CZECH TECHNICAL UNIVERSITY IN PRAGUE

Faculty of Electrical Engineering

Department of Electroenergetics

Bachelor Thesis

Hanseo Yoo

Classification of sleep stages from polysomnography and actigraphy

Thesis supervisor: Ing. Eduard Bakštein, Ph.D.

Study program: Electrical Engineering and Computer Science

December 2024

BACHELOR'S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Yoo Hanseo** Personal ID number: **498664**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Electrical Power Engineering**

Study program: **Electrical Engineering and Computer Science**

II. Bachelor's thesis details

Bachelor's thesis title in English:

Classification of sleep stages from polysomnography and actigraphy

Bachelor's thesis title in Czech:

Klasifikace spánkových stádií z polysomnografického záznamu a aktigrafie

Guidelines:

Polysomnography (PSG) is a gold standard diagnostic method for sleep disorders, during which a set of biosignals including ECG, EEG, spirogram and others are recorded. One of the tasks in the evaluation of the recordings is the classification of sleep stages, which is normally performed manually. Nowadays, basic non-medical grade sleep diagnostics is provided by the widely used wrist-worn smart devices, based on actigraphy measurement.
The aim of the project is to develop a classifier for automatic assessment of sleep stages based on PSG and actigraphy and to evaluate to which degree the different biosignals contribute to the classification accuracy .
Guidelines:
- Study and summarize the existing methods for sleep stage classification from actigraphy and polysomnography
- Implement a set of classifiers, based on actigraphy and biosignals from the provided PSG dataset, such as the ECG or respiratory activity
- Evaluate the classification accuracy achieved by models based on different subsets of the provided biological signals.

Bibliography / sources:

Nahlik, M. (2019), Sleep-Wake and Sleep Stage Detection from Wrist-Worn Actigraphy, diploma thesis, CTU Prague
Altini, M., & Kinnunen, H. (2021). The promise of sleep: A multi-sensor approach for accurate sleep stage detection using the oura ring. Sensors, 21(13), 1–21. https://doi.org/10.3390/s21134302
Boe, A.J., McGee Koch, L.L., O'Brien, M.K. et al. Automating sleep stage classification using wireless, wearable sensors. npj Digit. Med. 2, 131 (2019). https://doi.org/10.1038/s41746-019-0210-1

Name and workplace of bachelor's thesis supervisor:

Ing. Eduard Bakštein, Ph.D. Analysis and Interpretation of Biomedical Data FEE

Name and workplace of second bachelor's thesis supervisor or consultant:

Date of bachelor's thesis assignment: **18.09.2024** Deadline for bachelor thesis submission: **07.01.2025**

Assignment valid until: **15.02.2026**

Ing. Eduard Bakštein, Ph.D. Supervisor's signature	doc. Ing. Zdeněk Müller, Ph.D. Head of department's signature	prof. Mgr. Petr Páta, Ph.D. Dean's signature

III. Assignment receipt

The student acknowledges that the bachelor's thesis is an individual work. The student must produce her thesis without the assistance of others, with the exception of provided consultations. Within the bachelor's thesis, the author must state the names of consultants and include a list of references.

.

Date of assignment receipt — Student's signature

Declaration:

I hereby declare that this bachelor's thesis is the product of my own independent work and that I have clearly stated all information sources used in the thesis according to Methodological Instruction No. 1/2009 – "On maintaining ethical principles when working on a university final project, CTU in Prague".

Date	Author's signature
....	....

Abstract

The main goal of this study is to design and evaluate algorithms for identifying different sleep stages using a combination of actigraphy and polysomnography recordings.

Actigraphy is a simple and convenient method to analyze movement activity and therefore provides a great potential for tracking sleep. However, while actigraphy is accurate in sleep-wake classification, it is much less accurate for identification of sleep stages. Polysomnography is the gold standard method for sleep evaluation, combining multiple biosignals, such as the EEG and ECG. It is more complex and expensive compared to actigraphy, especially when the patients need to attach many sensors and need to visit a specialized center with trained personnel.

Previous studies have used physiological signals, such as actigraphy and polysomnography, to classify sleep stages. In this thesis, we compared an initial model that uses actigraphy data with extended models, which include additional biosignals, such as electrocardiography (ECG) and leg movement. From the analysis, we confirmed that the addition of ECG signals played an important role in improving the accuracy of sleep stage classification.

The performance of the model was evaluated in two different scenarios: using sample-based validation and a patient-based cross-validation. In sample-based validation, the combination of actigraphy, ECG, leg movement signals and time since sleep start showed the highest performance with an accuracy of 0.69 and an F1 score of 0.67 in the classification of 30 secend sleep segments into 5 sleep stages. On the other hand, when using actigraphy alone, the accuracy and F1 score were low at 0.47 and 0.33, respectively, confirming the difficulty of accurately distinguishing sleep stages from actigraphy only. In a patient-based validation scenario, all models showed low performance. Even with all features combined, the accuracy was only 0.45, and the F1 score was 0.42, which is substantially lower than the results from sample-based validation. This suggests that the model has difficulty with generalizing data to new patients, due to existing systematic differences in physiological signals between patients.

This study shows that the accuracy of sleep stage classification can be improved by combining polysomnography features such as ECG and leg movement signals into actigraphy. However, in the stricter patient-based validation scenario, the model results were not sufficient. Since this study did not include many features of polysomnography, such as electroencephalogram (EEG) and respiratory, additional research might be required to achieve high accuracy and generalizability to new patients.

Keywords: Actigraphy, Polysomnography, sleep stage classification

Abstrakt

Hlavním cílem této studie je navrhnout a vyhodnotit algoritmy pro identifikaci různých fází spánku z kombinace záznamů aktigrafie a polysomnografie.

Aktigrafie je jednoduchá metoda analýzy pohybové aktivity, poskytující velký potenciál pro sledování spánku. Zatímco je však aktigrafie přesná při klasifikaci spánku a bdění, je mnohem méně přesná při identifikaci spánkových stadií. Polysomnografie je referenční metodou hodnocení spánku, která kombinuje více biosignálů, například EEG a EKG. Ve srovnání s aktigrafií je složitější a nákladnější, zejména je třeba pacientům připojit velký počet senzorů a je třeba navštívit specializované centrum s vyškoleným personálem.

Předchozí studie používaly ke klasifikaci spánkových stadií fyziologické signály, jako je aktigrafie a polysomnografie. V této práci jsme porovnávali původní model využívající údaje z aktigrafie s rozšířenými modely, které zahrnovaly i další biosignály, jako je elektrokardiografie (EKG) a pohyby nohou. Na základě analýzy jsme potvrdili, že přidání EKG signálů hraje důležitou roli při zlepšování přesnosti klasifikace spánkových stadií.

Výkonnost modelu byla hodnocena ve dvou různých scénářích: pomocí křížové validace na základě vzorků a křížové validace na základě pacientů. Při validaci na základě vzorku vykazovala kombinace aktigrafie, EKG, signálů pohybů nohou a času od začátku spánku nejvyšší výkon s přesností 0,69 a skóre F1 0,67 při klasifikaci 30-sekundových spánkových segmentů do 5 spánkových stadií . Na druhou stranu při použití pouze aktigrafie byla přesnost a skóre F1 nízké, a to 0,47, resp. 0,33, což potvrzuje obtížnost přesného rozlišení spánkových stadií pouze s jedním údajem. Ve validačním scénáři, založeném na pacientech, vykazovaly všechny modely nízkou výkonnost. I při kombinaci všech funkcí byla přesnost pouze 0,45 a skóre F1 0,42, což je podstatně méně než výsledky validace na základě vzorku. To naznačuje, že model má potíže se zobecněním dat na nové pacienty, a to kvůli existujícím systematickým rozdílům ve fyziologických signálech mezi pacienty.

Tato studie ukazuje, že přesnost klasifikace spánkových stadií na základě aktigrafie lze zlepšit přidáním dalších signálů. jako je EKG a signály pohybů nohou. V přísnějším validačním scénáři, založeném na pacientech, však výsledky modelu nebyly dostatečné. Vzhledem k tomu, že tato studie nezahrnovala další signály, dostupné v polysomnografickém vyšetření, zejm. elektroencefalogram (EEG) a dýchání, může být v příštím výzkumu k dosažení vysoké přesnosti a zobecnitelnosti na nové pacienty využito těchto signálů.

Klíčová slova: EEG: aktigrafie, polysomnografie, klasifikace spánkových stadií.

Table of Contents

List of Abbreviations

ACT – Actigraphy
PSD – Power Spectral Density
MAD – Mean Amplitude Deviation
IQR – Interquartile range
EEG – Electroencephalography, brain activity
EOG – Electrooculogram, eye movements
EMG – Electromyography, muscle movements
ECG – Electrocardiography, heart activity

1. Introduction

Sleep is one of the most essential functions for human health and well-being. It is as essential as eating or breathing, and without enough sleep, our bodies and minds cannot function properly. During sleep, the body repairs itself, the brain consolidates memories and processes information, and our energy is restored. Lack of sleep can affect our mood, cognitive abilities, physical health, and overall quality of life. Therefore, understanding and ensuring good quality sleep is fundamental for maintaining optimal health and functioning. [11] To better understand the impact sleep has on our lives, reliable methods to track sleep duration and assess sleep quality are essential.

Therefore, we are focusing on understanding the complex figure of sleep by using tools like actigraphy and polysomnography (PSG). Actigraphy is a device designed for the measurement of movement activity that people typically wear around the wrist. This tool provides continuous measurement of the wearer’s movement, from which useful information about the timing and extent of physical activity, circadian rhythmicity, and sleep can be extracted [23][24]. Compared to actigraphy, polysomnography provides us with more detailed information, as it measures the activity of the brain through electro-encephalography (EEG), heart rate (electrocardiogram, ECG), eye movements (electro-oculography, EOG), and how relaxed our muscles are. Based on these parameters, a very detailed analysis of a person’s sleep may be performed.

Previously, sleep stages were mainly analyzed through complex equipment such as PSG, which is a laboratory-oriented technique and has limitations that make it difficult to apply to real life. Recently, studies using simpler wearable signals such as actigraphy have been attempted. With the widespread worldwide use of smartwatches and other wearable devices with actigraphy measurement ability, the methods and tools to obtain additional health insights – such as sleep analysis – gain importance, as well as wide application possibilities.

Based on previous studies, this study attempted to increase the accuracy of actigraphy-based sleep stage classification by adding PSG signals to the actigraphy data. First, we created a model that only used the actigraphy data, and then we gradually added PSG and prior data to compare how the performance changes. By doing this, we identified the differences in sleep stage classification accuracy provided by the additional physiological signals.

2. Background information

2.1 Sleep stage

Sleep is very important for our health and can be divided into several phases, called sleep stages. According to the National Heart, Lung, and Blood Institute (NHLBI), there are two main types of sleep: Rapid Eye Movement (REM) sleep and Non-Rapid Eye Movement (Non- REM) sleep) [4].

