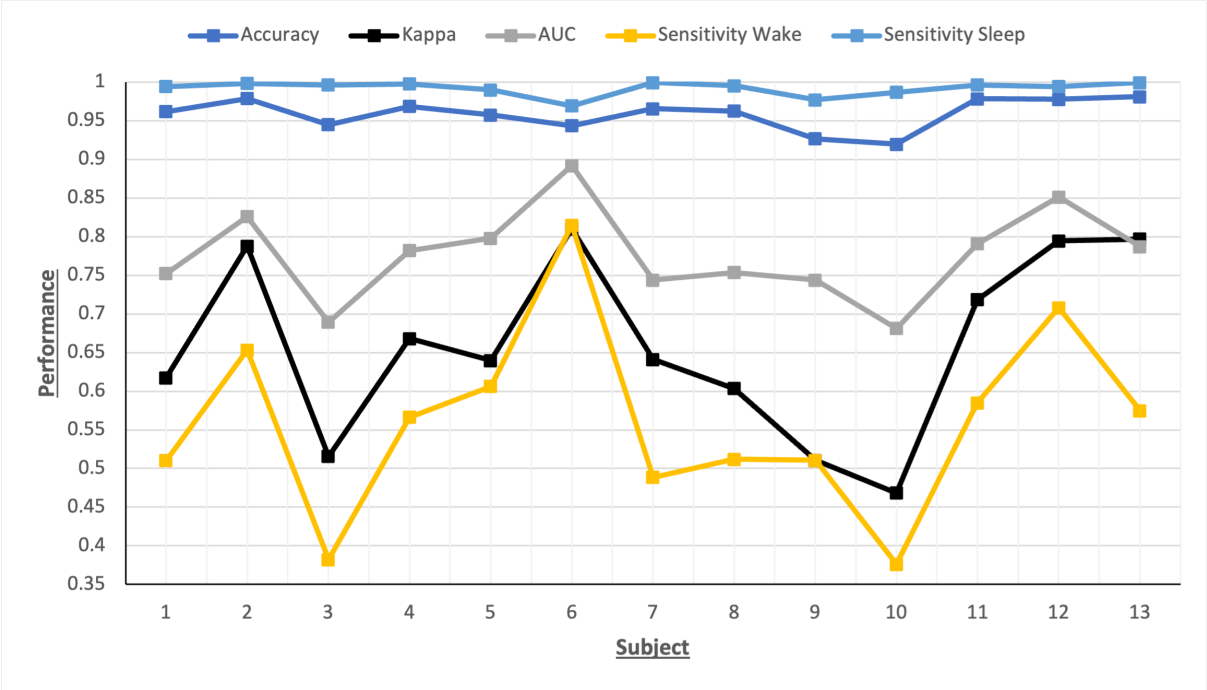


Figure 5.10: Performance of all subjects utilizing NSGA Subset 2.

Best Subset of Channels for Each Specific Subject

Following the NSGA analysis, a set of sub-optimal solutions was generated on the Pareto Front for each subject. From this selection, the optimal subset of channels corresponding to the highest-performing solution was extracted for each individual, consisting of eight to ten channels. Each subject's respective optimal subset was then employed to train and evaluate a unique model. The outcomes of these subject-specific models, utilizing the most favorable subsets determined by the

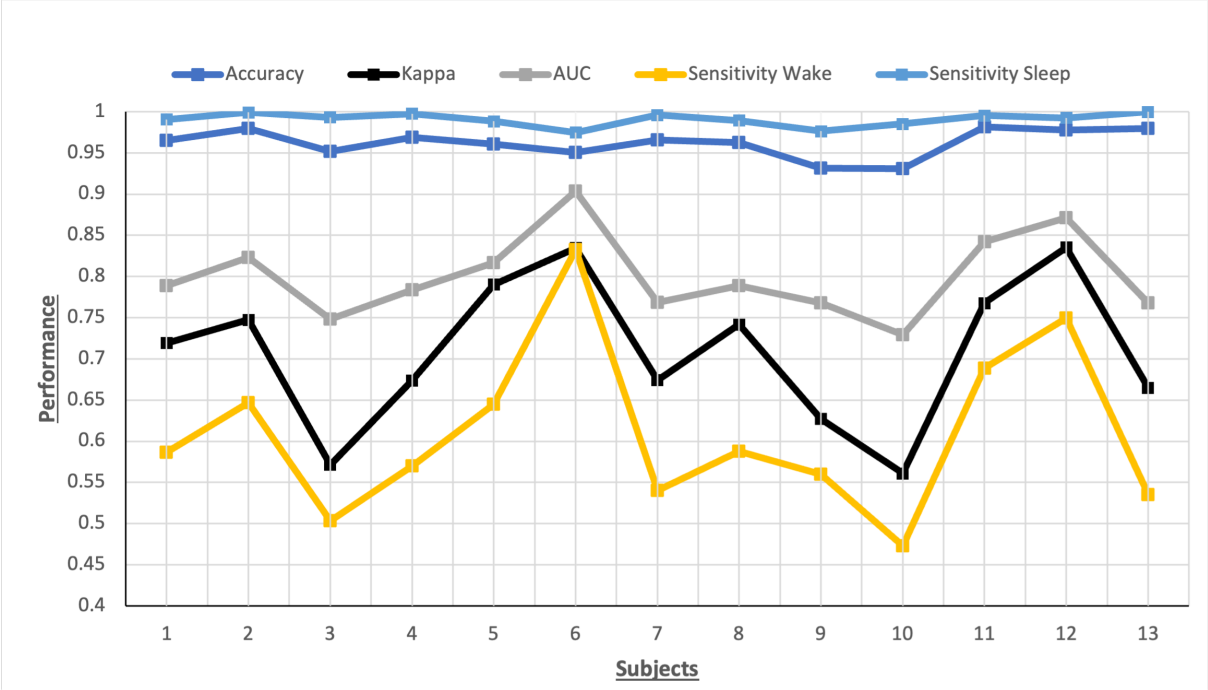


Figure 5.11: Performance of all subjects utilizing NSGA Subset 3.

NSGA analysis, are presented in Figure 5.12. The full performance for all subjects with corresponding best channels is included in Table A.10 in Appendix A.

By observation, Subject 10 appears as the worst performing subject with respect to the sensitivity of wake, below 0.5. This performance is closely related to Subject 3, with the lowest kappa value. Subject 6 is consistently the best-performing subject but is now closely followed by subject 13, with a sensitivity of wake above 0.8.

5.8 In-Depth Analysis of the Four Selected Subjects

The section presents a detailed analysis of selected subjects, aiming to draw a nuanced understanding that goes beyond the average trends.

A confusion matrix can reveal crucial information, considering the prevalence of either type-I or type-II errors. Using all 128 channels, Table 5.13 presents an instance of a confusion matrix for the four chosen subjects, Figure 5.13a for Subject 3, Figure 5.13b for Subject 6, Figure 5.13c for Subject 9 and Figure 5.13d for Subject 11. Note that the graphical representation from Figure 5.2 is based on the average of five executions. Conversely, the displayed confusion matrices are derived from a single execution and hence might not exactly correspond with the graphical results. The confusion matrices are presented to deepen the understanding of the classifier's performance. All matrices show that type-I errors, indicating false positives (false alarms), occur more frequently than type-II errors, which denote missed detections.

A comparison of the results obtained with the different subset of channels for Subject 3 is presented in Figure 5.14. The subsets from the NSGA analysis show an increasing trend. The AASM-recommended channels, in comparison, deliver poorer performance than both the DREEM channels and the NSGA

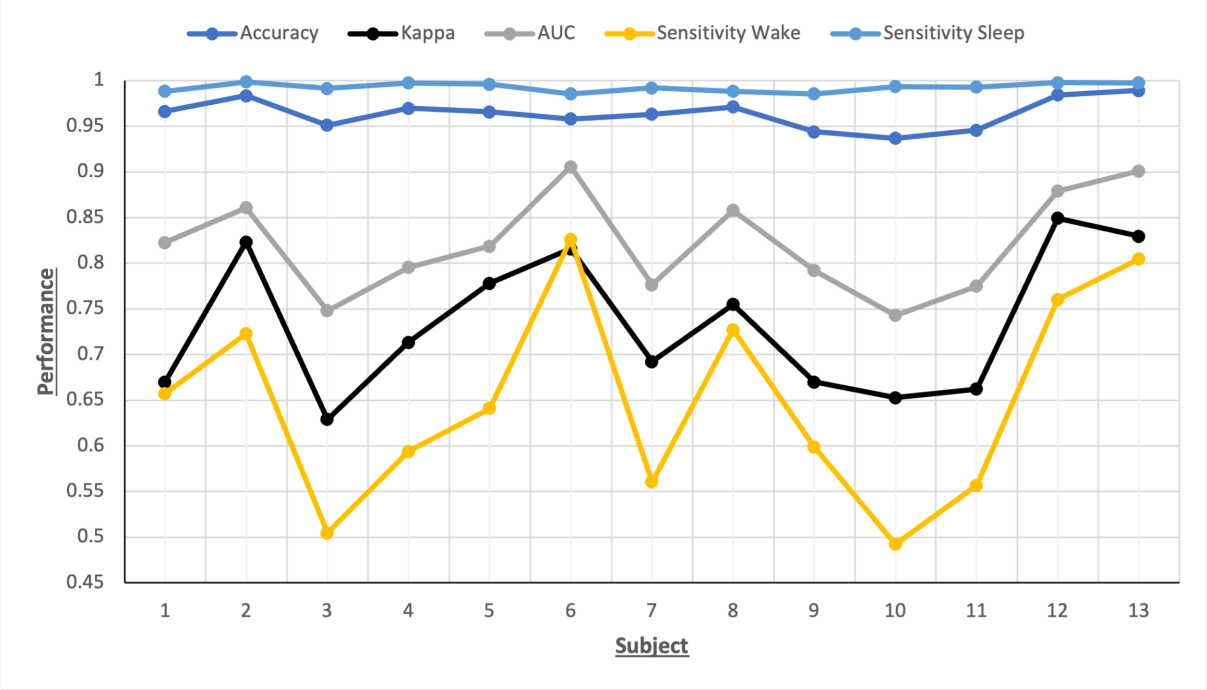


Figure 5.12: Performance of all subjects utilizing its corresponding best subset of channels from the NSGA analysis

P / R	Wake	Sleep
Wake	169	144
Sleep	19	3351

(a) Subject 3

P / R	Wake	Sleep
Wake	493	91
Sleep	37	3111

(b) Subject 6

P / R	Wake	Sleep
Wake	288	132
Sleep	43	3239

(c) Subject 9

P / R	Wake	Sleep
Wake	138	29
Sleep	19	3519

(d) Subject 11

Figure 5.13: Confusion matrices for the four chosen subjects for sleep-wake classification. P: Predicted class, R: Real class

subsets. The DREEM channels perform similarly to Subset 2. When relying on the bipolar single-channel pairs Fpz-Cz and Pz-Oz, results collapse significantly in terms of wake sensitivity and kappa score. In comparison with the other subsets, the kappa score is even lower than the sensitivity of wake, approximately zero. Such a low kappa score approximating zero implies that the degree of agreement is comparable to what would be expected by chance. Utilizing Subject 3s corresponding best subset from the NSGA analysis offers similar results as utilizing all 128 high-density EEG channels, but slightly lower but similar to the channels from Subset 3, although slightly lower yet comparable to Subset 3.

The results obtained with the different subsets of channels for Subject 6 are presented in Figure 5.15. The performance improves across the three NSGA subsets, although Subset 2 trails slightly behind Subset 3. The use of the AASM-recommended channels shows poorer performance compared to both the DREEM channels and the three suggested subsets from NSGA. The bipolar single-channel pair Pz-Oz shows the weakest performance but still remains above 0.6, significantly better compared to Subject 3. The Fpz-Cz bipolar pair outperforms the AASM-recommended channels, scoring above 0.7. The best subset from the NSGA analysis for Subject 6 performs similarly to the common Subset 3, just slightly lower than using all 128 channels.

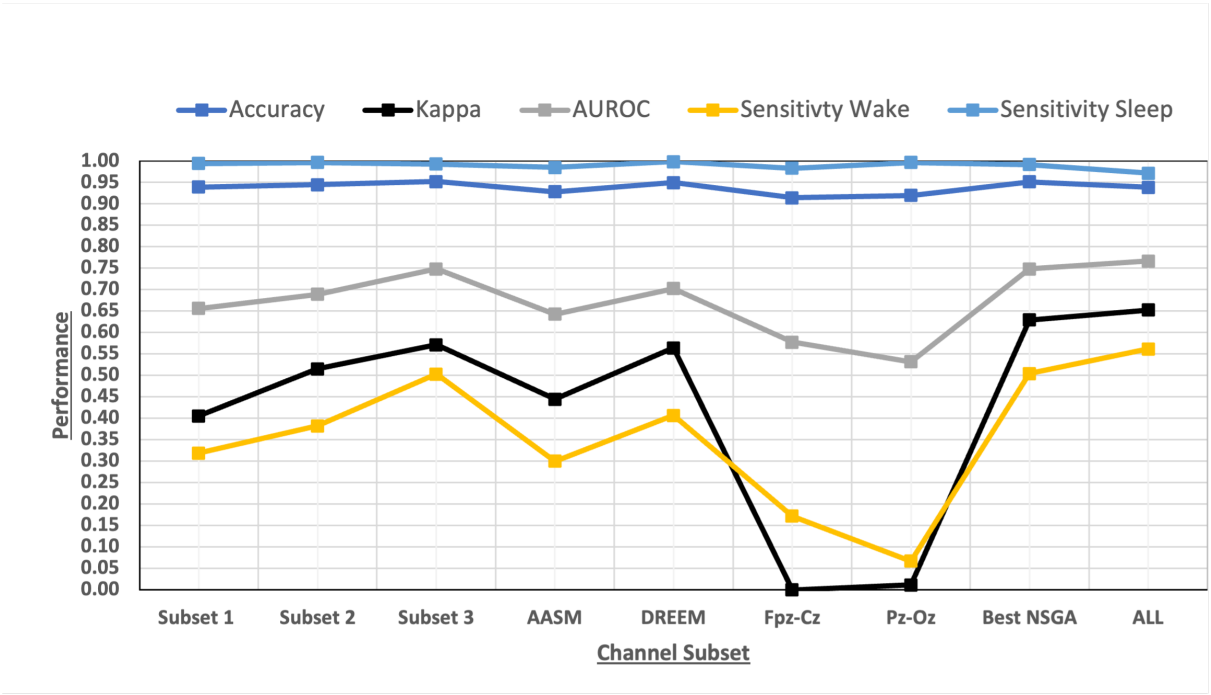


Figure 5.14: Performance of sleep-wake classification for Subject 3 based on different subsets of channels.

