Master's thesis

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

Margrethe Gryvill

Sleep classification of body-worn accelerometer data using machine learning

Master's thesis in Computer Science
Supervisor: Kerstin Bach
Co-supervisor: Aleksej Logacjov
December 2021

NTNU
Norwegian University of
Science and Technology

Margrethe Gryvill

Sleep classification of body-worn accelerometer data using machine learning

Master's thesis in Computer Science
Supervisor: Kerstin Bach
Co-supervisor: Aleksej Logacjov
December 2021

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

NTNU
Norwegian University of
Science and Technology

Abstract

Sleep is essential for maintaining good physical and mental health. One important aspect of sleep is the pattern of sleep stages, which can tell a lot about a person's sleep quality. In the field of sleep stage classification the goal is to differentiate between wakefulness and the various stages of sleep. This is preferably done with Polysomnography (PSG), which is a procedure where physiological signals, like brain activity, heart rate and muscle activity, are measured while the subject is sleeping. The procedure is usually performed in a hospital, and the data must be classified manually by a specialist. Because of its high cost and the discomfort experienced by many patients during the procedure, PSG is not optimal for all studies of sleep. Actigraphy has shown to be a successful alternative to PSG, especially when the goal is to separate wakefulness from sleep, without considering sleep stages. This is a method where body-worn sensors collect accelerometer data, and possibly other measurements, in a natural setting. The actigraphy data is suitable for classification with machine learning.

In this project, our main goal is to enable sleep analysis of the actigraphy data from the HUNT4 study. This is a population study from Norway, where 35,000 subjects participated in collection of actigraphy data from one sensor on the back and one on the thigh. Analysis of this data can ideally expand on the knowledge of sleep and health. Two separate datasets, with the same sensor placements as in HUNT4, are used in this project. The actigraphy is collected simultaneously as PSG, which gives sleep class labels to the actigraphy data. This is used by supervised machine learning methods in our experiments. There are several ways to categorize sleep. In this project we classify the sleep data as light sleep, deep sleep and rapid eye movement (REM) sleep.

Both sleep–wake classification and classification of sleep stages are tested in our experiments with four different machine learning algorithms: Random Forest, XGBoost, K-NN and SVM. As the equipment used in the HUNT4 data and our test datasets also measure skin temperature, we experiment with inclusion of temperature data. In addition, our test datasets include a sensor placed on the wrist, and the results of incorporating the wrist data is investigated as well. For sleep–wake classification the best results are achieved by XGBoost using accelerometer and temperature data from back, thigh and wrist sensors combined. This results in accuracy, F1-score, area under ROC-curve, sensitivity and specificity of 0.91, 0.94, 0.94, 0.97 and 0.72, respectively.

Sammendrag

Søvn er essensielt for å opprettholde god fysisk og psykisk helse. Et viktig aspekt ved søvn er mønsteret av søvnstadier, som kan si mye om en persons søvnkvalitet. Innenfor søvnstadie-klassifisering er målet å kunne skille mellom våkenhet og ulike stadier av søvn. Dette gjøres fortrinnsvis med Polysomnografi (PSG), som er en prosedyre hvor fysiologiske signaler, som hjerneaktivitet, hjerterytme og muskelaktivitet, måles mens pasienten sover. Prosedyren gjennomføres vanligvis på et sykehus, og dataene må klassifiseres manuelt av en spesialist. På grunn av den høye kostnaden og ubehaget som mange pasienter opplever ved metoden, er ikke PSG optimal for alle typer søvnstudier. Aktigrafi har vist seg å være et godt alternativ til PSG, spesielt når målet kun er å skille mellom våkenhet og søvn, uten å skille mellom søvnstadier. Dette er en metode hvor sensorer festet til kroppen samler akselerometerdata, og eventuelt andre målinger, i en naturlig setting. Aktigrafidata egner seg for klassifisering med maskinlæring.