Non-REM Sleep:
- Stage N1: The lightest stage, where we begin to relax and approach from wakefulness into sleep.
- Stage N2: A more restful sleep stage in which the heartbeat rate and muscular activity slow down.
- Stage N3: Known as deep sleep. It is a phase where it’s more complicated to wake up.

REM Sleep: Includes fast eye movements, and an active brain, but a relaxed body [4]. REM sleep contributes to learning and emotional equilibrium. The sleep stages change in 80 to 100 minutes cycles, each including Non-REM and REM stages. Usually, we run through these cycles several times per night, which helps our brain rest and facilitates different operations.

2.2 Actigraphy

Figure 1: Actigraphy device, ActiGraph wGT3X-BT [6]

Actigraphy is a technique to measure physical activity through recording of movement, typically using an actigraphy device worn on the wrist. Actigraphy is frequently used for studies of human rest-activity rhythm and sleep patterns, which can be derived from the actigraphy recordings [24]

The study in this thesis uses actigraphy for analysis of sleep stages, in particular, the ActiGraph GT3X+ to monitor activity profiles. The use of such a device is useful because it can easily be worn throughout the day and night without disturbing sleep. In this study, three-axial continuous acceleration data at 100Hz was collected while the participants were undergoing a whole-night sleep evaluation in a polysomnography laboratory. When compared to polysomnography, actigraphy is much cheaper and the participants have an opportunity to sleep in their beds, therefore the results become more natural and ecologically valid. Therefore, actigraphy is a practical tool for studying sleep, as it provides valuable insights into sleep patterns and movement without disturbing sleep, making it a great tool for researchers [6].

2.1.1 Actigraphy in sleep stage classification

In previous studies, actigraphy has been shown to accurately distinguish episodes of sleep from wakefulness [24] For example, Sadeh (1989) showed that actigraphy can be used to detect sleep and wake states. Later, Sadeh and Acebo (2002) discussed how actigraphy is useful for tracking sleep patterns. However, while actigraphy works well for distinguishing sleep from being awake, it is less accurate in identifying specific sleep stages. The studies of sleep structure using actigraphy have shown mixed results, Actigraphy is good at detecting when someone is asleep, but it is less accurate at identifying wakefulness because it can be affected by movement and individual differences (Martin & Hakim, 2011). Ancoli-Israel et al. (2003) also noted that while actigraphy is useful for studying circadian rhythms and identifying sleep problems, it is not precise enough to distinguish detailed sleep stages. These limitations suggest that actigraphy works best when used together with other methods, like PSG, to improve accuracy.

2.3 Polysomnography

Polysomnography (PSG) is a method to measure and evaluate a patient's sleep, and it is used in the study of sleep and as a diagnostic tool in sleep medicine.

PSG mainly involves Electroencephalogram (EEG) to monitor brain activity, Electrocardiogram (ECG), to monitor cardiac function, Oral and nasal airflow pressure and temperature (using a nasal pressure transducer and thermistor) to monitor breathing, an electrooculogram (EOG), to monitor for rapid eye movements, Electromyography (EMG), specifically the jaw muscle activity and anterior tibialis (restless legs), and video recording. [10] The resulting recording is called a polysomnogram and is commonly evaluated by a trained sleep expert for diagnosis and identification of sleep stages.

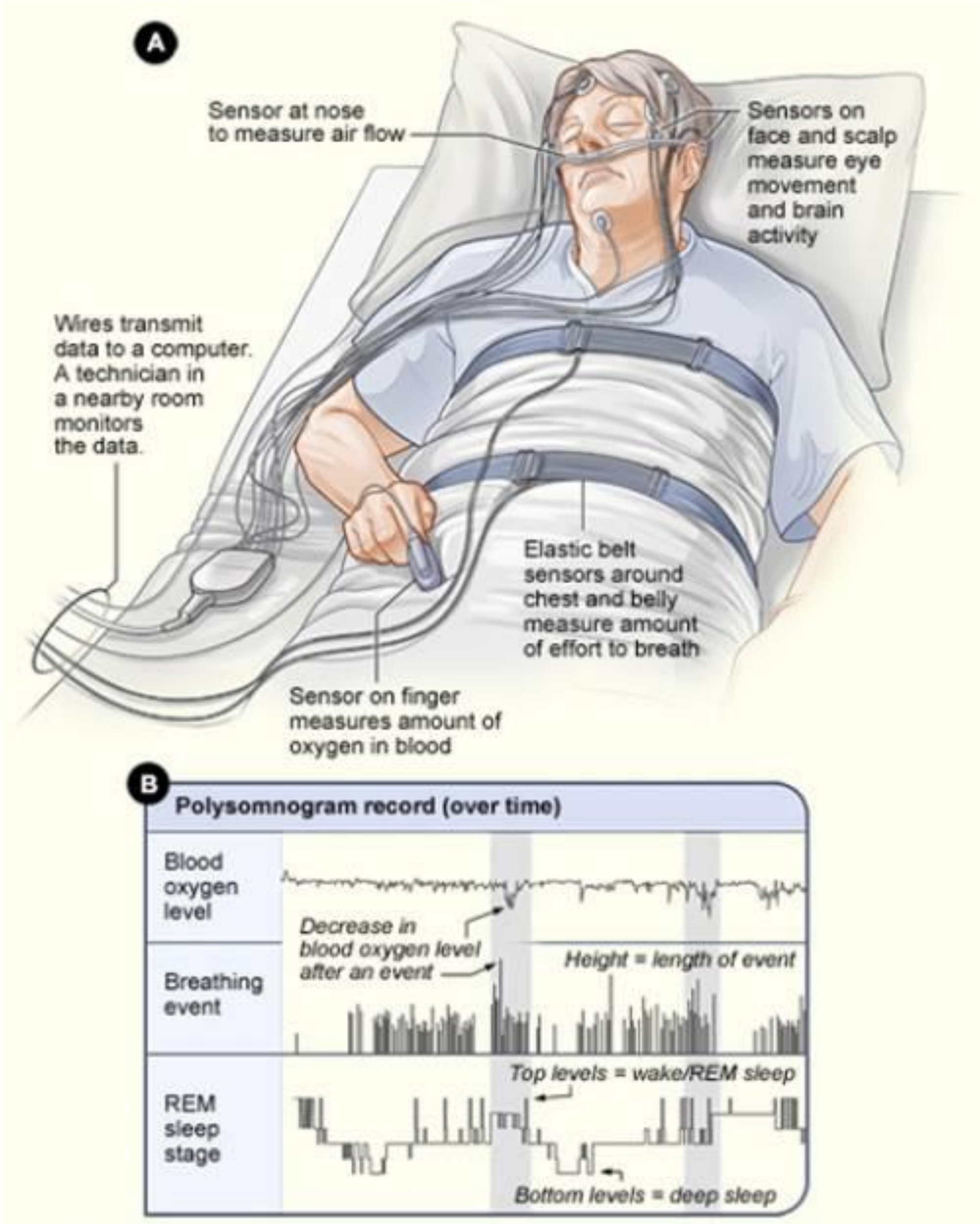

Figure 2: How Polysomnography works [10]

Simply put, PSG is a test to diagnose sleeping problems through various changes in the body's physiology that occur during sleep. During the test, the patient sleeps with the aforementioned sensors attached to their body. A technician attaches EEG electrodes to the head via an EEG cap, sensors to the chest, and nose to measure the degree of breathing. Other electrodes that measure eye movement, and muscle are also attached, and electrocardiogram electrodes are attached to the chest. Then, the patient goes to sleep, while his physiological functions are recorded. Sleep doctors analyse the recorded signals to evaluate sleep stages, changes in breathing, heart rhythms, eye movements and muscle contractions. The data collected in this way is used to analyse sleep quality or pathologies and identify the occurrence of sleep stages. The evaluation results are useful for detecting any diseases or disorders related to sleep.

Polysomnography is expensive and requires the patient to spend one night in a laboratory with several sensors attached. This may be burdensome for the patient, as well as affect the sleep quality, which may not correspond to the normal situation when the patient sleeps at home. Nevertheless, polysomnography is an important basis for accurately diagnosing various types of sleep disorders, assessing their severity, and determining the most appropriate treatment.

2.3.1 Electrocardiography

Electrocardiography (ECG) is a method to measure the electrical activity of the heart, measuring the timing of the different phases of heart activity. In this study, ECG was used as a part of PSG to analyse heart activity during sleep. From the ECG signal, we derived the measured the heart rate and heart rate variability by analyzing the recorded ECG. To identify the heartbeat, we calculated heart rate interval by finding the R peak and RR interval in the ECG signal, and based on this, we derived the mean heart rate and HRV. The BioSPPy library in Python (biosppy.signals.ecg) was used for R peak detection.

Heart Rate Variability

Heart rate variability (HRV) is used to understand the condition of the autonomic nervous system by measuring the change in the time interval between heart rates. HRV can be calculated as the standard deviation of the RR interval, which helps determine the regularity of the heartbeat or the extent of change. HRV is calculated as follows:

$$HRV = std(RR\ interval) \tag{1}$$

If the HRV value is too large (e.g. over 500ms), it can be considered an abnormal value and removed. [27]

Mean Heart Rate

The Mean Heart Rate is the heart rate per minute (bpm) and is used to understand changes in heart activity during sleep. The mean heart rate is calculated by the RR interval, which is the time interval between the two R peaks in the ECG signal. After calculating the mean RR interval, we can calculate the mean heart rate using the following equation:

$$Mean\ heart\ rate\ (BPM) = \frac{60000}{mean\ RR\ intervals(ms)} \tag{2}$$

60000 means a minute (in milliseconds) of time. In addition, if the mean heart rate is out of the normal range of 40 to 100 bpm, it is considered an abnormal value and it should be filtered. This is to increase the accuracy of the analysis by excluding abnormal results.

Analysis using ECG allows precise measurement of changes in heart rate during sleep, allowing a better understanding of body reactions in various sleep stages. [15]

2.3.2 Muscle activity

Leg movement data is typically captured during PSG to analyse physical activity during sleep for diagnostic purposes. During data preprocessing, we binarize the leg movement data in the following way: Movement is determined when the size of the movement exceeds a certain threshold, and the ratio of movement over a certain period (epoch) is calculated. The threshold used is determined by plotting the distribution of leg movement as a histogram.