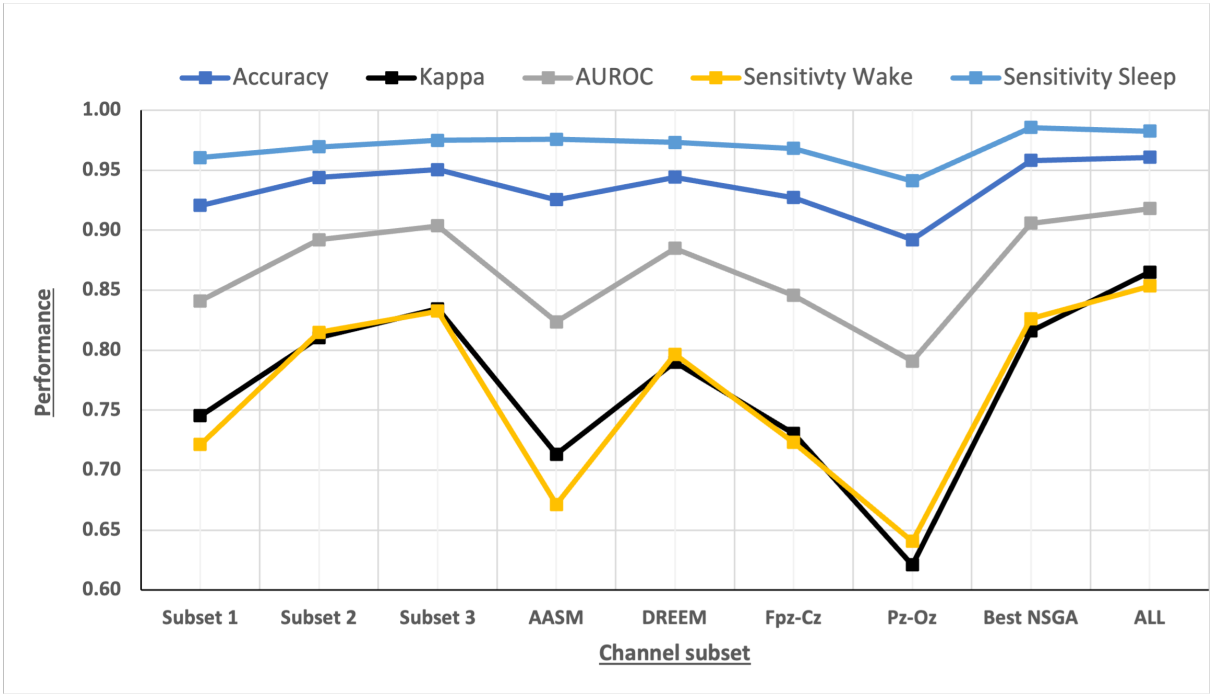


Figure 5.15: Performance of sleep-wake classification for Subject 6 based on different subsets of channels.

When examining Subject 9, Subset 1 yields a poor performance, as shown in Figure 5.16. Both the wake sensitivity and kappa score approximately 0.15. Subsets 2 and 3 from the NSGA analysis perform better and above 0.5, slightly below the best corresponding subset for Subject 9, and subsequently slightly below the use of all channels. The Fpz-Cz bipolar pair outperforms the AASM-recommended channel, both of which surpass the performance of the DREEM channels. The Pz-Oz bipolar pair achieves a similar wake sensitivity to Subset 1, yet the kappa score of 0 implies a level of agreement equal to chance.

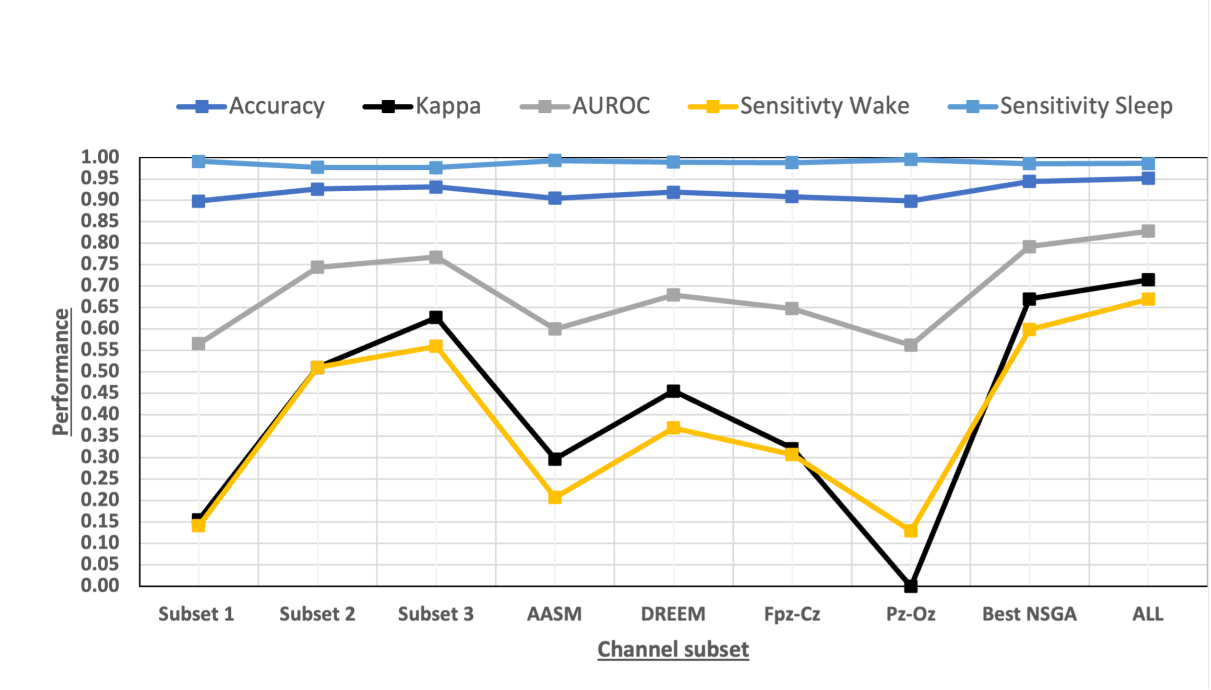


Figure 5.16: Performance of sleep-wake classification for Subject 9 based on different subsets of channels.

A comparison of the results obtained with the different subset of channels for Subject 11 is presented in Figure 5.17. Subject 11 does not seem to respond well to the two bipolar single-channel pairs, with both the kappa score and sensitivity of wake recording zero. This indicates that none of the wake instances in the test set were correctly identified. Subject 11's results diverge from the pattern seen in the other selected subjects. Both Subset 2 and Subset 3 from the NSGA analysis outperform the optimal NSGA channel subset for Subject 11 and the use of all 128 channels. The latter two achieve a similar level of performance, with a sensitivity of wake of around 0.55 and a kappa score of about 0.65.

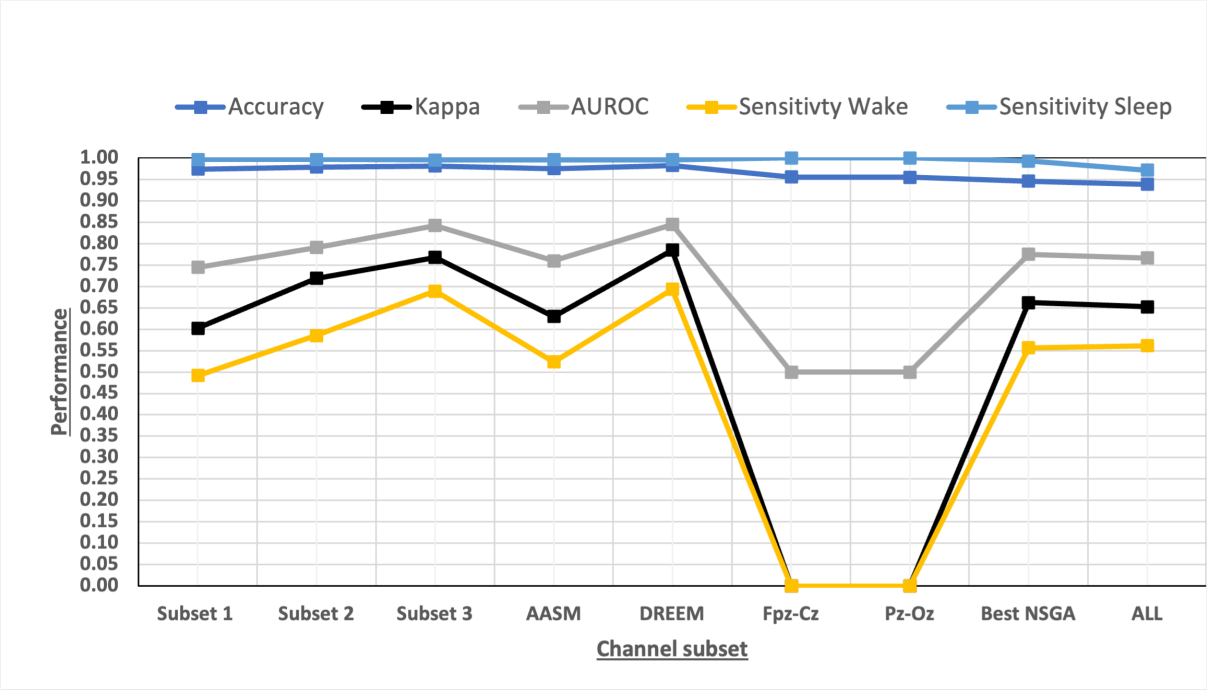


Figure 5.17: Performance of sleep-wake classification for Subject 11 based on different subsets of channels.

5.9 Comparison of Overall Results

Table 5.4 presents a comprehensive side-by-side comparison of average performance based on all thirteen subjects using various subsets of channels. 128 channels represent the use of all 128 high-density EEG channels. The next block consists of the subset of channels presented in the literature, starting with the AASM-recommended channels. The performance of the subset used in the DREEM device is presented together with the two bipolar single-channel pairs Pz-Oz and Fpz-Cz. Subsequently, the next block contains the performance utilizing the three different subsets created based on the results from the NSGA analysis. The last block presents the average performance when all subjects are trained on their corresponding best subset of channels from the NSGA analysis. Full performance for all subjects with the respective subset is available in Appendix A. Respectively; The DREEM subset in Table A.3, Fpz-Cz in Table A.5, Pz-Oz in Table A.6 and the AASM-recommended channels in Table A.4.

Utilizing all 128 high-density EEG channels obtained the highest performance in terms of accuracy, kappa, AUROC, and Sensitivity of wake. The closest subset of channels is the use of each subject's corresponding best subset of channels from the NSGA analysis, with a difference in accuracy, kappa, AUROC, and sensitivity of wake of 0.006, 0.044, 0.034, and 0.067, respectively.

Among the subsets, the bipolar single-channel pair of Pz-Oz registers the lowest performance, demonstrating a kappa score of 0.336 and a wake sensitivity of 0.290. From the subset of channels from the literature, the DREEM subset is the best performing, consistently displaying performance metrics between the NSGA Subset 2 and Subset 3.

Table 5.4: Average performance with all subjects for sleep-wake classification using different subsets of channels. The best NSGA corresponds to the average performance when all subjects are evaluated on its corresponding best subset of channels. This corresponds to the overall average from Figure 5.12.

Channels (No. of)	Accuracy	Kappa	AUROC	Sensitivity Wake	Sensitivity Sleep
128 channels	**0.970 ± 0.018**	**0.778 ± 0.086**	**0.855 ± 0.051**	**0.717 ± 0.098**	**0.992 ± 0.008**
3 AASM channels	0.951 ± 0.024	0.573 ± 0.149	0.716 ± 0.073	0.438 ± 0.146	0.994 ± 0.007
DREEM (5 channels)	0.959 ± 0.020	0.657 ± 0.134	0.777 ± 0.073	0.560 ± 0.148	0.994 ± 0.008
Pz-Oz (A19, A23)	0.940 ± 0.026	0.336 ± 0.290	0.641 ± 0.121	0.290 ± 0.250	0.992 ± 0.016
Fpz-Cz (C17, A1)	0.944 ± 0.020	0.414 ± 0.231	0.664 ± 0.095	0.336 ± 0.196	0.993 ± 0.010
NSGA Subset 1 (3 channels)	0.947 ± 0.024	0.507 ± 0.161	0.703 ± 0.067	0.414 ± 0.139	0.992 ± 0.010
NSGA Subset 2 (6 channels)	0.959 ± 0.020	0.652 ± 0.111	0.771 ± 0.059	0.549 ± 0.120	0.992 ± 0.009
NSGA Subset 3 (9 channels)	0.962 ± 0.016	0.708 ± 0.086	0.800 ± 0.048	0.609 ± 0.097	0.991 ± 0.008
Average: Best NSGA	**0.964 ± 0.016**	**0.734 ± 0.078**	**0.821 ± 0.055**	**0.650 ± 0.111**	**0.993 ± 0.005**

Chapter 6

Discussion, Conclusion and Further Work

This chapter aims to highlight important observations from the obtained results. Furthermore, potential weaknesses are pointed out together with implications for future research.

6.1 Importance and Challenges of Wake Detection

Experiencing frequent awakenings and periods of wakefulness during the night significantly impacts the quality of sleep. Sleep disorders such as insomnia or sleep apnea are characterized by frequent awakenings and longer periods of wakefulness during the night. It is critical to obtain reliable detection of wakeful periods to be able to discover and diagnose these conditions or give an indication of the person's quality of sleep to be able to perform the necessary precautions.

The alterations in sleep patterns due to sleep disturbances can complicate wake detection. For instance, for insomnia patients, frequent awakenings could lead to increased transitions between sleep and wake, requiring a wake detection method capable of accurately identifying these transitions. By utilizing small segment sizes, it is possible to better detect these transitions, increasing the detection of wake samples. With respect to patients suffering from sleep apnea, misclassification between light sleep stages and wakefulness might be present due to similarities in EEG signals. It might be expected that these samples are the ones most often misclassified in sleep-wake classification as well.

Effective detection of wakefulness and estimation of sleep quality can help in assessing effective treatment for such sleep disturbances. For instance, in treating sleep apnea, the goal is often to reduce the number of awakenings. Accurate detection of wakefulness can indicate whether the treatment is working or if another one should be applied.

Given the increasing amount of sleep disorders, accurate identification and estimation of sleep quality are important to obtain insight. Sleep disturbances or chronic sleep deprivation can lead to severe health repercussions, such as excessive daytime sleepiness, diminished cognitive function, and a range of adverse health outcomes, including cardiovascular diseases, diabetes, and depression. Increased monitoring or tracking of sleep duration and patterns can give greater insight into what cau-

tions should be taken to improve sleep quality and counteract the unfortunate side effects associated with inadequate or poor-quality sleep.

6.2 Potential Limitations Due to Dataset Representativeness

The WIP-IIIS dataset utilized in this study, while rich in information, does present some limitations in terms of its demographic scope. Notably, the age range encapsulated by the dataset is quite narrow, varying only by a single year. As a result, the lack of age diversity may not fully reflect the substantial alterations in sleep patterns that can occur across different age groups. Therefore, it would be beneficial to assess the performance and generalizability of the proposed models using data derived from a more diverse age demographic.

Moreover, the dataset comprises data from just thirteen subjects. While this sample provides a starting point for analysis, it may not be representative enough of the varied sleep patterns found among a broader population. For instance, comparing sleep patterns between healthy individuals with consistent sleep patterns and individuals suffering from insomnia, who experience frequent nighttime awakenings, could yield interesting insights. Such diversity could help enhance the robustness of the models and their applicability to different individual sleep scenarios.

6.2.1 Class Imbalance in the Dataset

The decision to work with imbalanced data was in the context of more accurately reflecting real-world sleep-wake cycles, characterized by significantly more sleep periods than wake periods. Working with this natural imbalance offers insight into how the proposed models and subset of channels perform under realistic conditions.

The class imbalance problem may lead to developing bias towards the majority class in the performance of the ML models. Therefore, exploring various sampling techniques might provide an understanding of how the imbalance affects the classification between sleep and wake. Such techniques might include oversampling the minority class or undersampling the majority class. These approaches might augment computational overhead in a real-time device, but might also increase performance.