I dette prosjektet er hovedmålet vårt å muliggjøre søvnanalyse av aktigrafidata fra HUNT4-studien. Dette er en befolkningsundersøkelse fra Norge, hvor 35.000 personer deltok i innsamling av aktigrafidata fra én sensor på ryggen og én på låret. Analyse av dataene kan forhåpentligvis gi ny kunnskap om søvn og helse. To separate datasett, med samme sensorplassering som i HUNT4, blir brukt i dette prosjektet. Aktigrafi-målingene har blitt utført samtidig som PSG, og dette gir oss merkelapper for søvnklassene i aktigrafidataene. Dette blir brukt av veiledede maskinlæringsmetoder i eksperimentene våre. Det finnes mange ulike måter å kategorisere søvn. I dette prosjektet klassifiserer vi søvndataene som lett søvn, dyp søvn og REM(Rapid eye movement - rask øyebevegelse)-søvn.

Både klassifisering av søvn og våkenhet, og søvnstadieklassifisering testes i våre eksperimenter med fire ulike maskinlæringsalgoritmer: Random Forest, XGBoost, K-NN og SVM. Ettersom utstyret som brukes i HUNT4-dataene og våre testdatasett også måler hudtemperatur, eksperimenterer vi med å inkludere temperaturdata. I tillegg inneholder testdatasettene en sensor plassert på håndleddet, og resultatet av å inkludere håndleddsdataene blir også undersøkt. For klassifisering av søvn og våkenhet oppnås de beste resultatene av XGBoost med både akselererometer- og temperaturdata fra rygg-, lår- og håndleddssensor kombinert. Dette gir nøyaktighet, F1-skår, areal under ROC-kurven, sensitivitet og spesifisitet på 0.91, 0.94, 0.94, 0.97 and 0.72, respektivt.

Preface

This master thesis was conducted at the Department of Computer Science at the Norwegian University of Science and Technology (NTNU). The project was executed in the fall of 2021, supervised by associate professor Kerstin Bach and Aleksej Logacjov.

I would like to thank Kerstin for her valuable guidance and mentoring through the project and thesis writing. In addition I would like to thank Aleksej for his helpful feedback and support.

Margrethe Gryvill
Trondheim, December 17, 2021

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Sleep is a crucial part of our everyday life. It affects our general well-being, as well as our physical and mental health. Sleep disorders have a negative impact on quality of life, and this is confirmed by the research of Reimer and Flemons [2003]. Specifically sleep apnea, insomnia, narcolepsy and restless leg syndrome are considered in their research, and even though the research papers they summarize use different measurements of life quality, the research consistently shows poorer life quality in people with sleep disorders, compared to the general population. Sleep loss is also shown to affect the community, not only through life quality, but also economically [Hillman and Lack, 2013]. These issues should be of major concern due to the extent of sleep problems. According to the Norwegian Institute of Public Health, sleep problems are experienced weekly by almost a third of the population[1]. To achieve more insight in the connection between sleep and health, more research in the field is needed. This can possibly lead to better treatment of sleep related problems, more knowledge about the effects of sleep, and help the general public in obtaining healthier sleep habits.

The gold standard of sleep analysis, Polysomnography (PSG), includes measurements of several body functions, including brain activity and heart rhythm. The sleep stages are then manually classified. These factors causes the method to have high accuracy, but also makes it costly. The fact that the procedure is usually conducted in a sleep laboratory or hospital can have a negative impact on the sleep quality of the subject, in addition to the discomfort of wearing the equipment. This is a weakness because the goal is usually to analyze the normal sleep patterns of the subject.

Another way to analyze sleep is actigraphy. An actigraph is a body-worn sensor, consisting of a three-dimensional accelerometer and possibly other sensors.

[1] `https://www.fhi.no/en/op/hin/mental-health/sleep-problems/`; accessed 2021-11-22

The body movements data can be analyzed in a variety of ways. Because of the inexpensive equipment and low intrusiveness of the method, it is preferable in some situations, for instance when the subjects are children, when data collection for several days is necessary, or when a large group of people is participating in a study.

The Trøndelag Health Study, HUNT[2], is the largest collection of health data in Norway. The data has been collected through four studies, from 1984 to 2019 and now is a dataset including 230,000 participants. The data is mainly collected through questionnaires and biological samples. In the last study, HUNT4, the subjects were also invited to participate in activity monitoring with actigraphs. The participants wore the sensors for a week, day and night. By analyzing this data, we can obtain valuable information about the sleep patterns of the population, and possibly connections between sleep and health in general.