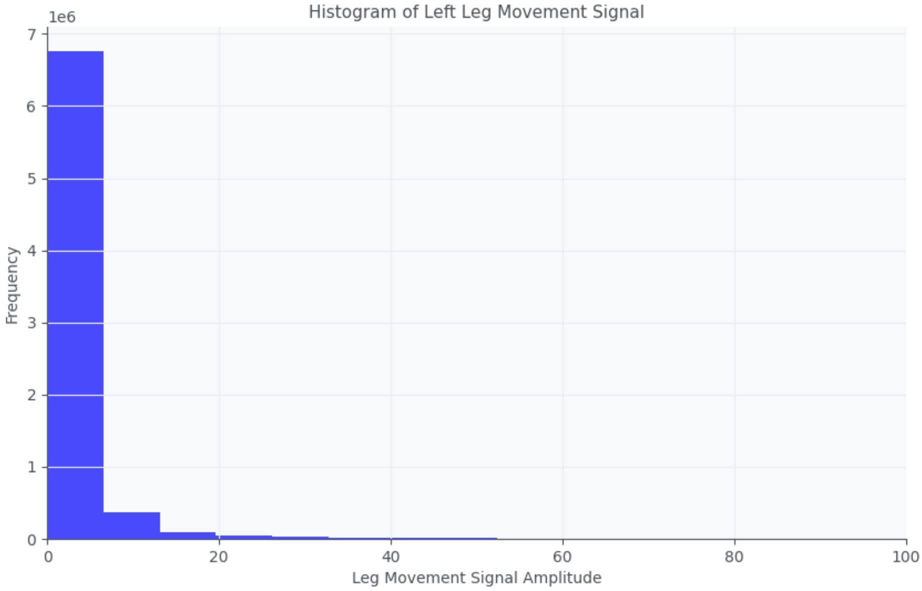

Figure 3: Histogram of Left Leg Movement Signal (Linear Scale)

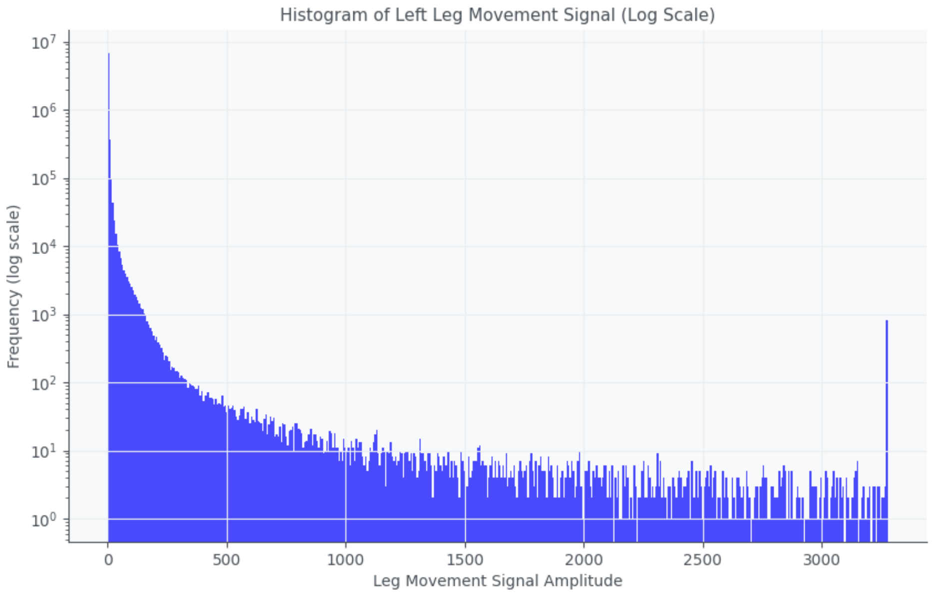

Figure 4: Histogram of Left Leg Movement Signal (Logarithmic Scale)

We chose the threshold value of 10, based on the leg movement activity histograms in Figures 3 and 4. The linear-scale histogram shows a significant concentration of data with amplitudes below 10, which means that values in this range are noise. The log-scale histogram shows that amplitudes of 10 or greater value represent meaningful movements while higher thresholds capture only rare occurrences. Setting the threshold at 10 ensures that the analysis includes sufficient leg movements for meaningful results while effectively filtering out noise during the periods of no movement, which are prevalent during sleep.

The purpose of this analysis using leg movement is to identify movement during sleep in a similar way to actigraphy. Based on leg movement data, this study was able to analyse the frequency of movement for each stage of sleep.

2.3.3 Hypnogram

Hypnogram is a graph showing a change according to time. Generally, the sleep step is divided into an Awake, S1, S2, S3, and REM sleep stage may show how these steps are converted depending on time. Hypnograms allow you to easily determine the duration and frequency of sleep stages.

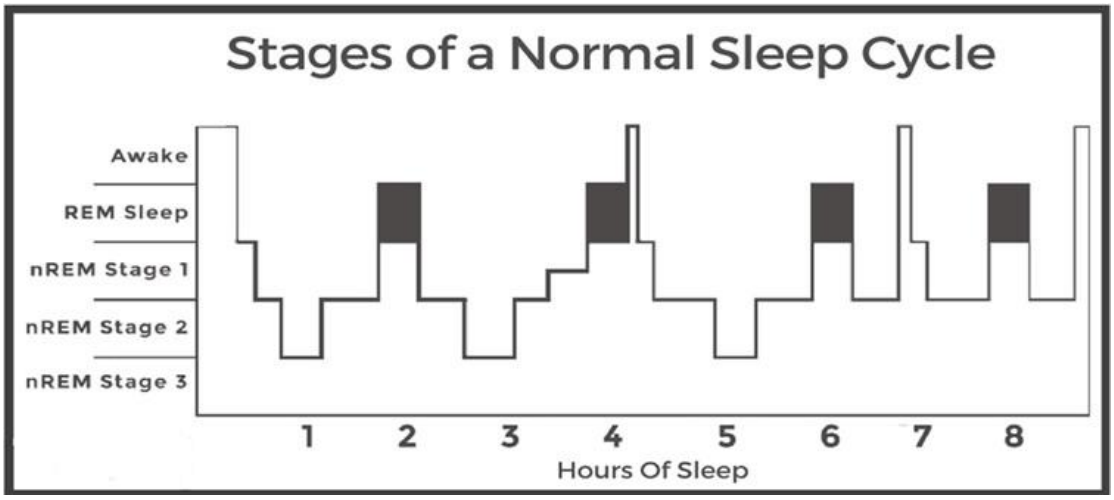

Figure 5: Hypnogram [14]

3 Technical background

3.1 Machine Learning Approaches

Machine learning is a technology creating algorithms to find patterns in data, learn themself, and predict outcomes for other data. Traditional programming requires humans to code all rules, but machine learning makes computers learn rules on their own using data. In other words, it is the way computers solve problems through algorithms based on data without instructions from humans. Machine learning can be categorized into supervised and unsupervised learning. Supervised learning uses labeled datasets to train algorithms to classify data or predict outcomes accurately. By comparing its predictions to the labeled outputs, the model improves its accuracy over time. In contrast, unsupervised learning analyzes and clusters unlabeled data, discovering hidden patterns without human intervention. These two approaches enable machine learning to solve a wide range of problems effectively. [25]

3.1.1 Logistic regression

Logistic regression is defined as a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation. [15] Logistic regression uses the Sigmoid function to transform the output value into a probability value between 0 and 1, based on which value the given data falls into which category.

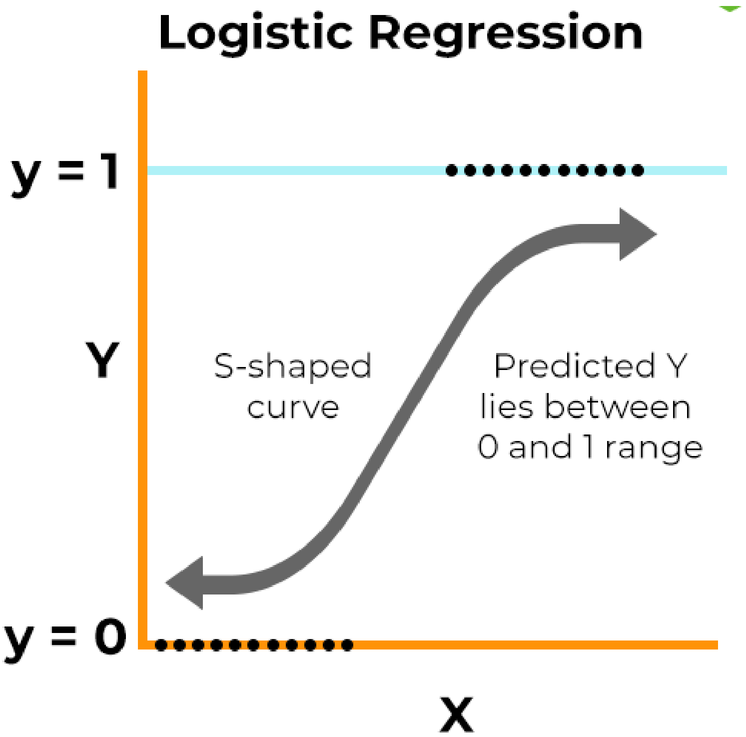

Figure 5: Logistic regression explanation [15]

The model has the following equation:

$$p = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 \cdots b_n x_n)}} \tag{3}$$

p… the probability of the outcome being 1, given input data x..
b0, b1.., these represent the regression coefficients, which indicate how much the individual dimensions of the input data influence the outcome.
x1, x2… these are the input data (features).

The sigmoid function converts the result of linear regression into probability. By thresholding the resulting probability p, we can perform the binary classification task: If the value is greater than 0.5, then it classifies it into one class and if it is less than 0.5, then it classifies it into another class. For example, in sleep analysis, logistic regression can be used to calculate the probability that a sleep stage is REM.

There are two ways to solve the multi-class problem. The first way is the One-vs-Rest (OvR) method, which trains a binary classifier by comparing each class to all other classes individually. The second way is the softmax method, which considers all classes simultaneously to calculate the probability distribution for each class [25]. In this analysis, softmax method was used for multi-class logistic regression.

3.1.2 Random forest

Random forest is a powerful machine learning algorithm that is widely used for data classification. The algorithm creates multiple decision trees combines the prediction results of each of them and then makes the final result. A decision tree starts at the root and splits the data at each internal node based on specific variable conditions (a threshold on a single variable) until it leads to a final prediction at each leaf node. A single decision tree is simple and easy to understand, but it often tends to overfit the data. Random forest combines many trees to overcome these weaknesses. Each tree is trained differently through random sampling of training data and random selection of features. The random forest then aggregates the results of individual trees. [16]

Random forest is good for learning different data patterns and can make accurate predictions, even if the data is complex. For instance, to classify deep sleep stages or awake stages, each tree from random forest can analyze the data differently, understand various patterns, and improve the accuracy of the final predictions. Also, the aggregation of multiple decision trees prevents overfitting and leads to more robust classification results.