However, an alternative and perhaps more robust solution are to naturally balance the dataset by collecting and including more wake samples. Unlike the synthetic balance achieved through undersampling or oversampling, this method contains real data. The process of undersampling risks excluding important information or nuances from the data, which might be critical for understanding patterns and correlations in the sleep data. Oversampling artificially increases the number of instances of the minority wake class by replicating existing instances, which could lead to overfitting. By intentionally collecting and incorporating more wake samples, the need for these artificial techniques can be avoided, resulting in a more naturally balanced dataset. This approach provides a richer and more representative view of sleep stage patterns and holds greater promise for practical applications in real-time sleep stage monitoring devices.

The methodology employed in this research embraced a subject-by-subject approach. The data from one night of sleep for one specific subject were split into training, validation, and test sets. The size

of the test set consists of 25% of the data. Given that the distribution of wake varies between 4-16 %, the splitting of the dataset includes a risk of only getting a few, or in the worst case, no wake samples within certain subsets. Such an occurrence could significantly influence the performance metrics of the models.

6.3 Performance Utilizing All 128 High-Density EEG Channels

The uniqueness of 128 channels of high-density EEG signals has the possibility to identify patterns that may be missed with the use of fewer data and fewer channels. 128 electrodes located across the scalp gives a greater range of data collection points, which may lead to more comprehensive and robust results. Since this model utilizes the maximum amount of information, possibly identifying all important and present patterns, this works as a benchmark model for further analyses and comparisons. On the other hand, training ML models with the use of high-density data require significant computational resources and time.

With respect to Figure 5.2, all subjects perform above 0.55 while the best-performing Subject 6 obtains above 0.85 for the sensitivity of wake. Subject 6 has the highest distribution of wake samples, highlighting the impact of the unbalanced dataset on the performance. However, this would imply an expected performance in the same pattern as the distribution of wake samples in the dataset, which is not the case. The worst-performing subject is Subject 3, with 8.5% wake samples, and the second-worst performing is subject 10, which has the third-highest distribution of 10.9%. On the other hand, Subject 11 is the second best-performing subject and has the third-lowest distribution of wake samples. This shows that the distribution of wake is not a direct match with the expected performance.

6.3.1 Analysis of Five-class Sleep Stage Classification Using 128 High-Density EEG Channels

With the access of high-density EEG data and five-class labeling data, the performance of the full five-class sleep stage classification using all 128 high-density EEG channels was presented in Figure 5.3. Stage N1 emerged as the most challenging sleep stage for accurate detection and classification by the proposed ML model.

By inspecting Table 5.1, both Subject 6 and Subject 9 obtained higher sensitivity of wake in the five-class sleep stage classification compared to the binary sleep-wake classification. This could be a result of the more evenly distributed dataset across the five classes, as opposed to the imbalanced distribution in the binary sleep-wake classification. When all sleep stages are consolidated into a single class, this could significantly increase the internal variability of that class. It could make it more challenging to distinguish wakefulness from similar wave patterns exhibited during sleep stages, particularly those like stage N1 that closely resemble the state of wakefulness. Such increased internal variability might, in turn, potentially impact the performance of the classification models.

The confusion matrices in Figure 5.4 present additional detail. For all subjects, the samples of wake that are not detected are most commonly classified as N1, representing the type-II error. Considering the Type-I error, when a sample is predicted as wake when it is not, it is most commonly classified

as N2, except for Subject 6, which is misclassified as N1. Subject 6 is considered the most balanced subject with its high distribution of wake samples, while the other subjects have N2 as their majority class. These misclassifications could be due to both the overrepresentation of N2 samples and similarities in sleep patterns. The corresponding confusion matrices for all chosen subjects is present in Figure 5.4a for Subject 3, 5.4b for Subject 6, Figure 5.4c for Subject 9 and Figure 5.4d for Subject 11.

Furthermore, the recurring misclassification involving stage N1 across all subjects could potentially be due to the similarities in wave patterns, given that N1 is a transition phase between wakefulness and sleep. The low rate of misclassification between wake and REM stages, both stages with similar high-frequency, low-amplitude activity, indicates promising performance by the classifier. This highlights the classifier's ability to distinguish between nuanced differences in EEG signals, despite apparent similarities.

6.4 NSGA-III Channel Selection Results

The averaged NSGA results, presented in Figure 5.2, detail the outcomes of sleep-wake classification for different numbers of channels. Using nine channels yields the best results, even slightly higher than the average results obtained from using all 128 high-density EEG channels. However, these numbers are based on different amounts of subjects. The nine-channel average is based on eight subjects and excludes subjects 2, 3, 6, 8, and 12. Subject 3 has consistently decreased the average with its poor performance, but on the other hand, Subject 6 has consistently been increasing the average performance.

6.4.1 Performance of Subsets of Channels Based on the NSGA-III Analysis

The best-obtained results from the presented subsets of channels are the model for each subject based on their corresponding optimal subset of channels from the NSGA analysis. As anticipated within the optimization framework, this outcome is promising. Furthermore, the exceptionally low deviation between the optimized performances and the benchmark performance from the 128 high-density channels is notably encouraging, with minimal differences in average accuracy, kappa, AUROC, and sensitivity of wake being 0.006, 0.044, 0.034, and 0.064, respectively.

Subset 1 of the NSGA analysis does not appear to provide optimal or sufficient information, given its lower performance compared to the three AASM-recommended channels. The subset consists of the three most occurring channels from the NSGA analysis, present in at least ten subjects. These results might indicate limitations to the proposed extraction method. Although Subject 6 achieves a sensitivity of wake exceeding 0.70 with Subset 1, the overall average was only 0.414. This might indicate that improvements in the optimization algorithm should be considered.

An incremental performance pattern is observed across the three subsets, with Subset 3 outperforming Subset 2, which again outperforms Subset 1. This suggests that a modest increase in the number of optimal channels can marginally increase the model's overall performance, revealing an important balance to consider between obtained performance and computational expense.

It is noteworthy that these subsets of channels demonstrated comparable or superior performance

to the well-established subsets from the literature. An important consideration is that these performances could potentially be improved further by incorporating additional parameters in the optimization of the chromosome. This could involve the deep learning model, the application of filters, or adjustments to other parameters.

A crucial point to consider is that the creation of these subsets is predicated on single-channel extraction, determined by the frequency of occurrence among subjects. This approach, however, might overlook the collaborative interaction between certain channels, which could contribute more effectively when grouped together than when functioning as standalone channels. This aspect is not considered in the current subset creation method. Ideally, future iterations would involve extracting subsets based on a general model based on all subjects, with greater variability. Although such an approach would require additional computational resources, it could provide a more comprehensive and accurate representation of important channels and channel interactions with respect to sleep-wake classification.

6.5 Performance of Channel Sets from the Literature

When comparing the achieved results with the leading state-of-the-art, it is evident that they do not reach the same level. However, the underlying assumptions behind these results necessitate careful consideration. The current state-of-the-art research often employs publicly available and fairly balanced datasets, occasionally tending towards a higher distribution of wake samples compared to sleep samples. This significantly contrasts the distribution of the dataset utilized in this research. The performance of the presented model, when using subsets of channels from the literature on the presented dataset, clearly underscores the influence of these assumptions. Although the performance using these subsets varies, all of them result in poorer performance than utilizing all 128 channels, which is expected due to the high dimensionality of all the 128 high-density EEG channels. The performance is also poorer than their corresponding reported performance in the state-of-the-art.

The AASM-recommended channels were barely present in the results of the channel importance analysis. This might indicate that different channels are of main importance when considering wake detection in sleep-wake classification compared to full five-class sleep stage classification. Alternatively, it may indicate that these channels do not provide sufficient information as standalone channels but work better together in collaboration. This reasoning also applies when considering the performance of the bipolar single-channel pairs. Fpz-Cz consistently outperforms Pz-Oz, even though the channels of Pz-Oz were more frequently occurring in the channel importance analysis. This suggests that these pairs work better together compared to standalone channels.

The channels used in the sleep-monitoring device DREEM were barely present in the results of the channel importance analysis. Still, they provided better results than the AASM-recommend channels, NSGA Subset 1, and the two bipolar single-channel pairs. This might indicate that the creators of the device have found a subset of channels working efficiently together, yielding promising results.

Comparing the positioning of the channels, the bipolar single-channel pair Pz-Oz is located in the back center of the brain, while Fpz-Cz is located between the center and the front of the brain. The AASM-recommended channels are positioned on the right side of the brain. The channels corre-

sponding to the DREEM device are scattered across the whole head. The channels from the NSGA analysis (see Figure 5.7) are also scattered across the whole head. This might indicate that the significance of specific brain regions for may not be limited to one single region sleep-wake classification, but rather multiple regions could be important or contribute different information. This may indicate that no single brain region is of significant importance alone with respect to sleep-wake classification. It might also imply that channel subsets spread across the brain may provide a more comprehensive understanding of wave patterns from the entire head, which could be important for accurate sleep-wake classification.

6.6 Selected Subjects for Detailed Analysis

Performance across different subsets of channels for individual subjects offers a deeper level of analysis and a richer comparison of varied outcomes. Examples of confusion matrices associated with sleep-wake classification utilizing all 128 high-density EEG channels are depicted in Figure 5.13. The data indicated that Type-I errors consistently occur more frequently than Type-II errors. As previously stated, the main objective is to detect wake samples. It is considered more critical with Type-II errors as this indicates a wake sample that is not detected, compared to the false alarm of Type-I errors. In a clinical context, a false alarm typically prompts further patient investigation, while a Type-II error might result in an undiagnosed condition.

Subject 3 demonstrates a relatively balanced distribution between the various sleep stages, as shown in Figure 4.3 and Figure 5.5a. Although frequent awakenings are present, Subject 3 also spends a substantial duration within each sleep stage. Regarding performance, the subject was consistently the worst performing subject in both sleep-wake and five-class sleep stage classification, using all 128 high-density EEG channels. This subject also consistently ranked among the poorest performers for all NSGA subsets. It was surprising to note that NSGA Subset 2 outperformed Subset 3. The wake distribution for Subject 3 was 8.4%, a middle-range value across all subjects. The poor performance might be due to the frequent transitioning between each sleep stage. Utilizing all 128 channels yields a sensitivity of wake below 0.60, suggesting that the proposed classifier struggles to accurately classify this subject. The bipolar channel pairs seemingly fail to gather sufficient information, as the classifier cannot detect wake samples using these channels.

In contrast, Subject 6 consistently outperformed other subjects across all channel subsets, highlighting the impact of a high wake distribution due to prolonged sleep latency. The hypnogram of Subject 6, as shown in Figure 5.5b, reveals a scenario where the patient stays awake for about half an hour before alternating between N1, light sleep, and awake for approximately an hour before falling asleep. This pattern explains the subject's high proportion of wake samples compared to other subjects. The use of Subset 2, Subset 3, DREEM channels, the best NSGA channels, and all 128 channels resulted in a minimum wake sensitivity of 0.8, signifying robust classifier performance. This indicates that including more wake samples that are not recorded in between the stages, but specifically before the sleep period might increase the performance of the classifier. In the context of at-home sleep-monitoring devices, patients could be instructed to wear the recording device for an hour prior to bedtime. This would allow for the recording of wake samples to naturally increase the distribution of wake and potentially improve performance.

However, a high distribution of wake samples does not ensure better performance. Subject 9, despite having the second-highest distribution of wake samples, consistently ranks among the worst performers. Frequent awakenings throughout the night, as seen in Figure 5.5c, could explain this poor performance. Conversely, Subject 11, which holds the lowest distribution of wake samples (refer to Figure 5.5d), outperformed expectations, ranking as the third-best subject when utilizing NSGA Subset 3. This evidence emphasizes that a higher distribution of wake samples does not necessarily equate to improved performance.

These disparities in performance could be attributed to the consistency and continuity of wake periods for each subject. Despite having a high overall wake distribution, the more fragmented wake periods for Subject 9 might make these periods more challenging to detect, with frequent sleep stage transitions. On the contrary, Subject 11 exhibits a more regular sleep pattern with less frequent awakenings. It tends to remain in the same stage for extended periods before transitioning, which could enhance classifier detectability.

It's important to note that such variability might be due to differences in the quality of EEG signals for each subject or differences in brainwave patterns. A comparison of the hypnograms of each subject further accentuates the diversity of sleep patterns across individuals. Even though none of these hypnograms align perfectly with the idealized hypnogram depicted in Figure 2.7, the sleep pattern of Subject 11 appears to be more closely related to the idealized pattern, with fewer awakenings and extended periods of REM sleep.

6.7 Limitations

The research presented has several limitations which should be taken into account when evaluating the findings. These constraints are due to various reasons, ranging from the choice and tuning of hyperparameters to the representativeness of the dataset and the objectives of the optimization process.

When working with ML, the selection and optimization of hyperparameters play a significant role in determining the model's efficiency, as several hyperparameters have to be set before training and evaluating the model. The hyperparameter optimization process can be complex due to the multitude of parameters and the high dimensionality in finding an optimal set and combination. In this research, some parameters were chosen prior to training, while others were chosen based on evaluating the best performance. Some of such hyperparameters refer to the number of epochs in the ML model, the distribution of size between the training, validation, and test set, and the number of generations in the optimization algorithm. Other combinations between the hyperparameters might increase performance.

Additionally, this research performs each experiment utilizing a subject-by-subject approach. An ideal scenario would involve creating a general model that encapsulates all subjects. A general model would give the ML model more data to learn characteristics and patterns from, making it more robust and relevant for subjects on a general basis. Moreover, it would allow the NSGA optimization to identify a common subset of crucial channels across all subjects. However, this approach would require comprehensive computational power and storage.

With only thirteen subjects with a relatively restricted age range, there is a limited representation of the variability of individual sleep patterns. A more comprehensive dataset, including individuals of different ages, lifestyles, and underlying health conditions, could yield a more insightful understanding of sleep pattern variability. Combined with a general model, the model might obtain higher robustness and generalizability.

Furthermore, the NSGA analysis is performed by optimizing four objectives: maximize accuracy, kappa, and AUROC while minimizing the number of EEG channels. However, other optimization objectives could potentially result in different outcomes. Other potential objectives of interest could include maximizing the sensitivity of wake or minimizing Type-I or Type-II errors.

Lastly, to better evaluate the performance of the ML model for the specific dataset, the model could benefit from being trained and evaluated on a publicly available dataset for comparison. Utilizing such a dataset would better indicate how the actual model performs on balanced and preprocessed data before applying the model to the real-world dataset.

6.8 Conclusion

With raw data from a highly unbalanced dataset, the proposed research has shown that it is possible to obtain promising performance for sleep-wake classification with deep learning methods, answering the first research question. The findings suggest that an increased distribution of wake samples gathered prior to sleep onset, as exemplified by Subject 6, might increase the performance of sleep-wake classification. An essential challenge for the classification lies in the relatively low distribution of wake samples, along with frequent and brief awakenings throughout the night.

In response to the second research question of the thesis, the proposed research successfully shows the possibility of achieving comparable sleep-wake classification performance with a substantially reduced subset of channels compared to the complete set of 128 high-density channels. While utilizing the full set of 128 high-density EEG channels brings computational complexity, this research highlights that subsets comprising fewer than ten channels can achieve similar performance in wake detection.

Notably, the results indicate that selecting a single universal set of channels may not yield optimal performance for every individual. Further research might consider customizing the channel selection process for each subject, aiming to maximize performance in sleep stage classification using EEG signals. This personalization becomes critical due to the individual variability of sleep patterns, as highlighted in this research. A unique subset that suits the entire dataset might be obtained by deploying a generalized model incorporating all subjects collectively.

Furthermore, this study demonstrates the potential existence of different subsets that outperform those suggested by the AASM or existing literature, mainly when considered on a subject-by-subject basis. Other subsets than the ones provided in this thesis might obtain even better performance. The research points to the possibility that the recommended subsets of channels for sleep stage classification might not necessarily be optimal for wake detection, a crucial consideration considering sleep disorders such as insomnia.