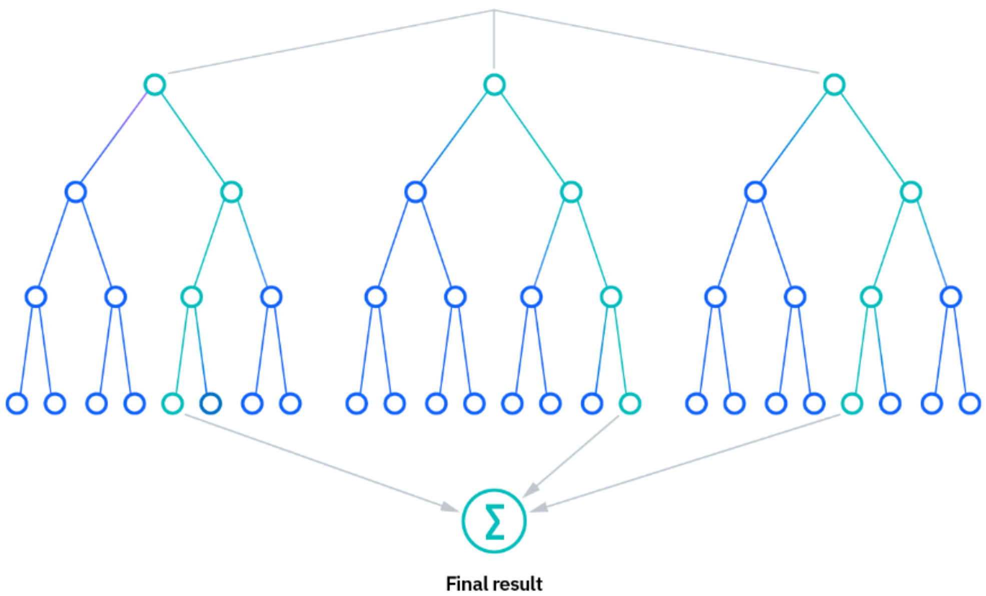

Figure 6: Random forest method explanation [16]

3.2 Classifier Evaluation

3.2.1 k fold Cross Validation

Cross validation is a method used to check how well a machine learning model works by splitting the data into parts. Normally, the dataset is divided into two sets: one is a train set and another one is a test set. After training the classifier (i.e. estimating the classifier parameters) on the train set, the model is evaluated on the test set to see how it performs on unseen data. Performing this split validation just once might not give a reliable result because it could be a lucky situation, dependent on chance. To obtain more reliable estimates of model performance, a more accurate strategy is to train and test the model multiple times. [17]

In k-fold Cross validation, the data is split into *k* groups. Each time, one group is used for testing, and the rest (*k*-1) are used for training. This process is repeated k times and results are aggregated over all *k* test sets, which gives a more reliable result than just testing once. For example, if *k* is set to 10, the method is called 10-fold Cross validation. This method is helpful as it shows how the model might perform on new data.

3.2.2 Confusion matrix and performance matrices

Confusion matrix

Confusion Matrix is a table that visually represents the relationship between predictive and real outcomes. This matrix provides information about how accurately the model predicted each class and what errors occurred. The matrix consists of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), where TP and TN represent when the model predicted correctly, and FP and FN represent when the model is predicted incorrectly. From this accuracy, precision, and reproducibility [18]

		Actual value	
		Positive	Negative
Predicted value	Positive	TP	FP
	Negative	FN	TN

Accuracy

Accuracy is an indicator of how accurately a model has classified the samples into correct classes. It represents the percentage of the values predicted by the model that match the actual values (classes) and is useful for checking overall performance. However, there is also a limitation that accuracy can be overestimated when the dataset is unbalanced with different ratios of examples in the negative and positive classes.

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{4}$$

Specificity

Specificity refers to the percentage of data that a model predicts to be negative that is actually negative, which shows how accurately the model predicts the negative class. It is useful when false positive predictions are important.

$$specificity = \frac{TN}{TN + FP} \tag{5}$$

Precision

Precision represents how much of the data that the model predicts as positive is actually positive. In other words, it shows how reliable the positive predictions are, and it is a useful indicator, especially when false positive predictions are important issues. Precision plays an important role in situations where false positive predictions can be costly.

$$precision = \frac{TP}{TP + FP} \tag{6}$$

Recall

Recall, often called sensitivity, represents how many positives the model accurately predicted among the actual positive data.

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

F1 score

The F1 score represents the inverse relationship of precision and recall and evaluates the balance between the two metrics. When an indicator is too high or too low, it is adjusted to provide a balanced performance evaluation. F1 score is useful for comprehensively evaluating the performance of a model in situations where both precision and recall are important and is especially useful in unbalanced datasets.

$$F1 = \frac{2Precision \cdot Recall}{Precision + Recall} \tag{8}$$

3.3 Used software

All calculations and analyses were made in the Python version 3. Specifically, machine learning models are evaluated for their performance using the Scikit-learn (sklearn) package, and Pandas, NumPy, and Matplotlib for data processing and visualization.

4 Methods

4.1. Problem definition

In this section, data from actigraphy and polysomnography is used to analyze and classify sleep stages. While actigraphy provides insights into patterns of movement during sleep, polysomnography recordings are represented in the form of a hypnogram, providing an expertlabel for sleep stages. The graphs show the relationship between these signals and the corresponding sleep stages and provide the basis for assessing the accuracy of sleep stage classification. This problem is addressed using machine learning methods in the following sections.

4.2 Experiment design

The data used in this thesis were collected from participants who underwent overnight polysomnography (PSG) while simultaneously wearing an ActiGraph GT3X+ device. Actigraphy data were recorded at 100 Hz and processed using ActiLife software (version 6.13.4) to extract movement information. Polysomnography data were initially recorded at 1000 Hz and then resampled to 250 Hz.

4.3 Dataset description

The data from polysomnography and actigraphy were provided by the National Institute of Mental Health in Prague, the Czech Republic, and contained data from 22 patients with a mean age of 36.6 years, and standard deviation age of 6.1 years. Among the patients, 19 were female, and 3 male.

The data was collected over 22 nights, with each participant contributing one full night of sleep recording. The dataset consists of approximately 173 hours of recording time, split into 20,766 epochs (30 second segments). On average, each patient contributed around 943 epochs.

4.4 actigraphy analysis and features

The figure below shows sleep data from the actigraphy device that was attached to the person's wrist. The graph in Figure 7shows the times the extent to which the person moved, which provides information about different parts of their sleep.

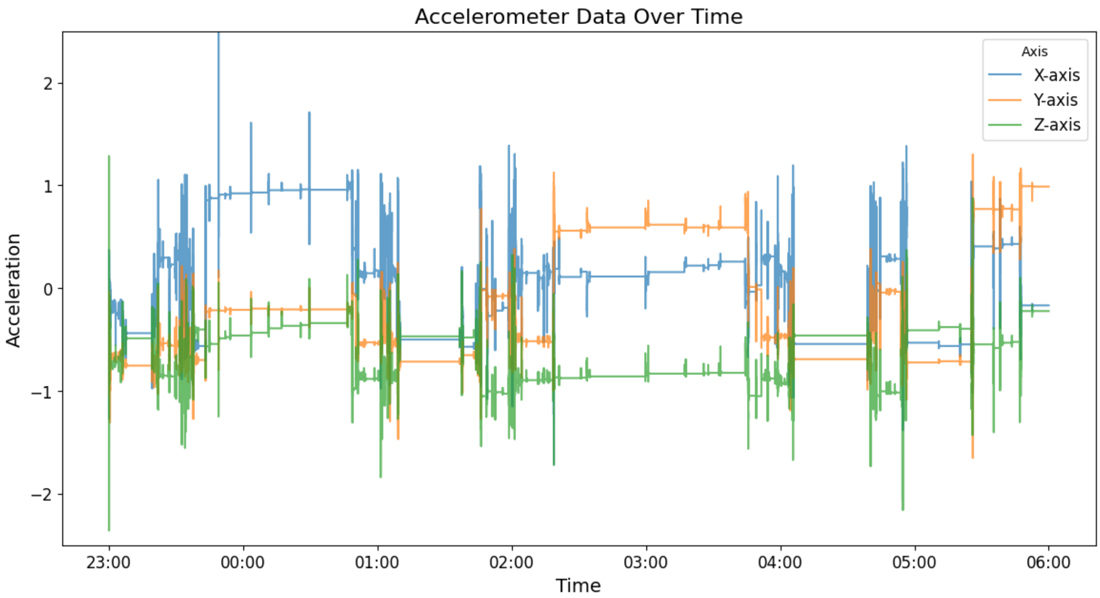

Figure 7: Accelerometer Data, data measured by actigraphy. Acceleration shown in [g] units

Figure 7 presents accelerometer data collected over time, using a 3-axis accelerometer to measure the movements of the person: the X-axis (blue), Y-axis (yellow), and Z-axis (green). The X-axis on the graph represents time, a day, from 23:00 on 9th April 2018, to 06:00 on 10th April. The Y-axis represents acceleration, where the values show how fast and in what direction the movement is happening. The spikes in the graph reflect periods of movement. Higher spikes can suggest more intense activity, like turning in bed or getting up. Changes in baseline in different axes represent changes in actigraphy orientation and thus different orientations of the ground acceleration.

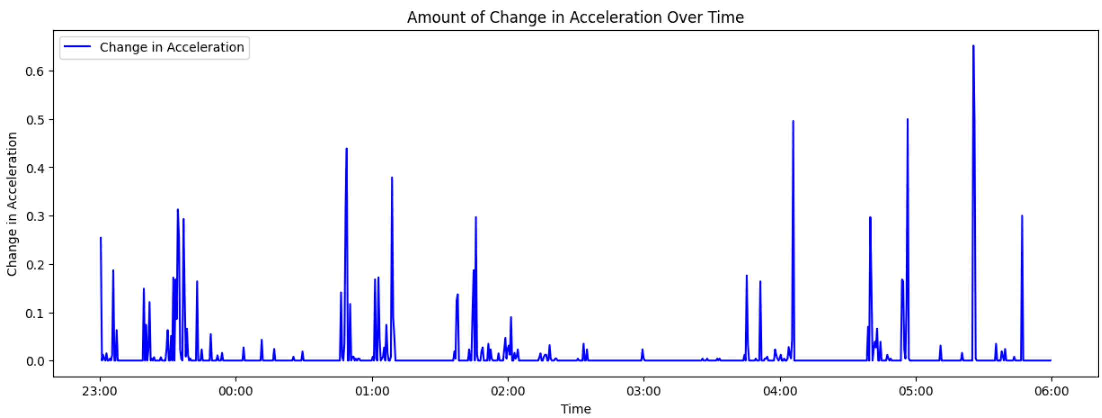

Figure 8: Amount of change in acceleration over time by theta z angle.

Figure 8 depicts the change in theta z variable. This axis is typically ranged in the direction of up and down movement of the forearm. For this sleep analysis, z-axis analysis is useful, because it measures vertical movements that can be associated with specific sleep behaviors. The spikes in the graph mean changes or movements detected on the vertical axis, which means the patient might have various activities or movements during sleep, for example flipping or changing posture. Theta z was calculated by dividing the Z-axis aceleration data into 30 second epochs and measuring the absolute difference between the median values of consecutive epochs.

4.5 Data processing

Bandpass filtering

We applied Bandpass filtering to Actigraphy data to analyse accurately. Bandpass filtering is a method that allows only a certain frequency range to pass through a signal and blocks the rest, thus elimination of unnecessary noise and leaving only meaningful signals. In this study, we used filters that leave only frequencies between 0.5 Hz and 4 Hz. In addition, the order of the filter is set to 4. These settings help to capture key body movement information effectively. The filtered data was then divided into each epoch and used to calculate the Actigraphy Mean.

We analyzed Actigraphy and Polysomnography data by dividing them into 30 second epochs. The 30 second epoch is suitable for sleep stage analysis and was used to efficiently observe body movements and physiological changes.

In the following part, Actigraphy was analysed by Mean Amplitude Deviation (MAD) method used in the study of [5] Altini & Kinnunen (2021). In previous studies, the accelerometer data was calculated in 5 second epochs to capture microscopic motion changes. However, in this study, MAD was calculated in 30 second epochs to match the epoch length of the PSG labels and motion changes at longer intervals were also analysed.

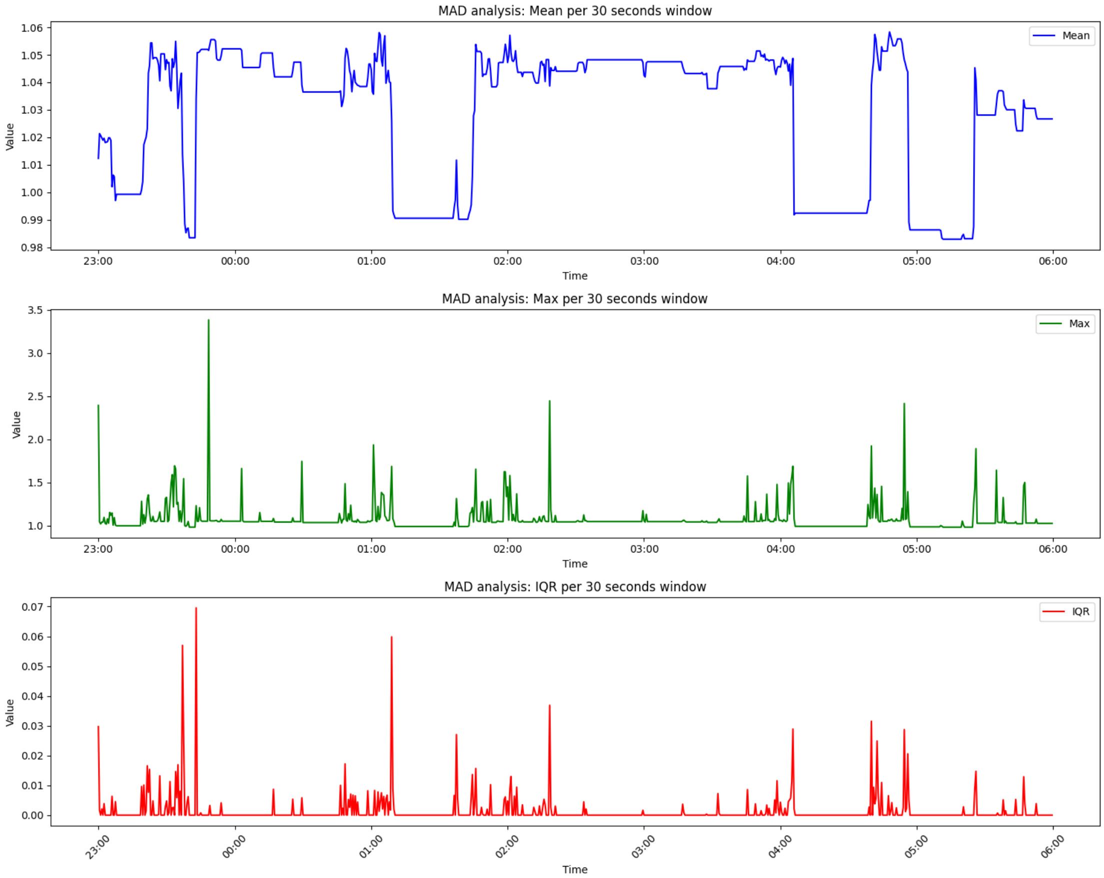

Figure 9: MAD method analysis

In this section, we computed short intervals of Mean Amplitude Deviation (MAD) to assess movement constancy. [7] One of the numbers we calculate to understand how much movement changes over time is called Mean Amplitude Deviation, MAD. This tool verifies movement after every 5 seconds. For MAD, we determine the average movement over these five seconds first by applying a term called mean value (R_ave). Then, we take each move within those 5 seconds and compare it with this average to see how much it deviates. Therefore, we sum up these differences and the average of them. This average is the MAD. It tells us whether the movements were average or considerably different. Figure 9 shows the results of the MAD analysis. The three plots in Figure 9 represent the mean, maximum, and interquartile range (IQR).

$$r_i = \sqrt{x_i^2 + y_i^2 + z_i^2},$$

r_i which means the magnitude of the acceleration vector.

$$R_{ave} = \frac{1}{N}\sum_{i=j}^{j+N-1} r_i$$

$$MAD = \frac{1}{N}\sum_{i=j}^{j+N-1} |r_i - R_{ave}| \quad (9)$$

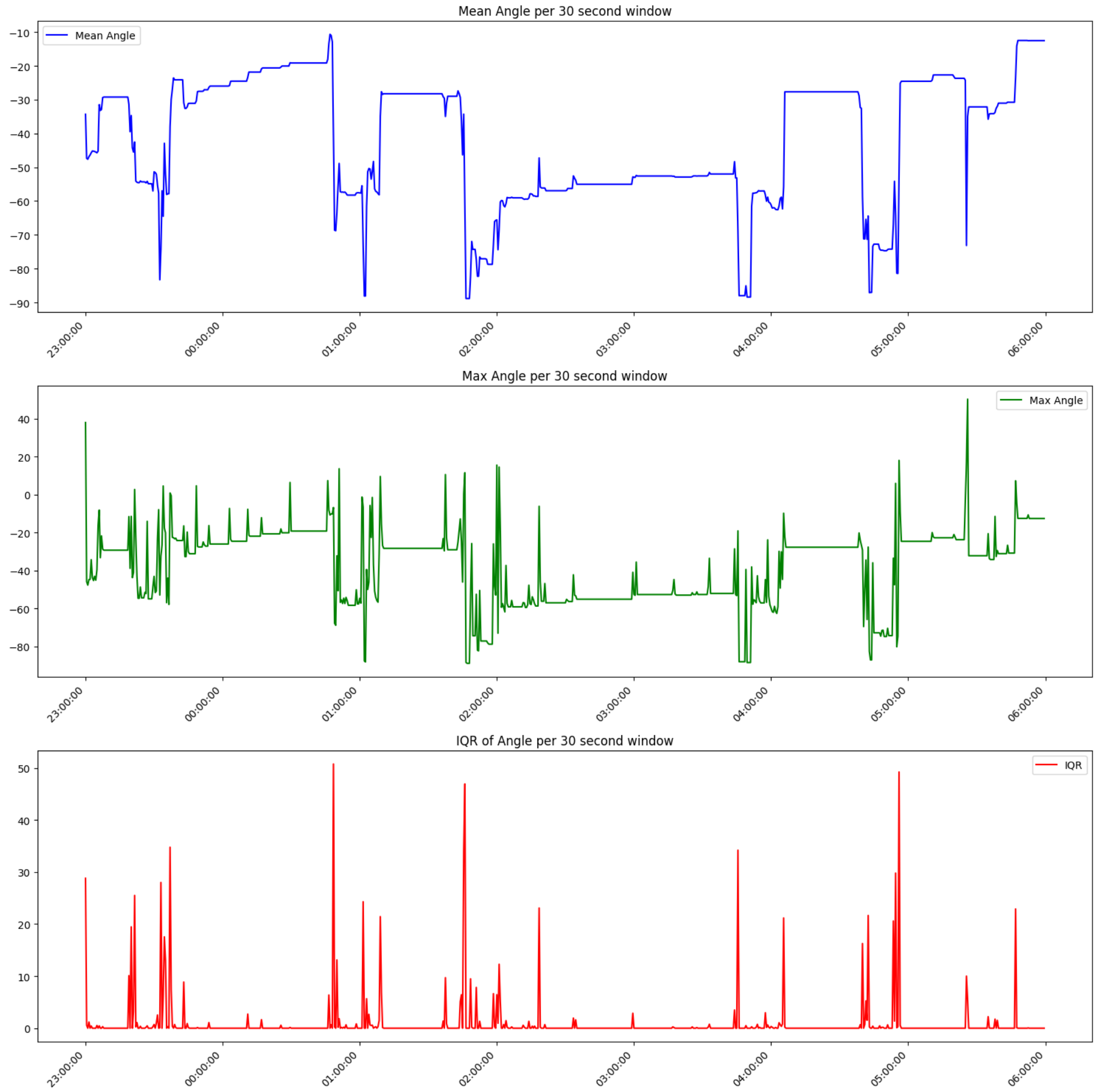

Figure 10: Wrist angel analysis

In these graphs, we observe the analysis of wrist angles from accelerometer data measured during the night, as shown as figure 10. This data allows us to know how much and when the patient moves during sleep. 3 orthogonal acceleration signals were used to calculate the angle of movement. This equation gives the value of angle [8].

$$angle = \left(tan^{-1} \frac{\theta_z}{\sqrt{\theta_x^2 + \theta_y^2}} \right) \cdot 180/\pi \tag{10}$$

where θ_x, θ_y, and θ_z are the median values of the three orthogonally positioned raw acceleration sensors in g-units derived based on a 30 second epoch. [8]

4.6 Actigraphy and Polysomnography Analysis

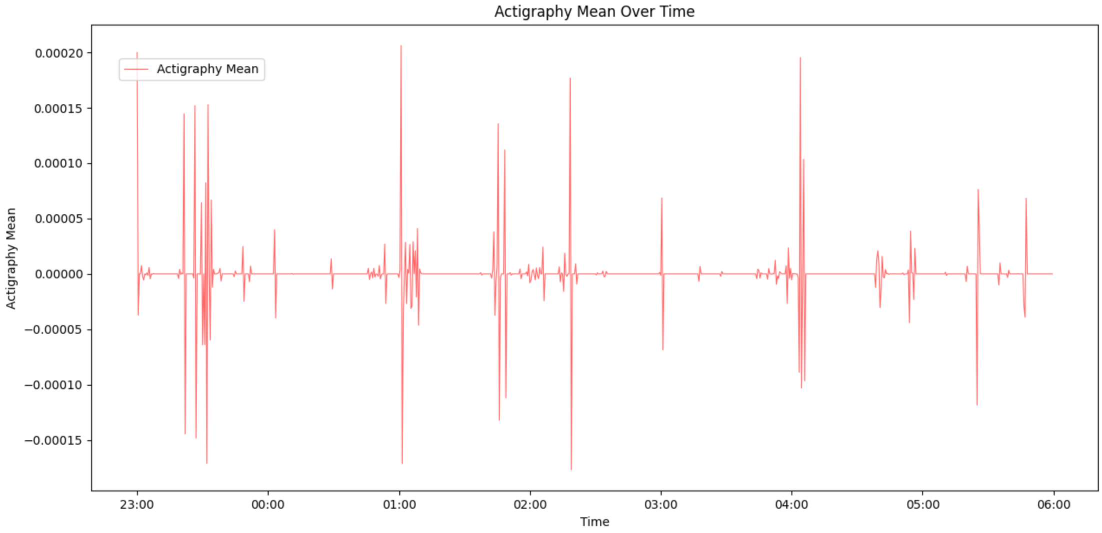

Figure 11: Actigraphy mean

This Figure 11 represents the average value of Actigraphy data divided by the 30 second epoch after bandpass filtering. Low frequency and high frequency noise that is not needed was removed through bandpass filtering, and only important frequency bands of body movement were extracted. The filtered data was divided into 30 second units to calculate the average of activities in each time interval, and this was visualized to clearly express the change in body movement over time. As you can see in the graph, you can see a pattern of rapid increase in activity during certain periods, which shows awake or high movement sections.