In conclusion, this research underscores the importance of continuous exploration in the field of ASSC, opening up new possibilities for real-time sleep monitoring at home. The outcomes from this research offer valuable insights and provide essential findings for further research.

6.9 Implications for Future Research

Throughout this research, experiments have been performed exclusively on a dataset consisting of healthy subjects. As a next step, it would be beneficial to validate the applicability of the established models and methodologies on datasets featuring patients diagnosed with sleep disorders. The generalizability and robustness of the proposed methods with corresponding results could be enhanced by incorporating data with a broader range of age variability as well, covering a more comprehensive spectrum of sleep patterns across different life stages.

The presence of an imbalance in the dataset makes it more difficult for the ML algorithms to properly learn how to discriminate between the defined classes. Further work should investigate methods for balancing the dataset, either by data augmentation or by collecting and including more wake samples in the recordings.

The application of the NSGA optimization technique in this research has demonstrated promising outcomes. However, exploring different objectives could yield further improvements. For instance, aiming to maximize the sensitivity of the wake or directly minimizing false negatives. These objectives focus on the accurate detection of wakefulness.

These suggestions should serve to give a comprehensive view of potential directions for further research. It is essential to continue exploring and challenging the established models and methodologies to ensure their robustness and effectiveness in diverse and complex real-world scenarios.

References

[1] Karoline Herleiksplass. Binary classification for sleep stages with deep learning, eegnet. 2022.

[2] Terrence J Sejnowski and Alain Destexhe. Why do we sleep? *Brain research*, 886(1-2):208–223, 2000.

[3] Luigi Fiorillo, Alessandro Puiatti, Michela Papandrea, Pietro-Luca Ratti, Paolo Favaro, Corinne Roth, Panagiotis Bargiotas, Claudio L Bassetti, and Francesca D Faraci. Automated sleep scoring: A review of the latest approaches. *Sleep medicine reviews*, 48:101204, 2019.

[4] Kwang Suk Park and Sang Ho Choi. Smart technologies toward sleep monitoring at home. *Biomedical engineering letters*, 9(1):73–85, 2019.

[5] Maurice M Ohayon. Epidemiological overview of sleep disorders in the general population. *Sleep Medicine Research*, 2(1):1–9, 2011.

[6] Charles L Nunn, David R Samson, and Andrew D Krystal. Shining evolutionary light on human sleep and sleep disorders. *Evolution, medicine, and public health*, 2016(1):227–243, 2016.

[7] Thomas Roth and Timothy Roehrs. Insomnia: epidemiology, characteristics, and consequences. *Clinical cornerstone*, 5(3):5–15, 2003.

[8] Maria Paola Mogavero, Lourdes M DelRosso, Francesco Fanfulla, Oliviero Bruni, and Raffaele Ferri. Sleep disorders and cancer: State of the art and future perspectives. *Sleep Medicine Reviews*, 56:101409, 2021.

[9] Faith S Luyster, Patrick J Strollo, Phyllis C Zee, and James K Walsh. Sleep: a health imperative. *Sleep*, 35(6):727–734, 2012.

[10] Sunil Sharma and Mani Kavuru. Sleep and metabolism: an overview. *International journal of endocrinology*, 2010, 2010.

[11] Emadeldeen Eldele, Zhenghua Chen, Chengyu Liu, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. An attention-based deep learning approach for sleep stage classification with single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:809–818, 2021.

[12] Colin A Espie. Insomnia: conceptual issues in the development, persistence, and treatment of sleep disorder in adults. *Annual review of psychology*, 53(1):215–243, 2002.

[13] Marina Nano, Pedro Fonseca, Sebastiaan Overeem, Rik Vullings, and Ronald M Aarts. Lying awake at night: cardiac autonomic activity in relation to sleep onset and maintenance. *Frontiers in Neuroscience*, 13:1405, 2020.

[14] Wenrui Zhao, Eus JW Van Someren, Chenyu Li, Xinyuan Chen, Wenjun Gui, Yu Tian, Yunrui Liu, and Xu Lei. Eeg spectral analysis in insomnia disorder: A systematic review and meta-analysis. *Sleep Medicine Reviews*, 59:101457, 2021.

[15] Ahmed S BaHammam, Divinagracia E Gacuan, Smitha George, Karen Lorraine Acosta, Seithikurippu Ratnas Pandi-Perumal, and Ravi Gupta. Polysomnography i: procedure and technology. *Synopsis of Sleep Medicine*, pages 443–456, 2016.

[16] Mustafa Radha, Pedro Fonseca, Arnaud Moreau, Marco Ross, Andreas Cerny, Peter Anderer, Xi Long, and Ronald M Aarts. A deep transfer learning approach for wearable sleep stage classification with photoplethysmography. *NPJ digital medicine*, 4(1):1–11, 2021.

[17] Mark Bear, Barry Connors, and Michael A Paradiso. *Neuroscience: exploring the brain, enhanced edition: exploring the brain*. Jones & Bartlett Learning, 2020.

[18] Lennart Heimer. *The human brain and spinal cord: functional neuroanatomy and dissection guide*. Springer Science & Business Media, 2012.

[19] Priyanka A Abhang, Bharti Gawali, and Suresh C Mehrotra. *Introduction to EEG-and speech-based emotion recognition*. Academic Press, 2016.

[20] Berry Richard B, Albertario Claude L, Harding Susan M, and et al. The aasm manual for the scoring of sleep and associated events: Rules, terminology and technical specifications version 2.5. *the American Academy of Sleep Medicine*, 2018.

[21] Michal Teplan et al. Fundamentals of eeg measurement. *Measurement science review*, 2(2):1–11, 2002.

[22] José Luis Cantero and Mercedes Atienza. Alpha burst activity during human rem sleep: descriptive study and functional hypotheses. *Clinical neurophysiology*, 111(5):909–915, 2000.

[23] Juri Kropotov. *Quantitative EEG, event-related potentials and neurotherapy*. Academic Press, 2010.

[24] Margitta Seeck, Laurent Koessler, Thomas Bast, Frans Leijten, Christoph Michel, Christoph Baumgartner, Bin He, and Sándor Beniczky. The standardized eeg electrode array of the ifcn. *Clinical neurophysiology*, 128(10):2070–2077, 2017.

[25] Biosemi 128 headcaps. `https://www.biosemi.com/headcap.htm`. Accessed: 2022-11-19.

[26] Laurent Koessler, Louis Maillard, Adnane Benhadid, Jean Pierre Vignal, Jacques Felblinger, Hervé Vespignani, and Marc Braun. Automated cortical projection of eeg sensors: anatomical correlation via the international 10–10 system. *Neuroimage*, 46(1):64–72, 2009.

[27] Aurora Saibene, Mirko Caglioni, Silvia Corchs, and Francesca Gasparini. Eeg-based bcis on motor imagery paradigm using wearable technologies: A systematic review. *Sensors*, 23(5), 2023.

[28] James W Antony, Monika Schönauer, Bernhard P Staresina, and Scott A Cairney. Sleep spindles and memory reprocessing. *Trends in neurosciences*, 42(1):1–3, 2019.

[29] Patel AK, Reddy V, Shumway KR, and et al. *Physiology, Sleep Stages*. StatPearls, 2022. `https://www.ncbi.nlm.nih.gov/books/NBK526132/`.

[30] Péter Halász, Mario Terzano, Liborio Parrino, and Róbert Bódizs. The nature of arousal in sleep. *Journal of sleep research*, 13(1):1–23, 2004.

[31] Radhika Basheer, Robert E Strecker, Mahesh M Thakkar, and Robert W McCarley. Adenosine and sleep–wake regulation. *Progress in neurobiology*, 73(6):379–396, 2004.

[32] Su H Hwang, Yu J Lee, Do U Jeong, and Kwang S Park. Unconstrained sleep stage estimation based on respiratory dynamics and body movement. *Methods of information in medicine*, 55(06):545–555, 2016.

[33] Fabien Lotte, Marco Congedo, Anatole Lécuyer, Fabrice Lamarche, and Bruno Arnaldi. A review of classification algorithms for eeg-based brain–computer interfaces. *Journal of neural engineering*, 4(2):R1, 2007.

[34] Mahnaz Arvaneh, Cuntai Guan, Kai Keng Ang, and Chai Quek. Optimizing the channel selection and classification accuracy in eeg-based bci. *IEEE Transactions on Biomedical Engineering*, 58(6):1865–1873, 2011.

[35] Bruce J Swihart, Brian Caffo, Karen Bandeen-Roche, and Naresh M Punjabi. Characterizing sleep structure using the hypnogram. *Journal of Clinical Sleep Medicine*, 4(4):349–355, 2008.

[36] Michelle A Short, Michael Gradisar, Leon C Lack, Helen Wright, and Mary A Carskadon. The discrepancy between actigraphic and sleep diary measures of sleep in adolescents. *Sleep medicine*, 13(4):378–384, 2012.

[37] Matthew D Weaver, Tracey L Sletten, Russell G Foster, David Gozal, Elizabeth B Klerman, Shantha MW Rajaratnam, Till Roenneberg, Joseph S Takahashi, Fred W Turek, Michael V Vitiello, et al. Adverse impact of polyphasic sleep patterns in humans: Report of the national sleep foundation sleep timing and variability consensus panel. *Sleep health*, 7(3):293–302, 2021.

[38] Maurice M Ohayon, Mary A Carskadon, Christian Guilleminault, and Michael V Vitiello. 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, 2004.

[39] Bryce A Mander, Joseph R Winer, and Matthew P Walker. Sleep and human aging. *Neuron*, 94(1):19–36, 2017.

[40] Charles C-H Hong, James H Fallon, Karl J Friston, and James C Harris. Rapid eye movements in sleep furnish a unique probe into consciousness. *Frontiers in Psychology*, 9:2087, 2018.

[41] J Allan Hobson. Rem sleep and dreaming: towards a theory of protoconsciousness. *Nature Reviews Neuroscience*, 10(11):803–813, 2009.

[42] Andrew SP Lim, Matthew Kowgier, Lei Yu, Aron S Buchman, and David A Bennett. Sleep fragmentation and the risk of incident alzheimer's disease and cognitive decline in older persons. *Sleep*, 36(7):1027–1032, 2013.

[43] Marina-Marinela Nano, Pedro Fonseca, Rik Vullings, and Ronald M Aarts. Measures of cardiovascular autonomic activity in insomnia disorder: A systematic review. *PloS one*, 12(10):e0186716, 2017.

[44] Michael H Bonnet and Donna L Arand. Hyperarousal and insomnia: state of the science. *Sleep medicine reviews*, 14(1):9–15, 2010.

[45] Maxine F Profitt, Samuel Deurveilher, George S Robertson, Benjamin Rusak, and Kazue Semba. Disruptions of sleep/wake patterns in the stable tubule only polypeptide (stop) null mouse model of schizophrenia. *Schizophrenia Bulletin*, 42(5):1207–1215, 2016.

[46] Stefan Cohrs. Sleep disturbances in patients with schizophrenia: impact and effect of antipsychotics. *CNS drugs*, 22:939–962, 2008.

[47] Ashura Williams Buckley, Alcibiades J Rodriguez, Kaitlin Jennison, Jack Buckley, Audrey Thurm, Susumu Sato, and Susan Swedo. Rapid eye movement sleep percentage in children with autism compared with children with developmental delay and typical development. *Archives of pediatrics & adolescent medicine*, 164(11):1032–1037, 2010.

[48] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[49] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[50] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[51] Hojjat Salehinejad, Sharan Sankar, Joseph Barfett, Errol Colak, and Shahrokh Valaee. Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078*, 2017.

[52] Yasen Jiao and Pufeng Du. Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quantitative Biology*, 4:320–330, 2016.

[53] Fadi Thabtah, Suhel Hammoud, Firuz Kamalov, and Amanda Gonsalves. Data imbalance in classification: Experimental evaluation. *Information Sciences*, 513:429–441, 2020.

[54] Nan Wang, Xibin Zhao, Yu Jiang, Yue Gao, and KLISS BNRist. Iterative metric learning for imbalance data classification. In *IJCAI*, volume 2018, pages 2805–2811, 2018.

[55] Hao Ding, Bin Wei, Zhaorui Gu, Zhibin Yu, Haiyong Zheng, Bing Zheng, and Juan Li. Ka-ensemble: towards imbalanced image classification ensembling under-sampling and over-sampling. *Multimedia Tools and Applications*, 79(21):14871–14888, 2020.

[56] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, 6(2):182–197, 2002.

[57] Nidamarthi Srinivas and Kalyanmoy Deb. Muiltiobjective optimization using nondominated sorting in genetic algorithms. *Evolutionary computation*, 2(3):221–248, 1994.

[58] Wen-Hung Liao and Chien-Ming Yang. Video-based activity and movement pattern analysis in overnight sleep studies. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008.

[59] Janna Mantua, Nickolas Gravel, and Rebecca MC Spencer. Reliability of sleep measures from four personal health monitoring devices compared to research-based actigraphy and polysomnography. *Sensors*, 16(5):646, 2016.

[60] DJ Mullaney, DF Kripke, and S Messin. Wrist-actigraphic estimation of sleep time. *Sleep*, 3(1):83–92, 1980.

[61] Sjoerd J van Hasselt, Simon Verhulst, Theunis Piersma, Niels C Rattenborg, and Peter Meerlo. A comparison of continuous and intermittent eeg recordings in geese: How much data are needed to reliably estimate sleep–wake patterns? *Journal of sleep research*, 31(3):e13525, 2022.

[62] Pierrick J Arnal, Valentin Thorey, Michael E Ballard, Albert Bou Hernandez, Antoine Guillot, Hugo Jourde, Mason Harris, Mathias Guillard, Pascal Van Beers, Mounir Chennaoui, et al. The dreem headband as an alternative to polysomnography for eeg signal acquisition and sleep staging. *BioRxiv*, page 662734, 2019.

[63] Pierrick J Arnal, Valentin Thorey, Eden Debellemaniere, Michael E Ballard, Albert Bou Hernandez, Antoine Guillot, Hugo Jourde, Mason Harris, Mathias Guillard, Pascal Van Beers, et al. The dreem headband compared to polysomnography for electroencephalographic signal acquisition and sleep staging. *Sleep*, 43(11):zsaa097, 2020.

[64] Mingyu Fu, Yitian Wang, Zixin Chen, Jin Li, Fengguo Xu, Xinyu Liu, and Fengzhen Hou. Deep learning in automatic sleep staging with a single channel electroencephalography. *Frontiers in Physiology*, 12:628502, 2021.

[65] Bob Kemp, A Zwinderman, B Tuk, H Kamphuisen, and J Oberyé. Sleep-edf database expanded. *physionet. org*, 2018.

[66] S Devuyst, T Dutoit, and M Kerkhofs. The dreams databases and assessment algorithm. *Zenodo, Genève*, 2005.

[67] Marcely Zanon Boito*, William Havard*, Mahault Garnerin, Éric Le Ferrand, and Laurent Besacier. MaSS: A Large and Clean Multilingual Corpus of Sentence-aligned Spoken Utterances Extracted from the Bible. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6486–6493, Marseille, France, May 2020. European Language Resources Association.

[68] Ahnaf Rashik Hassan, Syed Khairul Bashar, and Mohammed Imamul Hassan Bhuiyan. On the classification of sleep states by means of statistical and spectral features from single channel electroencephalogram. In *2015 International conference on advances in computing, communications and informatics (ICACCI)*, pages 2238–2243. IEEE, 2015.