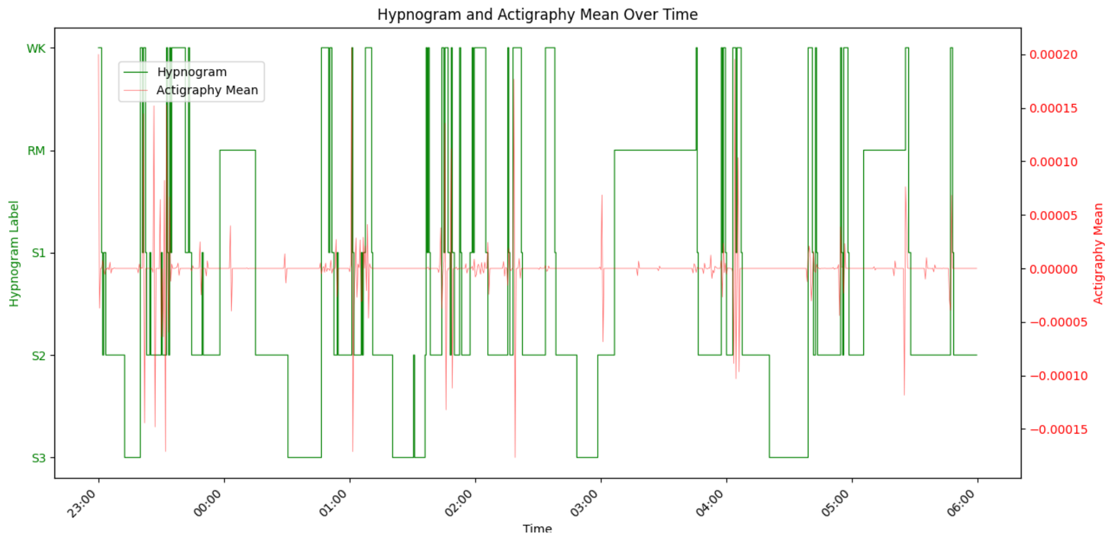

Figure 12: Hypnogram and Actigraphy mean

The figure 12 is a visualization of the Hypnogram and Actigraphy Mean together, and it shows the relationship between sleep stages and physical activity. The hypnogram contains WK, REM sleep, and sleep stages (S1, S2, S3), and you can see how they change by stages while the patient sleeps. Actigraphy Mean measures physical activity, representing the average body movement while the patient sleeps.

This graph shows that physical activity is higher during the awake stage, and the Actigraphy Mean decreases as sleep gets deeper. Especially for deep sleep stages, there is little movement, and activity increases during REM stage and awake stage.

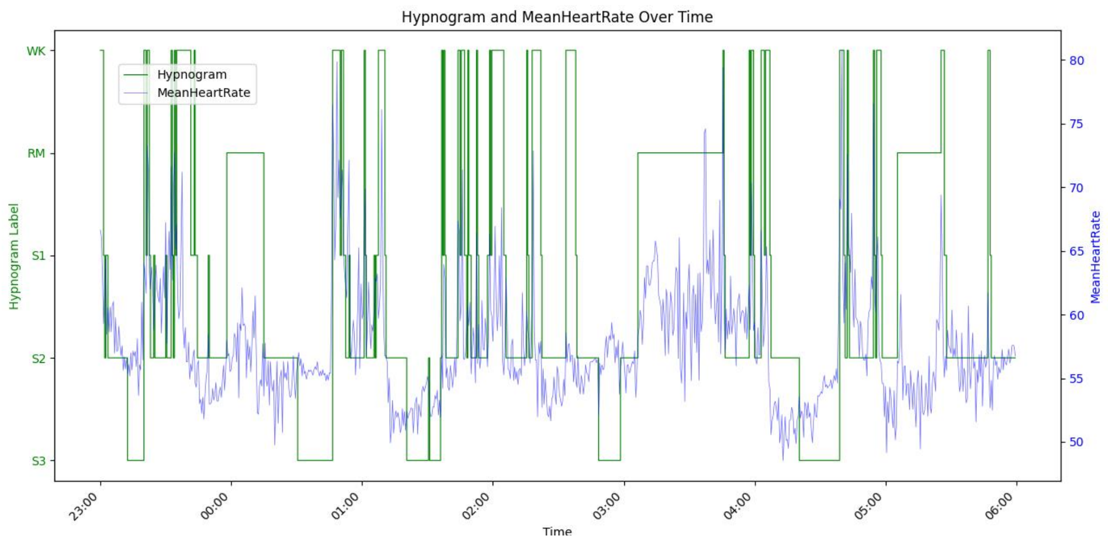

Figure 13: Hypnogram and Mean Heart Rate

The figure 13 represents changes in Hypnogram and Mean Heart Rate, which makes it easy to understand how patient heart rate changes depending on the stage of sleep. In the state of being awake stage (WK), the patient's heart rate is relatively high because their body is awake, active, and uses up a lot of energy. Also, in the REM sleep stage (RM), the heart rate remains relatively high. On the other hand, in deep sleep stages (S2, S3), the heart rate decreases, when the patient rests, the body temperature drops and the energy consumption decreases. The heart enters a stable and slow beating state, which is an important time for the human body to recover.

The Hypnogram + Mean Heart Rate and Hypnogram + Actigraphy Mean graphs show how ECG and actigraphy data change during different sleep stages, and it makes clear how each type of data helps predict sleep stages. ECG data are useful for representing sleep stage transitions through changes in heart rate, and actigraphy data can capture changes in sleep stages based on body movements.

4.7 Feature combinations

In this study, we compared the performance of our model using four feature combinations. The All feature combination includes all features, with leg movement frequency (left leg frequency, right leg frequency), ECG features (mean heart rate, HRV), actigraphy mean, and time since sleep start. The Act + Leg + ECG combination combines leg movement frequency, actigraphy mean, and ECG features. The Act + ECG combination is ECG features and actigraphy mean. Finally, Act only is a minimal combination of models that have only actigraphy mean. These various combinations allow us to evaluate which features are most effective in predicting sleep stages.

Feature	Model All	Model Act + ECG + Leg	Model Act + ECG	Model Act
Actigraphy	✓	✓	✓	✓
HRV	✓	✓	✓	
Mean heart rate	✓	✓	✓	
left leg frequency	✓	✓		
right leg frequency	✓	✓		

time since sleep start	✓			

Table 1: Features in each model

4.8 Sample-based method and patient-based method

Sample-based and patient-based methods are two approaches to validate the model. In sample-based, the whole data is divided into individual sample units, each of which is independently used for learning and validation. This helps the model to learn various data patterns, which is advantageous for the prediction of each sleep stage. On the other hand, patient-based evaluates whether the model can be generalized to new patients because data of each patient is divided separately for learning and validation. This approach tests how well the model works in real situations by considering the differences between individual patients.

5 Results

5.1 Dataset Overview

The dataset used in this study includes labels for five different sleep stages: Awake (WK), REM (RM), Stage 1 (S1), Stage 2 (S2), and Stage 3 (S3). These labels were taken from polysomnography (PSG) recordings. The dataset is composed of a total of 20,766 epochs, each representing a 30-second segment of data. The sleep stages are not evenly distributed in the dataset, with Stage 2 (S2) being the most frequent and Stage 1 (S1) being the least frequent, as shown in table 2.

Sleep stage label	Number of epochs	Percentage of Dataset
Stage 2 (S2)	9349	45%
Stage 3 (S3)	4857	23%
REM (RM)	3288	16%
Awake (WK)	2153	10%
Stage 1 (S1)	1119	5%

Table 2: Percentage of sleep stage label

We evaluated the performance of machine learning models and compared various feature combinations By analysing how each feature combination affects sleep stage prediction, we were able to find the best combination. This analysis helped to determine which data features and combinations were most beneficial for sleep research and to increase the accuracy of sleep stage predictions.

5.2 Model Effectiveness

We evaluated several machine learning models in the k-fold crossvalidation scheme. First, we used a logistic regression model for sleep stage prediction, but the accuracy was very low, as seen in table 3. The logistic regression model did not reflect the complex correlation between sleep stages well, so later analysis focused on getting the best performance with the Random Forest model. The Random Forest model was able to learn nonlinear relationships more effectively, resulting in higher accuracy in sleep stage prediction. Both models were evaluated using the Model All, which includes LegLeftFreq, LegRightFreq, MeanHeartRate, HRV, Actigraphy_Mean, and TimeSinceSleepStart.

	Accuracy	F1 Score
Logistic regression	0.514	0.473
Random forest	0.701	0.684

Table 3: Crossvalidation (test-set) accuracy of the selected machine learning models

5.3 Feature comparison

5.3.1 Feature comparison, sample-based

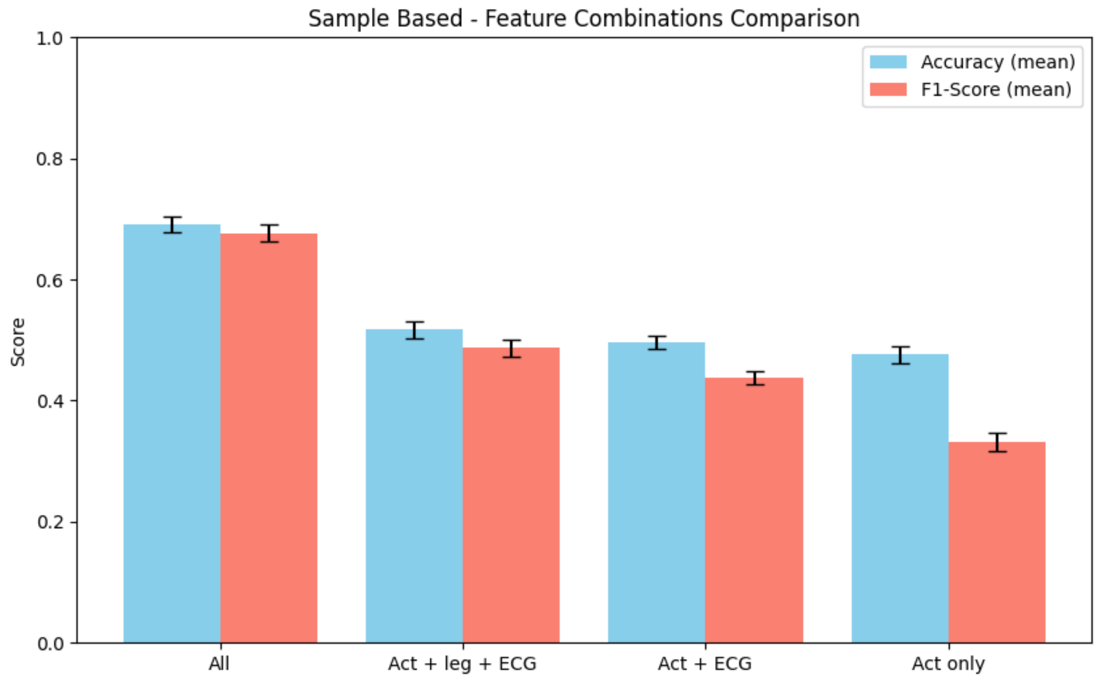

Figure 14: Feature Combinations Comparison, sample-based

The figure 14 shows the cross validation results of evaluating the performance of the sleep stage prediction sample-based model. The All combination has the highest accuracy and F1 score average, which means that when all features are combined, the model can predict sleep stages most accurately. On the other hand, when using Act only combination, the accuracy and F1 score are the lowest. It means it is difficult to distinguish sleep stages with a single feature alone.