[69] Ahnaf Rashik Hassan and Mohammed Imamul Hassan Bhuiyan. Computer-aided sleep staging using complete ensemble empirical mode decomposition with adaptive noise and bootstrap aggregating. *Biomedical Signal Processing and Control*, 24:1–10, 2016.

[70] Santosh Kumar Satapathy and D Loganathan. Automated classification of multi-class sleep stages classification using polysomnography signals: a nine-layer 1d-convolution neural network approach. *Multimedia Tools and Applications*, 82(6):8049–8091, 2023.

[71] Choel-Hui Lee, Hyun-Ji Kim, Jae-Wook Heo, Hakseung Kim, and Dong-Joo Kim. Improving sleep stage classification performance by single-channel eeg data augmentation via spectral band blending. In *2021 9th International Winter Conference on Brain-Computer Interface (BCI)*, pages 1–5. IEEE, 2021.

[72] Siddharth Biswal, Joshua Kulas, Haoqi Sun, Balaji Goparaju, M Brandon Westover, Matt T Bianchi, and Jimeng Sun. Sleepnet: automated sleep staging system via deep learning. *arXiv preprint arXiv:1707.08262*, 2017.

[73] Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, 2017.

[74] Christian Berthomier, Xavier Drouot, Maria Herman-Stoïca, Pierre Berthomier, Jacques Prado, Djibril Bokar-Thire, Odile Benoit, Jérémie Mattout, and Marie-Pia d'Ortho. Automatic analysis of single-channel sleep eeg: validation in healthy individuals. *Sleep*, 30(11):1587–1595, 2007.

[75] Guohun Zhu, Yan Li, and Peng Wen. Analysis and classification of sleep stages based on difference visibility graphs from a single-channel eeg signal. *IEEE journal of biomedical and health informatics*, 18(6):1813–1821, 2014.

[76] Wessam Al-Salman, Yan Li, Atheer Y Oudah, and Sadiq Almaged. Sleep stage classification in eeg signals using the clustering approach based probability distribution features coupled with classification algorithms. *Neuroscience Research*, 188:51–67, 2023.

[77] Turky Alotaiby, Fathi E Abd El-Samie, Saleh A Alshebeili, and Ishtiaq Ahmad. A review of channel selection algorithms for eeg signal processing. *EURASIP Journal on Advances in Signal Processing*, 2015:1–21, 2015.

[78] Håkon Stenwig, Andres Soler, Junya Furuki, Yoko Suzuki, Takashi Abe, and Marta Molinas. Automatic sleep stage classification with optimized selection of eeg channels. *bioRxiv*, pages 2022–06, 2022.

[79] Alexandra Piryatinska, Wojbor A Woyczynski, Mark S Scher, and Kenneth A Loparo. Optimal channel selection for analysis of eeg-sleep patterns of neonates. *Computer methods and programs in biomedicine*, 106(1):14–26, 2012.

[80] Luis Alfredo Moctezuma and Marta Molinas. Eeg channel-selection method for epileptic-seizure classification based on multi-objective optimization. *Frontiers in neuroscience*, 14:593, 2020.

[81] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

[82] Luis Alfredo Moctezuma, Takashi Abe, and Marta Molinas. Eeg-based 5- and 2-class cnn for sleep stage classification. In *The 22nd World Congress of the International Federation of Automatic Control*, 2023.

[83] Hisao Ishibuchi, Ryo Imada, Yu Setoguchi, and Yusuke Nojima. Performance comparison of nsga-ii and nsga-iii on various many-objective test problems. In *2016 IEEE Congress on Evolutionary Computation (CEC)*, pages 3045–3052. IEEE, 2016.

[84] Magnus Själander, Magnus Jahre, Gunnar Tufte, and Nico Reissmann. Epic: An energy-efficient, high-performance gpgpu computing research infrastructure. *arXiv preprint arXiv:1912.05848*, 2019.

Appendix A

Tables of results

Table A.1: Full performance for all subjects utilizing all 128 high-density EEG channels for sleep-wake classification.

Subject	Accuracy	Kappa	AUROC	Sensitivity Wake	Sensitivity Sleep
1	0.984 ± 0.003	0.852 ± 0.036	0.894 ± 0.018	0.792 ± 0.036	0.997 ± 0.002
2	0.984 ± 0.002	0.847 ± 0.015	0.874 ± 0.012	0.750 ± 0.026	0.998 ± 0.002
3	0.939 ± 0.003	0.652 ± 0.024	0.767 ± 0.012	0.562 ± 0.026	0.971 ± 0.001
4	0.973 ± 0.006	0.758 ± 0.053	0.816 ± 0.020	0.634 ± 0.036	0.997 ± 0.005
5	0.968 ± 0.004	0.711 ± 0.031	0.824 ± 0.022	0.653 ± 0.046	0.996 ± 0.002
6	0.961 ± 0.006	0.865 ± 0.017	0.918 ± 0.005	0.853 ± 0.017	0.982 ± 0.010
7	0.976 ± 0.002	0.776 ± 0.035	0.835 ± 0.018	0.673 ± 0.037	0.997 ± 0.000
8	0.971 ± 0.004	0.704 ± 0.053	0.836 ± 0.053	0.680 ± 0.111	0.992 ± 0.006
9	0.951 ± 0.010	0.715 ± 0.155	0.828 ± 0.060	0.670 ± 0.128	0.986 ± 0.010
10	0.941 ± 0.005	0.642 ± 0.034	0.789 ± 0.013	0.592 ± 0.025	0.985 ± 0.003
11	0.988 ± 0.003	0.845 ± 0.073	0.911 ± 0.021	0.826 ± 0.041	0.996 ± 0.000
12	0.987 ± 0.005	0.874 ± 0.153	0.910 ± 0.061	0.823 ± 0.127	0.997 ± 0.004
13	0.990 ± 0.002	0.872 ± 0.051	0.907 ± 0.033	0.816 ± 0.068	0.998 ± 0.002
AVERAGE	**0.970**	**0.778**	**0.854**	**0.717**	**0.992**
STD	**0.018**	**0.086**	**0.051**	**0.098**	**0.008**

Table A.2: Full performance for all subjects utilizing all 128 high-density EEG channels for five class sleep stage classification

Subject	Accuracy	Kappa	Sensitivity Wake	Sensitivity N1	Sensitivity N2	Sensitivity N3	Sensitivity REM
1	0.852 ± 0.006	0.780 ± 0.008	0.817 ± 0.014	0.219 ± 0.094	0.897 ± 0.016	0.931 ± 0.022	0.865 ± 0.028
2	0.820 ± 0.007	0.746 ± 0.015	0.743 ± 0.034	0.333 ± 0.091	0.882 ± 0.056	0.862 ± 0.081	0.948 ± 0.026
3	0.788 ± 0.010	0.716 ± 0.013	0.507 ± 0.072	0.112 ± 0.055	0.902 ± 0.021	0.900 ± 0.049	0.953 ± 0.011
4	0.785 ± 0.007	0.671 ± 0.007	0.665 ± 0.035	0.321 ± 0.049	0.914 ± 0.032	0.662 ± 0.050	0.978 ± 0.017
5	0.759 ± 0.015	0.647 ± 0.029	0.681 ± 0.098	0.362 ± 0.104	0.910 ± 0.038	0.772 ± 0.031	0.845 ± 0.133
6	0.823 ± 0.008	0.759 ± 0.016	0.869 ± 0.124	0.608 ± 0.094	0.784 ± 0.020	0.932 ± 0.030	0.950 ± 0.012
7	0.814 ± 0.009	0.755 ± 0.012	0.644 ± 0.086	0.189 ± 0.081	0.842 ± 0.034	0.848 ± 0.031	0.929 ± 0.022
8	0.868 ± 0.013	0.834 ± 0.018	0.758 ± 0.132	0.345 ± 0.036	0.930 ± 0.020	0.901 ± 0.027	0.920 ± 0.015
9	0.830 ± 0.010	0.767 ± 0.013	0.705 ± 0.068	0.286 ± 0.115	0.876 ± 0.026	0.828 ± 0.035	0.979 ± 0.007
10	0.736 ± 0.013	0.572 ± 0.022	0.690 ± 0.052	0.149 ± 0.083	0.869 ± 0.036	0.934 ± 0.044	0.970 ± 0.023
11	0.893 ± 0.005	0.835 ± 0.008	0.764 ± 0.063	0.271 ± 0.069	0.921 ± 0.017	0.896 ± 0.040	0.968 ± 0.014
12	0.780 ± 0.011	0.715 ± 0.016	0.789 ± 0.015	0.470 ± 0.129	0.833 ± 0.083	0.903 ± 0.022	0.975 ± 0.022
13	0.830 ± 0.012	0.762 ± 0.019	0.809 ± 0.061	0.185 ± 0.058	0.901 ± 0.022	0.877 ± 0.040	0.929 ± 0.036
AVERAGE	**0.814**	**0.735**	**0.726**	**0.296**	**0.882**	**0.865**	**0.939**
STD	**0.044**	**0.073**	**0.093**	**0.136**	**0.041**	**0.077**	**0.042**

Table A.3: Full performance for all subjects utilizing the DREEM channels for sleep-wake classification.

Subject	Accuracy	Kappa	AUROC	Sensitivity Wake	Sensitivity Sleep
1	0.960 ± 0.001	0.599 ± 0.024	0.714 ± 0.024	0.429 ± 0.052	0.998 ± 0.004
2	0.982 ± 0.002	0.809 ± 0.029	0.846 ± 0.020	0.692 ± 0.040	0.999 ± 0.001
3	0.950 ± 0.003	0.564 ± 0.008	0.703 ± 0.008	0.406 ± 0.017	0.999 ± 0.002
4	0.969 ± 0.001	0.667 ± 0.007	0.770 ± 0.004	0.541 ± 0.008	1.000 ± 0.0
5	0.960 ± 0.004	0.680 ± 0.042	0.785 ± 0.029	0.574 ± 0.060	0.996 ± 0.002
6	0.944 ± 0.003	0.784 ± 0.014	0.894 ± 0.016	0.817 ± 0.038	0.970 ± 0.006
7	0.964 ± 0.003	0.631 ± 0.023	0.734 ± 0.011	0.471 ± 0.022	0.998 ± 0.001
8	0.962 ± 0.004	0.678 ± 0.019	0.784 ± 0.037	0.579 ± 0.082	0.989 ± 0.010
9	0.920 ± 0.005	0.456 ± 0.037	0.679 ± 0.020	0.370 ± 0.043	0.989 ± 0.003
10	0.920 ± 0.005	0.347 ± 0.049	0.654 ± 0.025	0.313 ± 0.056	0.995 ± 0.006
11	0.982 ± 0.002	0.785 ± 0.022	0.844 ± 0.015	0.693 ± 0.030	0.996 ± 0.001
12	0.976 ± 0.003	0.765 ± 0.029	0.834 ± 0.026	0.672 ± 0.055	0.995 ± 0.005
13	0.982 ± 0.002	0.779 ± 0.052	0.855 ± 0.055	0.715 ± 0.114	0.994 ± 0.004
AVERAGE	**0.959**	**0.657**	**0.777**	**0.559**	**0.994**
STD	**0.020**	**0.134**	**0.073**	**0.148**	**0.008**

Table A.4: Full performance for all subjects utilizing the AASM channels sleep-wake classification

Subject	Accuracy	Kappa	AUROC	Sensitivity Wake	Sensitivity Sleep
1	0.963 ± 0.001	0.575 ± 0.022	0.728 ± 0.012	0.457 ± 0.023	0.998 ± 0.001
2	0.981 ± 0.001	0.807 ± 0.023	0.824 ± 0.015	0.649 ± 0.030	1.000 ± 0.0
3	0.929 ± 0.005	0.445 ± 0.104	0.643 ± 0.053	0.300 ± 0.118	0.985 ± 0.012
4	0.965 ± 0.002	0.642 ± 0.015	0.741 ± 0.008	0.482 ± 0.017	1.000 ± 0.0
5	0.960 ± 0.003	0.717 ± 0.030	0.773 ± 0.017	0.547 ± 0.035	0.999 ± 0.001
6	0.925 ± 0.004	0.713 ± 0.017	0.823 ± 0.007	0.671 ± 0.012	0.976 ± 0.003
7	0.960 ± 0.004	0.541 ± 0.046	0.696 ± 0.023	0.392 ± 0.046	1.000 ± 0.0
8	0.950 ± 0.002	0.415 ± 0.039	0.659 ± 0.028	0.324 ± 0.060	0.995 ± 0.005
9	0.905 ± 0.004	0.297 ± 0.047	0.600 ± 0.024	0.207 ± 0.056	0.993 ± 0.009
10	0.920 ± 0.008	0.556 ± 0.114	0.660 ± 0.056	0.330 ± 0.118	0.991 ± 0.008
11	0.975 ± 0.002	0.630 ± 0.016	0.760 ± 0.009	0.524 ± 0.017	0.996 ± 0.001
12	0.951 ± 0.003	0.407 ± 0.060	0.634 ± 0.027	0.275 ± 0.054	0.993 ± 0.001
13	0.980 ± 0.001	0.709 ± 0.030	0.766 ± 0.017	0.532 ± 0.035	1.000 ± 0.0
AVERAGE	**0.951**	**0.573**	**0.716**	**0.438**	**0.994**
STD	**0.024**	**0.149**	**0.073**	**0.146**	**0.007**

Table A.5: Full performance for all subjects utilizing the bipolar single-channel pair Fpz-Cz for sleep-wake classification

Subject	Accuracy	Kappa	AUROC	Sensitivity Wake	Sensitivity Sleep
1	0.958 ± 0.002	0.529 ± 0.023	0.694 ± 0.012	0.388 ± 0.024	0.999 ± 0.001
2	0.971 ± 0.002	0.669 ± 0.022	0.764 ± 0.012	0.531 ± 0.025	0.997 ± 0.001
3	0.914 ± 0.004	0.000 ± 0.156	0.577 ± 0.066	0.172 ± 0.146	0.983 ± 0.015
4	0.950 ± 0.003	0.446 ± 0.011	0.665 ± 0.009	0.334 ± 0.021	0.997 ± 0.003
5	0.947 ± 0.004	0.478 ± 0.066	0.700 ± 0.037	0.403 ± 0.076	0.997 ± 0.003
6	0.927 ± 0.003	0.730 ± 0.007	0.846 ± 0.008	0.723 ± 0.021	0.968 ± 0.006
7	0.949 ± 0.007	0.516 ± 0.159	0.605 ± 0.060	0.212 ± 0.120	0.998 ± 0.001
8	0.940 ± 0.003	0.184 ± 0.024	0.562 ± 0.008	0.124 ± 0.016	1.000 ± 0.0
9	0.909 ± 0.005	0.321 ± 0.039	0.647 ± 0.020	0.307 ± 0.045	0.988 ± 0.006
10	0.926± 0.002	0.523 ± 0.017	0.690 ± 0.007	0.391 ± 0.014	0.990 ± 0.003
11	0.955 ± 0.003	0.000 ± 0.000	0.500 ± 0.000	0.000 ± 0.000	1.000 ± 0.000
12	0.951 ± 0.002	0.393 ± 0.060	0.614 ± 0.026	0.231 ± 0.056	0.996 ± 0.003
13	0.972 ± 0.007	0.589 ± 0.072	0.774 ± 0.076	0.558 ± 0.163	0.990 ± 0.012
AVERAGE	**0.944**	**0.414**	**0.664**	**0.336**	**0.993**
STD	**0.020**	**0.231**	**0.095**	**0.196**	**0.009**

Table A.6: Full performance for all subjects utilizing the bipolar single-channel pair Pz-Oz for sleep-wake classification

Subject	Accuracy	Kappa	AUROC	Sensitivity Wake	Sensitivity Sleep
1	0.941 ± 0.009	0.274 ± 0.130	0.548 ± 0.041	0.096 ± 0.082	1.000 ± 0.000
2	0.965 ± 0.003	0.670 ± 0.045	0.772 ± 0.035	0.554 ± 0.074	0.990 ± 0.006
3	0.920 ± 0.003	0.012 ± 0.113	0.532 ± 0.038	0.067 ± 0.080	0.997 ± 0.004
4	0.934 ± 0.002	0.000 ± 0.000	0.500 ± 0.000	0.000 ± 0.000	1.000 ± 0.000
5	0.959 ± 0.001	0.676 ± 0.011	0.784 ± 0.009	0.574 ± 0.020	0.994 ± 0.001
6	0.892 ± 0.003	0.621 ± 0.024	0.791 ± 0.020	0.641 ± 0.046	0.941 ± 0.007
7	0.934 ± 0.001	0.028 ± 0.005	0.506 ± 0.001	0.011 ± 0.003	1.000 ± 0.000
8	0.958 ± 0.004	0.566 ± 0.042	0.749 ± 0.036	0.508 ± 0.076	0.990 ± 0.007
9	0.899 ± 0.008	0.187 ± 0.115	0.562 ± 0.041	0.129 ± 0.087	0.996 ± 0.005
10	0.925 ± 0.004	0.502 ± 0.044	0.665 ± 0.021	0.332 ± 0.043	0.997 ± 0.003
11	0.955 ± 0.004	0.000 ± 0.000	0.500 ± 0.000	0.000 ± 0.000	1.000 ± 0.000
12	0.961 ± 0.004	0.411 ± 0.092	0.653 ± 0.045	0.307 ± 0.091	0.999 ± 0.001
13	0.974 ± 0.001	0.609 ± 0.018	0.773 ± 0.018	0.553 ± 0.038	0.993 ± 0.003
AVERAGE	**0.940**	**0.336**	**0.641**	**0.290**	**0.992**
STD	**0.026**	**0.290**	**0.121**	**0.250**	**0.016**

Table A.7: Full performance for all subjects utilizing the NSGA-1 subset of channels, corresponding to: D25, A19, B14

Subject	Accuracy	Kappa	AUROC	Sensitivity Wake	Sensitivity Sleep
1	0.951 ± 0.004	0.428 ± 0.037	0.684 ± 0.012	0.377 ± 0.023	0.991 ± 0.005
2	0.975 ± 0.002	0.719 ± 0.020	0.786 ± 0.011	0.572 ± 0.022	0.999 ± 0.001
3	0.939 ± 0.001	0.405 ± 0.025	0.656 ± 0.010	0.318 ± 0.021	0.994 ± 0.001
4	0.959 ± 0.003	0.602 ± 0.028	0.701 ± 0.013	0.402 ± 0.027	1.000 ± 0.0
5	0.948 ± 0.003	0.516 ± 0.023	0.705 ± 0.015	0.414 ± 0.031	0.997 ± 0.002
6	0.921 ± 0.007	0.745 ± 0.019	0.841 ± 0.010	0.721 ± 0.028	0.961 ± 0.012
7	0.954 ± 0.004	0.393 ± 0.047	0.666 ± 0.021	0.332 ± 0.041	1.000 ± 0.0
8	0.956 ± 0.004	0.550 ± 0.067	0.700 ± 0.032	0.406 ± 0.064	0.994 ± 0.002
9	0.899 ± 0.005	0.156 ± 0.054	0.566 ± 0.023	0.141 ± 0.048	0.991 ± 0.003
10	0.908 ± 0.006	0.325 ± 0.079	0.626 ± 0.024	0.261 ± 0.049	0.990 ± 0.000
11	0.974 ± 0.002	0.602 ± 0.031	0.744 ± 0.015	0.492 ± 0.030	0.997 ± 0.001
12	0.951 ± 0.017	0.481 ± 0.035	0.713 ± 0.060	0.443 ± 0.144	0.983 ± 0.026
13	0.975 ± 0.003	0.674 ± 0.097	0.750 ± 0.049	0.505 ± 0.100	0.996 ± 0.002
AVERAGE	**0.947**	**0.507**	**0.703**	**0.414**	**0.992**
STD	**0.024**	**0.161**	**0.067**	**0.139**	**0.010**

Table A.8: Full performance for all subjects utilizing the NSGA-2 subset of channels, corresponding to: D25, A19, B14, A14, B29, D4

Subject	Accuracy	Kappa	AUROC	Sensitivity Wake	Sensitivity Sleep
1	0.962 ± 0.001	0.617 ± 0.025	0.752 ± 0.019	0.510 ± 0.039	0.995 ± 0.002
2	0.979 ± 0.003	0.788 ± 0.040	0.826 ± 0.025	0.653 ± 0.049	0.999 ± 0.002
3	0.945 ± 0.003	0.516 ± 0.021	0.689 ± 0.012	0.382 ± 0.024	0.996 ± 0.002
4	0.969 ± 0.003	0.668 ± 0.026	0.782 ± 0.014	0.567 ± 0.029	0.998 ± 0.002
5	0.958 ± 0.007	0.640 ± 0.058	0.798 ± 0.040	0.606 ± 0.082	0.990 ± 0.007
6	0.944 ± 0.007	0.810 ± 0.023	0.892 ± 0.020	0.815 ± 0.053	0.969 ± 0.015
7	0.966 ± 0.003	0.641 ± 0.024	0.744 ± 0.013	0.489 ± 0.026	1.000 ± 0.000
8	0.963 ± 0.004	0.604 ± 0.032	0.754 ± 0.015	0.512 ± 0.030	0.995 ± 0.003
9	0.927 ± 0.006	0.511 ± 0.061	0.744 ± 0.045	0.511 ± 0.098	0.977 ± 0.008
10	0.920 ± 0.003	0.469 ± 0.008	0.681 ± 0.006	0.376 ± 0.015	0.987 ± 0.003
11	0.979 ± 0.003	0.719 ± 0.044	0.791 ± 0.027	0.585 ± 0.055	0.997 ± 0.001
12	0.978 ± 0.001	0.795 ± 0.020	0.851 ± 0.027	0.708 ± 0.057	0.994 ± 0.003
13	0.981 ± 0.004	0.797 ± 0.066	0.787 ± 0.040	0.575 ± 0.080	1.000 ± 0.000
AVERAGE	**0.959**	**0.660**	**0.776**	**0.561**	**0.992**
STD	**0.019**	**0.113**	**0.057**	**0.116**	**0.009**

Table A.9: Full performance for all subjects utilizing the NSGA-3 subset of channels, corresponding to: D25, A19, B14, A14, B29, D4, B11, B31

Subject	Accuracy	Kappa	AUROC	Sensitivity Wake	Sensitivity Sleep
1	0.965 ± 0.005	0.719 ± 0.051	0.789 ± 0.024	0.587 ± 0.047	0.991 ± 0.003
2	0.979 ± 0.003	0.747 ± 0.030	0.823 ± 0.016	0.646 ± 0.032	0.999 ± 0.001
3	0.952 ± 0.004	0.571 ± 0.027	0.748 ± 0.012	0.503 ± 0.024	0.993 ± 0.001
4	0.969 ± 0.002	0.674 ± 0.007	0.784 ± 0.003	0.570 ± 0.006	0.997 ± 0.001
5	0.961 ± 0.013	0.790 ± 0.067	0.817 ± 0.031	0.645 ± 0.068	0.988 ± 0.016
6	0.951 ± 0.005	0.835 ± 0.012	0.904 ± 0.006	0.832 ± 0.021	0.975 ± 0.010
7	0.966 ± 0.002	0.674 ± 0.037	0.768 ± 0.029	0.540 ± 0.060	0.996 ± 0.003
8	0.962 ± 0.005	0.742 ± 0.050	0.788 ± 0.042	0.587 ± 0.089	0.989 ± 0.006
9	0.931 ± 0.003	0.627 ± 0.023	0.768 ± 0.154	0.560 ± 0.033	0.976 ± 0.004
10	0.931 ± 0.002	0.561 ± 0.029	0.729 ± 0.021	0.473 ± 0.047	0.985 ± 0.005
11	0.981 ± 0.004	0.768 ± 0.067	0.842 ± 0.047	0.689 ± 0.096	0.995 ± 0.002
12	0.978 ± 0.003	0.835 ± 0.020	0.871 ± 0.017	0.750 ± 0.038	0.993 ± 0.005
13	0.980 ± 0.004	0.665 ± 0.072	0.768 ± 0.043	0.535 ± 0.087	1.000 ± 0.000
AVERAGE	**0.962**	**0.708**	**0.800**	**0.609**	**0.991**
STD	**0.016**	**0.086**	**0.048**	**0.097**	**0.008**

Table A.10: Full performance for all subjects utilizing their corresponding best subset of channels from the NSGA analysis

Subject	Accuracy	Kappa	AUROC	Sensitivity Wake	Sensitivity Sleep	Channels
1	0.966 ± 0.005	0.670 ± 0.030	0.823 ± 0.026	0.657 ± 0.677	0.988 ± 0.008	A14, A19, B11, B14, B25, B29, D4, D13, D21
2	0.984 ± 0.002	0.823 ± 0.019	0.861 ± 0.08	0.723 ± 0.016	0.999 ± 0.001	A14, B11, B29, D4, D19, D25, D26, D32
3	0.951 ± 0.006	0.629 ± 0.017	0.748 ± 0.022	0.504 ± 0.055	0.992 ± 0.012	A14, A23, B11, B14, B31, C9, C13, C14, D21, D25
4	0.970 ± 0.003	0.713 ± 0.035	0.796 ± 0.019	0.594 ± 0.039	0.997 ± 0.002	A14, B11, B29, D4, D19, D25, D26, D32
5	0.966 ± 0.003	0.778 ± 0.039	0.819 ± 0.030	0.641 ± 0.061	0.996 ± 0.003	A22, A30, B11, B14, B29, B31, C4, C13, D6, D25
6	0.958 ± 0.005	0.816 ± 0.019	0.906 ± 0.017	0.826 ± 0.036	0.985 ± 0.004	A14, A19, A21, B2, B11, B14, B29, D25
7	0.963 ± 0.003	0.692 ± 0.052	0.776 ± 0.027	0.560 ± 0.055	0.992 ± 0.001	A13, A14, A19, A30, B11, B14, B31, D4, D19, D25
8	0.971 ± 0.005	0.755 ± 0.046	0.858 ± 0.050	0.727 ± 0.106	0.988 ± 0.008	A14, A19, B1, B32, C30, D4, D9, D15, D23, D25
9	0.944 ± 0.003	0.670 ± 0.029	0.792 ± 0.035	0.599 ± 0.080	0.986 ± 0.010	A14, A19, B14, B31, C8, D4, D6, D21, D25
10	0.937 ± 0.003	0.653 ± 0.031	0.743 ± 0.020	0.492 ± 0.042	0.994 ± 0.003	A14, A16, A19, B29, B31, C8, C29, D4, D15, D25
11	0.946 ± 0.003	0.662 ± 0.042	0.775 ± 0.030	0.557 ± 0.059	0.993 ± 0.001	A11, A14, B14, C8, C13, C16, D4, D17, D21, D32
12	0.985 ± 0.002	0.849 ± 0.021	0.879 ± 0.012	0.760 ± 0.023	0.998 ± 0.001	A19, B23, B29, B31, C13, D4, D5, D9, D21, D25
13	0.989 ± 0.002	0.830 ± 0.031	0.901 ± 0.021	0.805 ± 0.041	0.998 ± 0.001	A21, A22, B11, B14, C22, C28, C30, D16, D25
AVERAGE	**0.964**	**0.734**	**0.821**	**0.650**	**0.993**	
STD	**0.016**	**0.078**	**0.055**	**0.111**	**0.005**	

Appendix B

BI Conference Paper: *Automatic sleep-wake scoring with optimally selected EEG channels from high-density EEG*

Automatic sleep-wake scoring with optimally selected EEG channels from high-density EEG

Karoline Seljevoll Herleiksplass[1], Luis Alfredo Moctezuma[2], Junya Furuki, Yoko Suzuki[2], Takashi Abe[2], and Marta Molinas[1]

[1] Department of Engineering Cybernetics, Norwegian University of Science and Technology, Trondheim, Norway. karolshe@stud.ntnu.no
[2] International Institute for Integrative Sleep Medicine (WPI-IIIS), University of Tsukuba, Tsukuba, Ibaraki, Japan.

Abstract. This paper presents a sleep versus wake-classification model based on high-density electroencephalographic (EEG) sleep data. In a second stage, an optimization algorithm is applied to select the minimal set of EEG channels according to their contribution to classification performance. The performance of these subsets are compared to the ones recommended by the American Academy of Sleep Medicine (AASM) and with the 128 available channels in the tested dataset. We focus on accurate classification of the wake stage because of its importance as a metric of sleep quality and in diagnosing several sleep disorders. The performed experiment demonstrates that high-density EEG channels can achieve an accuracy and kappa score of 0.970 and 0.778 using the proposed method and dataset. But most importantly, the results from the optimization algorithm, Non-dominated Sorting Genetic Algorithm (NSGA)-III, indicate that using 3 to 9 channels can yield an average accuracy ranging between 0.955 to 0.975 and a kappa score between 0.678 and 0.795. The three most important channels detected by the optimization algorithm were D25, A19 and B14 from the 10-10 international system. This represents a significant improvement compared to the results obtained using the AASM-recommended channels, which yielded an average accuracy and kappa score of 0.951 and 0.573, respectively.

Keywords: Sleep staging · Electroencephalography (EEG) · Channel selection · Multi-objective optimization · Convolutional Neural Networks (CNN)

1 Introduction

Having high-quality sleep is crucial for maintaining good, both physical and mental health. Sleep disturbances and sleep disorders can lead to various health problems and are becoming a growing concern that requires medical attention. Sleep interruptions can be due to many different factors and result in nighttime awakenings. Patients suffering from the sleep disorder insomnia have difficulties sleeping through the night and have trouble getting back to sleep after unwanted

awakenings. The frequent occurrence of nighttime awakenings may lead to or signify a risk of developing sleep disorders.

Understanding the quality of sleep and the identification and measurement of wake periods throughout sleep is essential for diagnosing and treating many sleep disorders. Particularly key metrics that can offer significant insights into the overall sleep architecture include identifying and measuring the length and frequency of wakeful periods throughout a night of sleep. As such, accurate and reliable methods for capturing wake periods are essential for ensuring effective sleep monitoring and management [5, 11, 15].

Polysomnographic (PSG) recordings provide measurement for electroencephalography (EEG), electromyography (EMG), electrocardiography (ECG), electrooculography (EOG), and respiratory parameters. PSG are used in clinical settings to monitor and classify sleep stages or look for specific sleep patterns. This analysis can then be used to monitor and diagnose sleep disorders. The data is normally split into 30 seconds epochs, and each epoch is classified into one sleep stage by a sleep expert. This process is considered complex and time-consuming. This known challenge has motivated research using machine learning models for automatic sleep stage classification [6].

The need to handle large amounts of data that require significant storage space has prompted research into using smaller amounts of data while still achieving acceptable sleep stage classification accuracy. As EEG signals are commonly used in sleep research, there has been a desire to use a few channels to achieve real-time sleep staging. The American Academy of Sleep Medicine (AASM) recommends using three EEG channels: A28, B22, and C4 from the international 10-10 configuration [12].