5.3.2 Feature comparison, patient-based

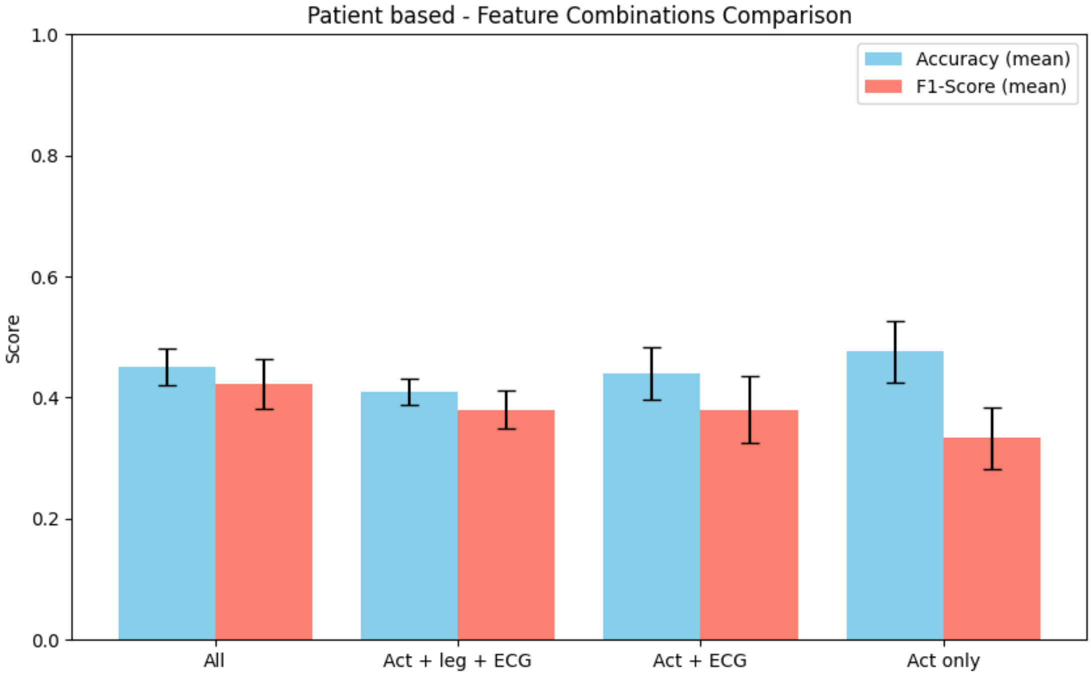

Figure 15: Feature Combination Comparison, patient-based

The figure 15 shows that the performance of each combination tends to be similar to that of the sample unit, but much lower. The All combination still has the highest average accuracy and F1 score, showing that the model makes the best predictions when using all features, even with differences between patients. However, the accuracy and F1 score were particularly lower in the Act only combination, which means that it is difficult to accurately predict sleep stages with only a single Actigraphy data when there are large differences between patients.

Error range

This error range on the graph was produced by calculating the standard deviation of the performance indicators (accuracy and F1 score) at 10-Fold Cross Validation and then add ±1 standard deviation for the mean value. This was added to show the reliability and stability of the model's ability visually. The error range indicates the variability between the folds, which means that the smaller the performance is the more reliable.

5.5 Confusion Matrix Analysis

In the following, we analyze and compare the confusion matrices obtained through the sample-based and patient-based cross validation methods to analyze the performance of the sleep stage prediction model in more detail. Since the two validation methods differ in the generalization performance of the model, it is important to analyze how each method affects the sleep stage prediction. For this analysis, the Random Forest model was evaluated using the Model All.

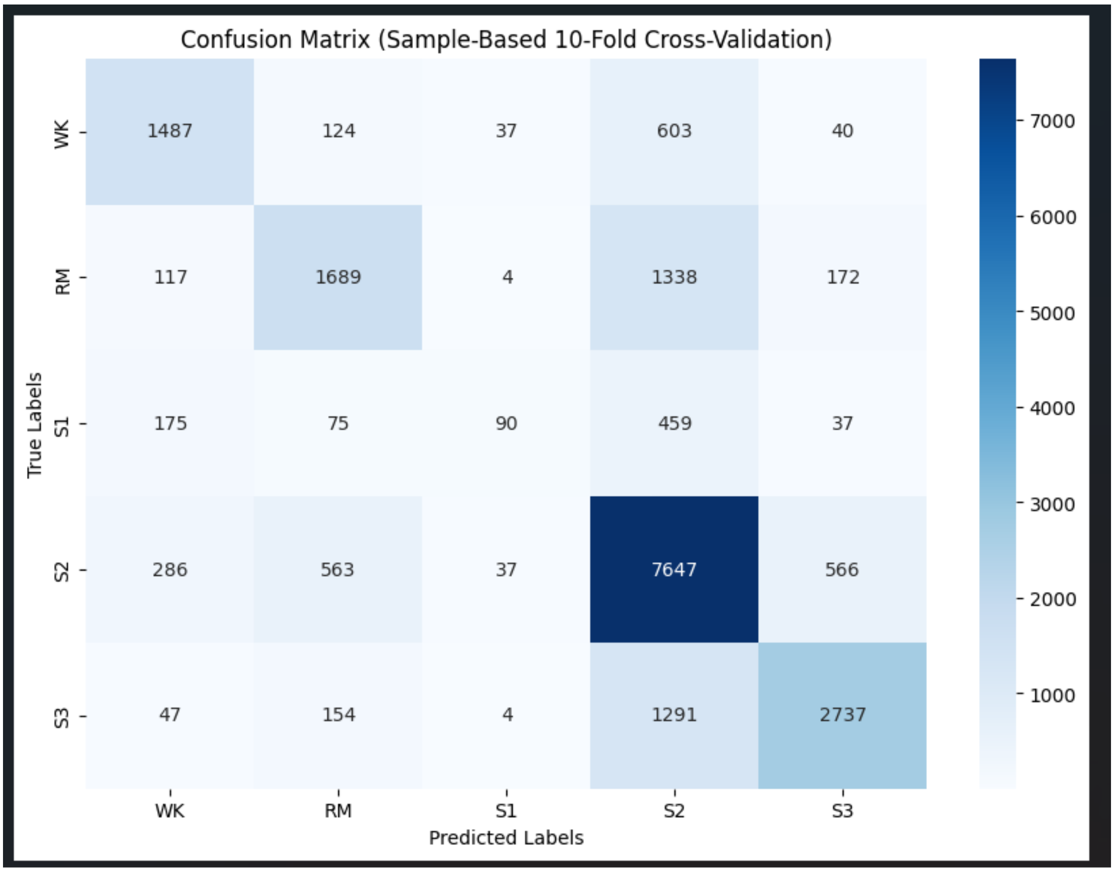

Figure 16: Model All, Confusion Matrix, sample-based

This confusion matrix, figure 16, shows the relationship between the predicted sleep stages (WK, RM, S1, S2, S3) and the actual label using a random forest algorithm. It is sample-based and uses 10-fold Cross Validation to evaluate the quality of the model, and each cell represents the frequency at how much predicted value and actual labels match. A higher diagonal value means that the model accurately predicted that sleep stage. For example, the S2 stage predicts accurately 7647 times, and the WK and RM stages also show a little bit high accuracy. On the other hand, there is a big confusion between the S3 and S2 stages, and the S3 stage is often incorrectly predicted as S2. This suggests that the model has some difficulty in distinguishing deep sleep stages.

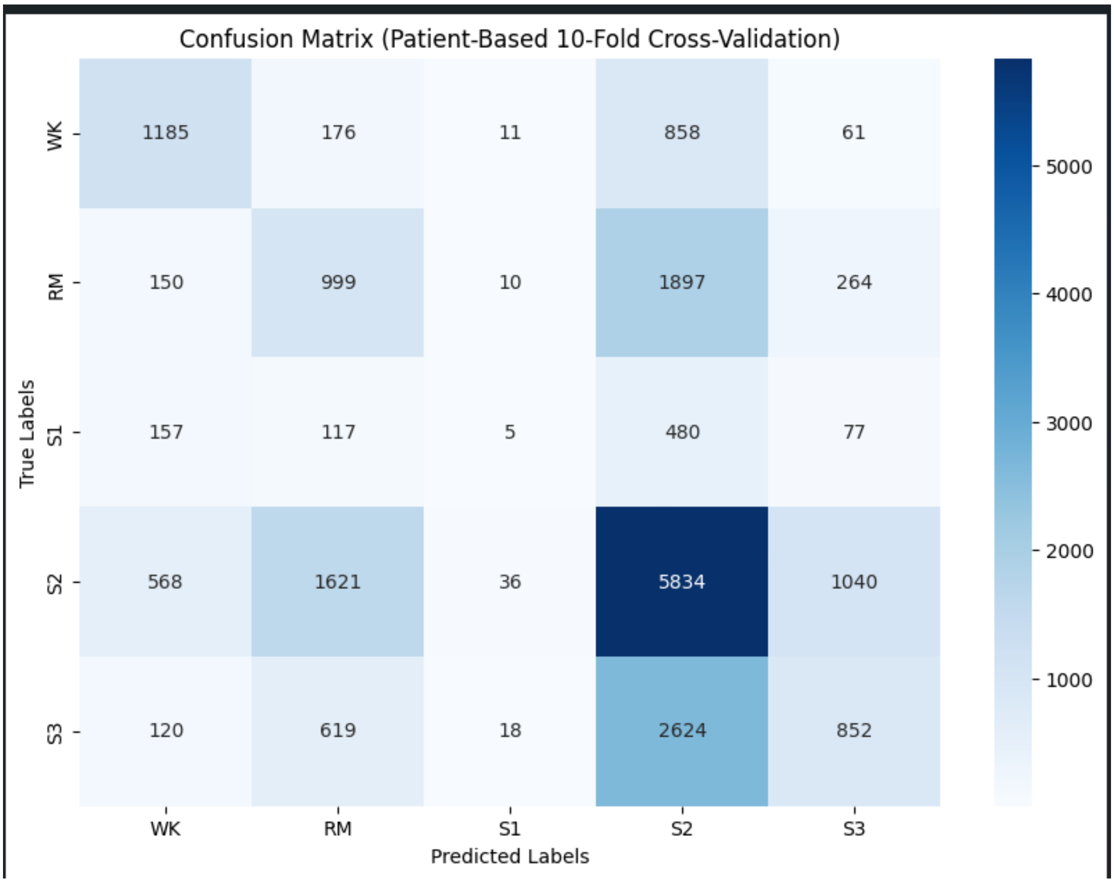

Figure 17: Model all confusion matrix, patient-based

This confusion matrix in figure 17, is the result of predicting sleep stages with a random forest model and was verified through patient-based 10 fold Cross Validation. Higher diagonal values show how often the model correctly predicted each sleep stage. The S2 stage is predicted 5834 times accurately, and it can be shat the WK and RM stages also have a little bit high concordance rate. On the other hand, confusion occurs a lot in the S3 and S2 stages and S3 is incorrectly predicted as S2.