Methods for capturing sleep data that are more widely available than PSG have been proposed. The Dreem Headband is a more user-friendly, inexpensive, and comfortable alternative to PSG, allowing sleep data to be collected at home. The Dreem Headband measures movement, head position, respiration trace, and rate using five EEG channels and a 3D accelerometer. The following EEG channels have been used with the 10-10 international convention: F7, F8, Fp1, O1 and O2. In research, the Dreem Headband has obtained an overall accuracy of 83.5% and an overall average kappa score of 74.8 [1, 2]

Previous research on binary sleep classification has employed varying numbers of channels, including single-channel analysis. For sleep-wake classification with single-channel EEG, works have reported accuracies up to 97.9% [16]. By using 8 channels, Ronzhina et al., for instance, obtained a sleep-wake accuracy of 96.06% and a Cohen's kappa score (kappa) of 0.666. Zhu et al. obtained an accuracy of 97.9% and a kappa score of 0.96 with their model, utilizing single-channel-pair (Pz-Oz) EEG data. Both works used EEG data from the publicly available Sleep-EDF Database, which consists of 30-second epochs, 3919 samples of wake, and 3562 samples of sleep [3, 7, 13, 16].

Recently, a convolutional neural network (CNN) called EEGNet has been applied to EEG-related applications [8]. The EEGNet has been combined with bidirectional long short-term memory (BiLSTM) for sleep staging, and

authors reported 87% of accuracy for 5-class classification [9]. For this, authors have randomly mixed selected delta, theta, and beta frequencies for data augmentation to overcome the class imbalance problem. In general, the current state-of-the-art reports accuracies around $87\% \pm 4$.

This study aims to utilize raw data without pre-processing from an unbalanced dataset to achieve the most realistic representation of a real-world situation. It is desirable to employ deep learning techniques to avoid being dependent on prior sleep knowledge of the unprocessed data.

In this work, we explore sleep-wake classification using the EEGNet. Our main focus is testing a multi-objective optimization technique, called Non-dominated Sorting Genetic Algorithm (NSGA)-III, successfully used to reduce the number of required EEG channels for sleep-wake classification.

2 Materials and methods

2.1 Dataset

The dataset was collected at the International Institute for Integrative Sleep Medicine (WPI-IIIS) at the University of Tsukuba, Japan, and the data collection was approved by the ethics committee of the University. It consists of EEG recordings from 14 subjects (5 female and 9 male, 22.5 ± 0.9 years), who were sleeping for about 8 hours while their data were collected using 128 EEG channels (BioSemi 128 configuration) with a sample rate of 1024 Hz, three EOG channels, three EMG channels, and two mastoid channels. The sleep stages were manually labelled every 30 seconds by a sleep expert from WPI-IIIS. In the present study, we considered 13 subjects, excluding one of the subjects that interrupted the data collection during the night.

The data were downsampled to 128 Hz, and each 30-second epoch was divided into 2-second segments. By utilizing a small segment size, it is possible to increase the number of instances of the minority wake class, which may contribute to better learning of patterns from the minority class. Additionally, this will allow us to detect the transition between different sleep stages swiftly. Similar comparisons have been conducted in the following research [10].

The dataset was divided into 50% as training data, 25% as validation data and 25% as test data. The data utilized in this study was in its raw form and was not subjected to any preprocessing. This is in contrast to publicly available datasets that are typically preprocessed and consider disturbances such as noise. Utilizing raw data with deep learning methods obviates the need for prior sleep knowledge compared to feature extraction methods. The dataset is highly unbalanced, which may affect the machine learning models, and the distribution between wake and sleep is presented in Table 1.

2.2 Classification Model

EEGNet is a compact CNN developed for EEG-based brain-computer interface (BCI) [8]. Its architecture aims to be generalizable across different BCI

Table 1: The percentage distribution of wake and sleep stages for all subjects in the WPI-IIIS dataset.

	Subject												
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Wake	6.6	5.7	8.4	6.8	8.3	16.5	6.6	6.7	11.0	10.9	4.6	5.9	4.2
Sleep	93.4	94.3	91.6	93.2	91.7	83.5	93.4	93.3	89.0	89.1	95.4	94.1	95.8

paradigms and is claimed to be robust enough to learn features for a wide range of different BCI tasks. The architecture of EEGNet includes two convolutional blocks followed by a classification layer. The first block is designed to learn temporal frequency filters from the input signal, while the second block extracts more abstract features from the EEG data. The network's design is based on a combination of spatial and temporal filtering techniques to produce interpretable features that can be applied across several BCI paradigms. EEGNet was developed to encapsulate the well-known features of EEG signals, including filter-bank construction and optimal spatial filtering, while reducing the number of trainable parameters [8].

EEGNet has the ability to capture both spatial and frequency information from the EEG signals. Together with the small amount of trainable parameters, the model can be more efficient to train and is less likely to overfitting compared to other neural network architectures. EEGNet has also shown to be robust to variations in the sampling rate and robust to limited training data. These qualifications look promising for use in sleep-wake classification with the presented dataset.

The training of the model is performed subject-by-subject, and default values from the implementation are used for the dropout rate (0.5) and batch size (16). The optimizer used is ADAM and the loss function is binary cross-entropy, as default. Early stopping is used for the number of epochs to prevent the model from overfitting.

2.3 Channel Selection

NSGA is a multi-objective optimization algorithm that can optimize complex problems based on one or more objectives. NSGA is a genetic algorithm that is based on the natural phenomenon of natural selection and the survival of the fittest through evolution, including mutation and crossover. A genetic algorithm works by generating different possible solutions, defined as chromosomes. A population is a collection of several chromosomes. The algorithm evaluates the result of each chromosome in the population based on the defined objectives and constraints. The chromosomes are ranked based on their level of non-domination. If no other chromosome has a better evaluation value, the chromosome is considered non-dominated. The non-dominated chromosomes are used to create the next generation using crossover and random mutations. The process is then

repeated for several generations, each producing a set of non-dominated solutions [4].

NSGA-II, an NSGA extension, modifies the survival of the chromosomes by selecting frontwise, using a Manhattan Distance-based crowding distance metric, and modifies the mating process by selecting by rank and crowding distance. NSGA-III further modifies NSGA-II, using initialized reference directions for optimization. If dividing the front, survivors are chosen by prioritizing the most underrepresented reference direction and the solution with the smallest normalized objective space's perpendicular distance. NSGA-III seeks a non-dominated solution for each reference line. [4].

We have implemented NSGA-III (NSGA) for EEG channel selection to space the reference points evenly. The optimization problem consisted of four objectives: maximize the 1) model accuracy, and from the test set the 2) Cohen's kappa coefficient (kappa) and 3) the area under the receiver operating characteristic curve (AUROC) score, all this, while 4) minimizing the number of EEG channels used during the classification process. The NSGA-III configuration represents each channel in a 128-binary array, 1 if the channel will be used for the classification process, and 0 if not. NSGA-III generated 10 chromosomes (10 possible EEG channel combinations) per generation, and the process was repeated for 150 generations.

Throughout the experimental phase of the investigation, it was observed that utilizing as few as 8 channels yielded encouraging outcomes. Consequently, the logical next step was to explore the feasibility of solutions that employed fewer than 10 channels. To this end, optimization was conducted under the constraint of limiting the number of channels to no more than 10. Similar findings were reached through experimentation with one or two channels, as the sensitivity of wake was observed to be low.

Given the configuration of our experiments, we performed all the experiments using GPUs on the NTNU IDUN computing cluster [14]. The cluster has more than 70 nodes and 90 general-purpose GPUs (GPGPUs). Each node contains two Intel Xeon cores and at least 128 GB of main memory, and is connected to an Infiniband network. Half of the nodes are equipped with two or more Nvidia Tesla P100 or V100 GPGPUs. Idun's storage is provided by two storage arrays and a Lustre parallel distributed file system.

3 Results

3.1 Sleep-wake classification using high-density EEG

After the data labeling step, the EEG signal segments were used as input to the CNN to classify between sleep and wake segments. The presented results are the average based on the training and evaluation of five separate models with different training, validation, and test subset splits. In this experiment, we use the 128 available channels and the three AASM-recommended EEG channels. Table 2 presents the average results among all subjects. The results

Table 2: Average performance with all the subjects for sleep-wake classification using 128 channels and the AASM-recommended channels.

	Accuracy	**Kappa**	**AUROC**	**Sensitivity Wake**	**Sensitivity Sleep**
128 channels	0.970±0.018	0.778±0.086	0.855±0.051	0.717±0.098	0.992±0.008
AASM channels	0.951±0.024	0.573±0.149	0.716±0.073	0.438±0.146	0.994±0.007

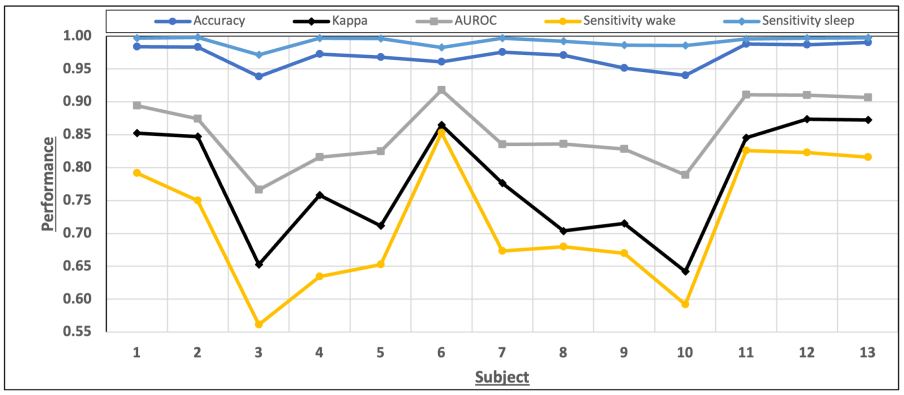


Fig. 1: Performance of sleep-wake classification for all subjects using 128 channels.

for each subject with all channels are presented in Figure 1 and with AASM-recommended channels in Figure 2.

The distribution of sleep data affects the classification model's overall performance. A comparison of the distribution in Table 1 with the results presented in Figure 1 and Figure 2 reveals that Subject 6 attained the highest sensitivity for wake in both cases and also had the highest distribution of wake class. Conversely, Subject 9 had the second-highest distribution of wake class but displayed one of the weakest performances. Subject 13, who had the lowest distribution of sleep, demonstrated better performance than many of the subjects with higher distributions of wake.

3.2 EEG channel selection using NSGA-III for sleep-wake classification

This experiment aims to explore the feasibility of achieving comparable outcomes by utilizing fewer channels compared to the standard 128 high-density EEG channels. The average results among all subjects from NSGA are presented in Table 3. Given that not all subjects generated NSGA results on the Pareto front for each number of channels, the average reported in the table is based on varying numbers of subjects. When the same subject yielded multiple results for

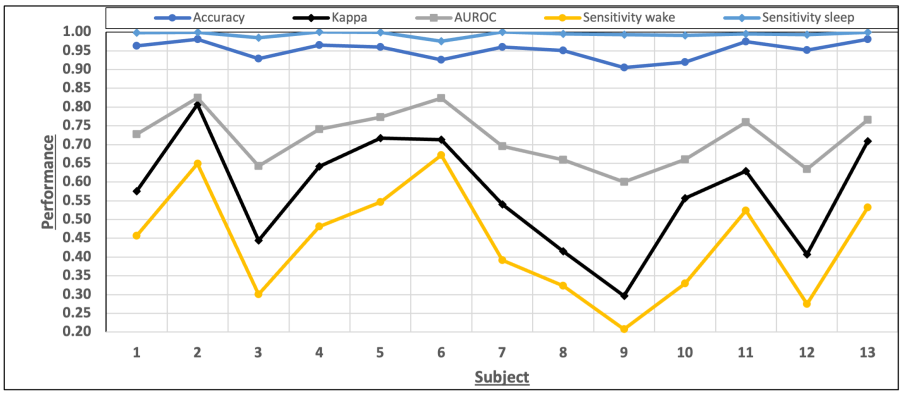


Fig. 2: Performance of sleep-wake classification for all subjects using the AASM-recommended channels; A28, B22, and C4.

Table 3: Average results obtained with 3-10 channels selected by NSGA for sleep-wake classification, comparing the performance using the 128 channels and the 3 AASM-recommended channels.

No. channels	**Accuracy**	**Kappa**	**AUROC**
3	0.955±0.023	0.678±0.097	0.790±0.075
4	0.966±0.019	0.719±0.100	0.811±0.066
5	0.965±0.020	0.714±0.129	0.814±0.081
6	0.970±0.017	0.757±0.101	0.832±0.065
7	0.972±0.014	0.781±0.085	0.849±0.060
8	0.973±0.014	0.783±0.075	0.855±0.056
9	0.975±0.015	0.795±0.075	0.855±0.047
10	0.974±0.013	0.790±0.067	0.853±0.047
128	0.970±0.018	0.778±0.086	0.855±0.051
AASM-recommended	0.951±0.024	0.573±0.149	0.716±0.073

the same number of channels, the best result was selected for inclusion in the analysis.

All the various channel combinations resulted in an accuracy of at least 0.955 according to Table 3. As discussed, accuracy alone does not reflect the model's ability to detect wake instances, which is the main objective of this study. With 9 channels, the results were slightly better for all metrics compared to 10 channels. The results are slightly decreasing from 9 channels and down to 3 channels. The difference between the maximum and minimum values for accuracy, kappa and AUROC were 0.2, 0.117 and 0.065, respectively. The use of 9 channels obtains the best results with an accuracy of 0.975 and a kappa score of 0.795, which is slightly higher than the use of all 128 high-density EEG channels with an accuracy of 0.97 and a kappa score of 0.778. All the NSGA results outperformed the AASM-recommended channels.

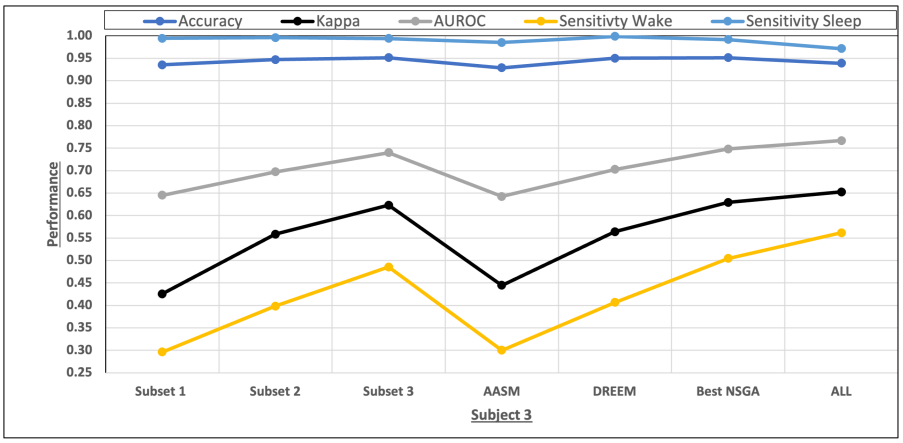


Fig. 3: Performance of sleep-wake classification for subject 3 based on different subsets of channels.

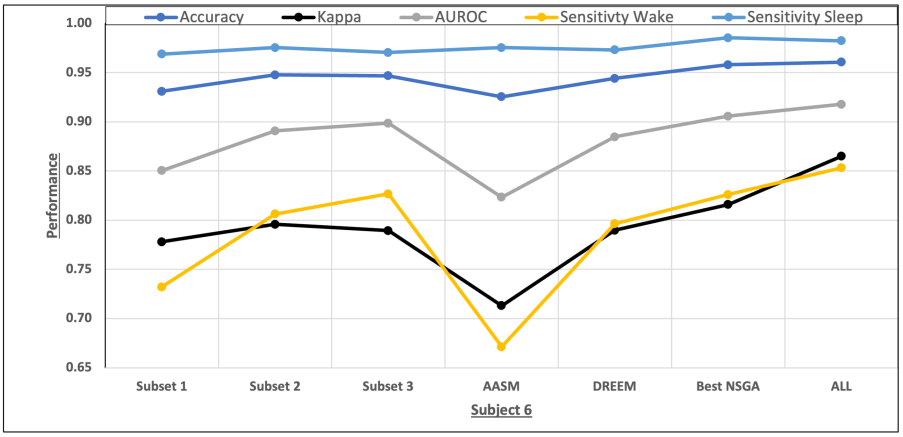


Fig. 4: Performance of sleep-wake classification for subject 6 based on different subsets of channels.