Sample-based cross validation randomly splits data samples into a training set and test set. This means that data from the same patient can be included in both sets, which helps the model learn and evaluate similar data from the same patient. As a result, the model can make a good result on these familiar patterns, but it may struggle to generalize to new patients.

In contrast, patient-based cross validation separates the training and test sets by the patient, so patients in the training set are completely different from those in the test set. This setup is similar to testing the model on a patient it has never seen before, providing a tougher test of the model's ability to generalize. Because of this, the model's performance in patient-based cross validation is often lower than in sample-based cross validation.

6 Discussion

In this study, we compared and analyzed the accuracy and F1 score of sleep stage prediction by applying various feature combinations and machine learning models. In the sample-based validation scenario, the full feature Model All improved accuracy by 46.8% compared to using actigraphy data alone. The model All showed the highest performance, and especially, the combination of ECG data and actigraphy data significantly contributed to sleep stage classification.

We evaluated the performance of machine learning models and compared various feature combinations. By analyzing how each feature combination affects sleep stage prediction, we were able to find the best combination. This analysis helped to determine which data features and combinations were most beneficial for sleep research and to increase the accuracy of sleep stage predictions.

After analyzing the reasons why certain features are advantageous for predicting sleep stages, the ECG data seem to show well the depth of sleep through heart rate. For example, the reason why the "Act + ECG" combination recorded a higher accuracy and F1 score than "Act only" can be understood as that the change in heart rate plays an important role in REM sleep stage or waking state (WK). In addition, confusion between deep sleep stages such as S3 and S2 might arise because the physiological differences between the two stages might be small. Autonomic nervous system responses and body movements in deep sleep stages (S2 and S3) might show similar patterns, making it difficult for the model to distinguish these differences. [26]

Sample-based validation is advantageous because it uses data from the same subject for training and testing, and makes the model evaluate how well it can predict data from subjects that were available for training. On the other hand, patient-based validation uses data from different subjects in each fold to evaluate whether the model can generalize to new subjects. Both validation methods showed a certain level of predictive performance, but the accuracy of both cases was not high. A possible reason for the lower accuracy could be the use of 30 second windows, which might introduce a small desynchronization between the actigraphy data and PSG data.

7 Conclusion

This study aimed to analyze the physiological changes according to various sleep stages and evaluate the performance of the sleep stage prediction model by combining actigraphy and some polysomnography (PSG) data based on this. Actigraphy is a simple way to understand sleep but has limitations in that it is not as accurate as polysomnography. According to this study, it attempted to improve the accuracy of sleep stage classification by combining additional Polysomnography data such as ECG and leg movement signals with actigraphy data.

As a result of the analysis, it was confirmed that the addition of ECG data played an important role in improving the sleep stage classification performance, and it showed mean heart rate and Heart rate variability help show the characteristics of each sleep stage clearly. Comparing the various feature combinations, the model combining all features showed the highest predictive performance, suggesting that the model combining multiple physiological signals is more effective in classifying sleep stages.

However, this study does not include other features from polysomnography for example EEG, respiratory signals, and other, which might reduce the accuracy of the model. Future studies will need polysomnography features or other physiological data to improve the accuracy of the model. These additional studies may contribute to increasing the accuracy and reliability of sleep stage predictions.

List of Figures

List of Equations

List of Table

References

[1] Song, Tzu-An, et al. “Ai-Driven Sleep Staging from Actigraphy and Heart Rate.” *PLOS ONE*, Public Library of Science, https://doi.org/10.1371/journal.pone.0285703. Accessed 11 Nov. 2024.

[2] Almutairi, H., Hassan, G. M., & Datta, A. (2023, December 15). *Machine-learning-based-approaches for sleep stage classification utilising a combination of physiological signals: A systematic review*. MDPI. https://doi.org/10.3390/app132413280

[3] Satapathy, Santosh Kumar, et al. “Machine Learning-Empowered Sleep Staging Classification Using Multi-Modality Signals - BMC Medical Informatics and Decision Making.” *BioMed Central*, BioMed Central, https://doi.org/10.1186/s12911-024-02522-2

[4] “Advancing Heart, Lung, Blood, and Sleep Research & Innovation.” *National Heart Lung and Blood Institute*, U.S. Department of Health and Human Services, www.nhlbi.nih.gov/. Accessed 12 Dec. 2023.

[5] Altini, Marco, and Hannu Kinnunen. “The promise of sleep: A multi-sensor approach for accurate sleep stage detection using the Oura Ring.” *Sensors*, vol. 21, no. 13, 2021, p. 4302, https://doi.org/10.3390/s21134302..

[6] “WGT3X-BT: ActiGraph Wearable Devices.” *wGT3X-BT | ActiGraph Wearable Devices*, https://theactigraph.com/actigraph-wgt3x-bt. Accessed 12 Dec. 2023.

[7] Vähä-Ypyä, Henri, et al. “Validation of cut-points for evaluating the intensity of physical activity with accelerometry-based mean amplitude deviation (MAD).” *PLOS ONE*, vol. 10, no. 8, 2015, https://doi.org/10.1371/journal.pone.0134813. Accessed 3 Jan. 2024.

[8] Van Hees, Vincent T., et al. “A novel, Open access method to assess sleep duration using a wrist-worn accelerometer.” *PLOS ONE*, vol. 10, no. 11, 2015, https://doi.org/10.1371/journal.pone.0142533. Accessed 3 Jan. 2024.

[9] Tilmanne, Joëlle, et al. “Algorithms for sleep–wake identification using actigraphy: A comparative study and new results.” *Journal of Sleep Research*, vol. 18, no. 1, 2009, pp. 85–98, https://doi.org/10.1111/j.1365-2869.2008.00706.x. Accessed 9 Jan. 2024.

[10] *Polysomnography (PSG)* (2023) *PsychDB*. Available at: https://www.psychdb.com/neurology/polysomnography, Accessed 11 November 2024.

[11] Lowrie, J. and Brownlow, H. (2020) *The impact of sleep deprivation and alcohol on driving: A comparative study - BMC public health, BioMed Central*. https://doi.org/10.1186/s12889-020-09095-5. Accessed 11 November 2024.

[12] Fois, Chiara, et al. "The Utility of Polysomnography for the Diagnosis of NREM Parasomnias: An Observational Study over 4 Years of Clinical Practice." *Journal of Neurology*, U.S. National Library of Medicine, Feb. 2015, www.ncbi.nlm.nih.gov/pmc/articles/PMC4330461/.

[13] Sieciński, S., Kostka, P. S., & Tkacz, E. J. (2020, August 13). *Heart rate variability analysis on electrocardiograms, seismocardiograms and gyrocardiograms on Healthy Volunteers*. Sensors (Basel, Switzerland). https://pmc.ncbi.nlm.nih.gov/articles/PMC7472094/

[14] Mistry, D. (2022, November 21). *What actually happens when we sleep?* Sleep Psychiatrist. Retrieved December 18, 2024, from https://sleeppsychiatrist.com/blog/what-actually-happens-when-we-sleep/

[15] Pedersen, M., Farnsworth, B., Fischer, E., & Mowla, L. (2024, April 17). *Heart rate variability - how to analyze ECG data*. iMotions. https://imotions.com/blog/learning/best-practice/heart-rate-variability/

[16] Ibm. (2024, October 25). *What is Random Forest?*. IBM. https://www.ibm.com/topics/random-forest, Accessed 25 October 2024

[17] Brownlee, J. (2023, October 3). *A gentle introduction to k-fold cross-validation*. MachineLearningMastery.com. https://machinelearningmastery.com/k-fold-cross-validation/ Accessed 3 October 2024

[18] Ph.D., J. M., & Kavlakoglu, E. (2024, August 12). *What is a confusion matrix?*. IBM. https://www.ibm.com/topics/confusion-matrix#:~:text=A%20confusion%20matrix%20 Accessed 10 November

[20] Nahlik, M. (2019), Sleep-Wake and Sleep Stage Detection from Wrist-Worn Actigraphy, diploma thesis, CTU Prague

[22] Boe, A.J., McGee Koch, L.L., O'Brien, M.K. et al. Automating sleep stage classification using wireless, wearable sensors. npj Digit. Med. 2, 131 (2019). https://doi.org/10.1038/s41746-019-0210-1

[23] A. Sadeh, J. Alster, D. Urbach, and P. Lavie, “Actigraphically based automatic bedtime sleep-wake scoring: validity and clinical applications,” *Journal of Ambulatory Monitoring*, vol. 2, no. 3, pp. 209–216, 1989.

[24] A. Sadeh and C. Acebo, “The role of actigraphy in sleep medicine,” *Sleep Medicine Reviews*, vol. 6, no. 2, pp. 113–124, 2002, doi: 10.1053/smrv.2001.0182. https://doi.org/10.1053/smrv.2001.0182

[25] “Supervised vs. Unsupervised Learning: What’s the Difference?” *IBM*, 12 Mar. 2021, www.ibm.com/think/topics/supervised-vs-unsupervised-learning. Accessed 11 Dec. 2024.

[26] Kandel, E.R., Schwartz, J.H., Jessell, T.M., Siegelbaum, S.A., Hudspeth, A.J. (Eds.). (2021). *Principles of Neural Science: The Interaction Between Sleep and Autonomic Nervous System Responses*. In: Neural Control of Sleep Stages. Springer, Cham. https://doi.org/10.1007/978-3-030-62263-3_7

[27] Sattler, Svenja. "The Role of the Immune System Beyond the Fight Against Infection." *Frontiers in Immunology*, vol. 8, 2017, p. 1359. *Frontiers*, https://doi.org/10.3389/fimmu.2017.01359.