NSGA-III found that the most relevant channel for sleep versus wake classification based on all 13 subjects was D25, which was present in 12 subjects. Both channels A19 and B14 were present in 10 subjects. Channel A14, B29 and D4 were present in 9 subjects, while channel B11 and B31 were present in 8 subjects. These results and the frequency of appearance for the other channels are illustrated in the heatmap of Figure 5. The channels not present in the heatmap were not part of the solutions. The AASM-recommended channels, A28 and B22, were not included in any of the optimal subsets found by NSGA. Channel C4 was only present for one of the subjects.

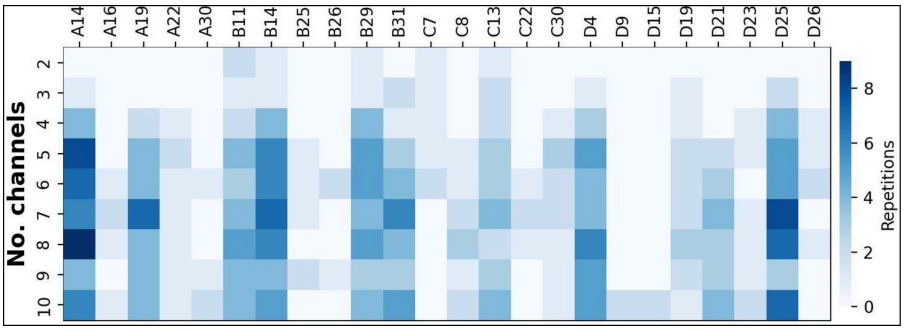


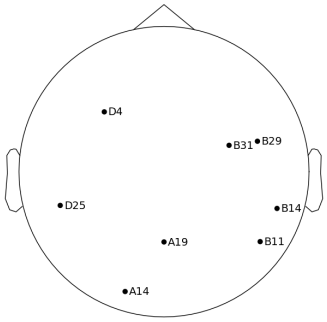
Fig. 5: Heatmap of channel importance in sleep-wake classification based on the results from NSGA for all subjects.

Based on the most important channels identified, D25 and A14 are located in the back left region of the brain, while B29 is located in the right frontal region. In addition, D4 is located in the left frontal region and B11 is in the right back region of the brain. It can be inferred from these observations that by including four of these important channels, the entire brain is sufficiently covered.

3.3 Comparison of Results for Subject-Specific Analysis

From the results of all the channels in Figure 1, subject 6 obtained the best overall results, while subject 3 obtained the worst overall results. These two subjects are chosen for the following analysis. Using the most common channels from the NSGA analysis, the obtained results are compared with those using all channels, AASM-recommended channels, and the subset with the best-obtained channels for that specific subject from the NSGA analysis. The three subsets of the most common channels from the NSGA analysis is listed in Table 6.

The results for subject 3 are presented in Figure 3, and the results for subject 6 are presented in Figure 4. The accuracy and sleep sensitivity values are consistently considered good for both subjects, above 0.95 and above 0.90, respectively, for all cases. Accuracy, kappa, and AUROC are metrics that pertain to the overall performance of the classification model. At the same time, sensitivity to the individual classes is an isolated metric, we observe that the kappa score is the metric that best captures the sensitivity of sleep in the overall performance of the model. As expected, the best results are obtained using all 128 high-density EEG channels. What is promising is that the results with the corresponding best NSGA results are in the same range, and this is also the case for the overall subset based on the most frequently appearing channels from the NSGA results. In the case of subject 3, the subset consisting of 3 channels selected by NSGA performed equally poorly as the 3 AASM-recommended channels, whereas for subject 6, the AASM-recommended channels were clearly the worst-performing subset.



	Channels
Subset 1	D25, A19, B14
Subset 2	D25, A19, B14, A14, B29
Subset 3	D25, A19, B14, A14, B29, D4, B11, B31

Fig. 6: Positioning and subsets of channels obtained from the NSGA-III analysis.

4 Discussion and conclusions

In this study, we have presented experimental results of the performance of a recently proposed CNN architecture, EEGNet, that has shown higher performance than other approaches across different EEG-based paradigms. We present a set of experiments using high-density EEG and the AASM-recommended subset of channels. The obtained performance is higher when using the 128 available channels in the dataset. However, due to the non-portability and high computational cost, we have implemented a multi-objective algorithm to select the most relevant channels to be usable in wearable sleep devices to classify sleep versus wake EEG signal segments.

In the context of wake-sleep classification, accurately detecting the minority wake class is a critical factor, as wake periods during the night play a vital role in sleep quality and are often present in several sleep disorders. Notably, relying solely on high accuracy as a performance metric can be misleading, particularly when the dataset is highly unbalanced, as is often the case in sleep-wake classification tasks. Therefore, the kappa score has been proposed as a more consistent metric for capturing the model's ability to detect wake sensitivity. Despite this, accuracy remains the most commonly reported performance metric in classification problems.

Using EEGNet and the 128 available channels produces highly favorable results, as evidenced by the findings presented in Table 2 and Figure 1. However, this approach entails a significant computational demand. Should the objective be to accurately classify and detect sleep-wakefulness in real-time, it is imperative that reliance on substantial computational resources or memory usage be avoided, given that these programs are intended for deployment on compact hardware devices.

The results obtained from utilizing NSGA-III demonstrate that EEG signal classification with significantly fewer channels can achieve comparable performance to that obtained from using the original 128 channels. Various

subsets were obtained by extracting the frequency of appearance of all channels from the NSGA-III results. Notably, all of these subsets surpassed the AASM-recommended channels in terms of performance. Moreover, the subset derived from NSGA-III for the two chosen subjects individually outperformed the results of using the subsets of most common channels from Table 6.

Further work can be performed to tune and optimize the many hyperparameters in the optimization problem. Other parameters that may affect the results and which could be included in the NSGA search are the sampling frequency, the segment size, or the various parameters of the classification algorithm.

By comparing the results to the leading state-of-the-art, they are not as high. However, the underlying assumptions of these results must be taken into consideration. The current state-of-the-art uses a publicly available and fairly balanced dataset, even with a higher proportion of wake samples than sleep samples. This is a significant difference from the distribution of wake-sleep samples in the dataset used in this study, which ranges from 4.2 to 16.4% of wake samples. Future work should be conducted using raw data from sleep labs, such as in this experiment, but with a more balanced dataset to determine if the results can match the state-of-the-art with a different methodology. An experiment for future work could involve collecting more wakefulness data from patients before they go to sleep to balance out the data.

Several experiments from the literature report results with a single EEG channel pair. Further work should be performed on testing the same type of channels with the unbalanced and raw dataset used in this study to see if the performance is promising.

These findings suggest that selecting a single set of channels may not provide optimal performance for every individual. Further research is of interest in customizing the channel selection process for each subject to achieve optimal performance in sleep stage classification using EEG signals. By using a general model of all subjects together, may obtain a unique subset for the whole dataset.

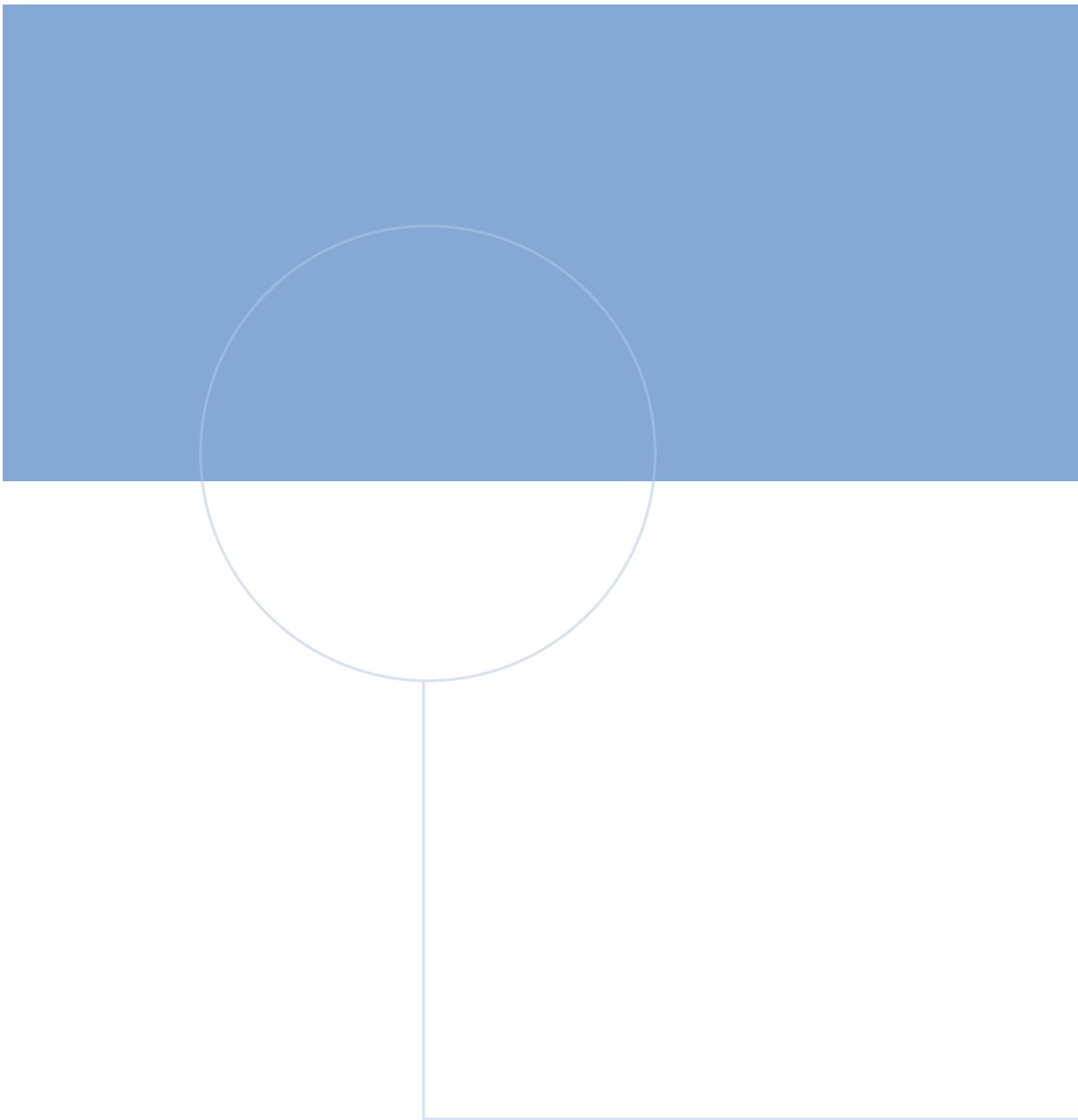
One of the main problems for sleep staging is the high difference in the number of instances across the sleep stages, which makes it more difficult for the algorithms to learn how to discriminate between the defined classes. For this purpose, further work must include the use of methods for data augmentation, as well as considering a balanced dataset created with data from multiple subjects.

Acknowledgement This work was supported by the Japan Society for the Promotion of Science (JSPS) Postdoctoral Fellowship for Research in Japan: Fellowship ID P22716, and JSPS KAKENHI: Grant number JP22K19802 and JP20K03493.

References

1. Arnal, P.J., Thorey, V., Ballard, M.E., Hernandez, A.B., Guillot, A., Jourde, H., Harris, M., Guillard, M., Van Beers, P., Chennaoui, M., et al.: The dreem headband

as an alternative to polysomnography for eeg signal acquisition and sleep staging. BioRxiv p. 662734 (2019)
2. Arnal, P.J., Thorey, V., Debellemaniere, E., Ballard, M.E., Bou Hernandez, A., Guillot, A., Jourde, H., Harris, M., Guillard, M., Van Beers, P., et al.: The dreem headband compared to polysomnography for electroencephalographic signal acquisition and sleep staging. Sleep **43**(11), zsaa097 (2020)
3. Berthomier, C., Drouot, X., Herman-Stoïca, M., Berthomier, P., Prado, J., Bokar-Thire, D., Benoit, O., Mattout, J., d'Ortho, M.P.: Automatic analysis of single-channel sleep eeg: validation in healthy individuals. Sleep **30**(11), 1587–1595 (2007)
4. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: Nsga-ii. IEEE transactions on evolutionary computation **6**(2), 182–197 (2002)
5. Espie, C.A.: Insomnia: conceptual issues in the development, persistence, and treatment of sleep disorder in adults. Annual review of psychology **53**(1), 215–243 (2002)
6. Fiorillo, L., Puiatti, A., Papandrea, M., Ratti, P.L., Favaro, P., Roth, C., Bargiotas, P., Bassetti, C.L., Faraci, F.D.: Automated sleep scoring: A review of the latest approaches. Sleep medicine reviews **48**, 101204 (2019)
7. Hassan, A.R., Bashar, S.K., Bhuiyan, M.I.H.: On the classification of sleep states by means of statistical and spectral features from single channel electroencephalogram. In: 2015 International conference on advances in computing, communications and informatics (ICACCI). pp. 2238–2243. IEEE (2015)
8. Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J.: Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. Journal of neural engineering **15**(5), 056013 (2018)
9. Lee, C.H., Kim, H.J., Heo, J.W., Kim, H., Kim, D.J.: Improving sleep stage classification performance by single-channel eeg data augmentation via spectral band blending. In: 2021 9th International Winter Conference on Brain-Computer Interface (BCI). pp. 1–5. IEEE (2021)
10. Moctezuma, L.A., Abe, T., Molinas, M.: Eeg-based 5- and 2-class cnn for sleep stage classification. In: The 22nd World Congress of the International Federation of Automatic Control (2023)
11. Nano, M., Fonseca, P., Overeem, S., Vullings, R., Aarts, R.M.: Lying awake at night: cardiac autonomic activity in relation to sleep onset and maintenance. Frontiers in Neuroscience **13**, 1405 (2020)
12. Richard B, B., Claude L, A., Susan M, H., al, e.: The aasm manual for the scoring of sleep and associated events: Rules, terminology and technical specifications version 2.5. the American Academy of Sleep Medicine (2018)
13. Ronzhina, M., Janoušek, O., Kolářová, J., Nováková, M., Honzík, P., Provazník, I.: Sleep scoring using artificial neural networks. Sleep medicine reviews **16**(3), 251–263 (2012)
14. Själander, M., Jahre, M., Tufte, G., Reissmann, N.: EPIC: An energy-efficient, high-performance GPGPU computing research infrastructure (2019)
15. Zhao, W., Van Someren, E.J., Li, C., Chen, X., Gui, W., Tian, Y., Liu, Y., Lei, X.: Eeg spectral analysis in insomnia disorder: A systematic review and meta-analysis. Sleep Medicine Reviews **59**, 101457 (2021)
16. Zhu, G., Li, Y., Wen, P.: Analysis and classification of sleep stages based on difference visibility graphs from a single-channel eeg signal. IEEE journal of biomedical and health informatics **18**(6), 1813–1821 (2014)



NTNU

Norwegian University of
Science and Technology