In our previous work the goal was to compare machine learning methods for sleep–wake classification [Gryvill, 2020]. In this thesis the main focus will be on comparing different sets of input data, both for sleep in general and sleep stages.

1.1 Goals and Research Questions

This work has three main goals.

Goal 1 Understand the state-of-the-art of sleep detection with data from body-worn accelerometers.

This goal will be achieved by answering the first research question.

Research question 1.1 What are the major contributions within sleep detection with accelerometer data using machine learning methods during the last three years?

Our work builds on the work of Hay [2019], going through the related work up to 2018. For this reason we will focus on new research, restricting our literature search to papers written in 2018 and later.

By the results of the first goal, the two other goals will be investigated.

Goal 2 Improve sleep–wake classification using machine learning on actigraphy data from the back, thigh and wrist.

Two labeled datasets consisting of accelerometer and temperature measurements from three body-worn sensors, placed on the back, thigh and wrist, are available. By training and testing machine learning models on these data, we will answer the following research questions.

[2] `https://www.ntnu.edu/hunt/about-hunt`; accessed 2020-10-14

Research question 2.1 How does including *temperature data* in the dataset impact the classification results?

Research question 2.2 How does including *wrist sensor data* in the dataset impact the classification results?

Goal 3 Improve sleep stage classification using machine learning on actigraphy data from the back, thigh and wrist.

The sleep stages normally follow a cyclic pattern, with a distinct distribution of each stage. By analyzing the sleep stages of a subject, unusual patterns and rhythms can be discovered, and reveal sleep disorders or other health issues.

For this type of classification we want to investigate the following research questions.

Research question 3.1 How does including *temperature data* in the dataset impact the classification results for light, deep and REM sleep, respectively?

Research question 3.2 How does including *wrist sensor data* in the dataset impact the classification results for light, deep and REM sleep, respectively?

1.2 Research Methods

To answer research question 1.1, a structured literature review is performed. Research questions 2.1, 2.2, 3.1 and 3.2 are answered by a pipeline written in Python and experiments using machine learning methods. The experiments include a comparison of machine learning methods with 1) only back and thigh accelerometer data, 2) back and thigh accelerometer and temperature data, and 3) back, thigh and wrist data, both accelerometer and temperature. These experiments are conducted for sleep–wake classification, light sleep – non-light sleep, deep sleep – non-deep sleep, and REM – non-REM sleep classification.

1.3 Contributions

Through this work our main contributions are the implementation of binary classification methods for general sleep and wake and the sleep stages light, deep and REM sleep. The implemented pipeline includes feature generation from raw data, and hyperparameter tuning with grid search and leave-one-group-out cross-validation. Our comparisons of sleep classification with combinations of back, thigh and wrist data, both accelerometer data and temperature, are not previously done, to our knowledge. This has resulted in knowledge about how wrist actigraphy data affects the results of sleep classification with machine learning, compared with classification from actigraphs placed at the back and thigh. Additionally, our models are trained on a dataset where all subjects have a sleep disorder, and then tested on a dataset of healthy sleepers. The similar results for the two datasets can be important in future work, as a majority of labeled sleep data is collected from people with sleep disorders.

1.4 Thesis Structure

This thesis consists of seven chapters, where chapter 1 is this introduction. Chapter 2 introduces necessary background theory for our work. This includes a presentation of the HUNT4 study in section 2.1, descriptions of the sleep recording methods PSG and actigraphy in section 2.2 and 2.3, respectively, followed by explanation of the relevant machine learning algorithms in 2.4. In chapter 3 the related literature is described in section 3.2 after a description of the search process for the literature in 3.1, and chapter 4 contains a description of the datasets we have used and an explanation of how the development of the models was conducted. In this section we also describe the segmentation and feature generation in section 4.3, and the tuning of hyperparameters for the machine learning methods in section 4.5. In chapter 5 the experiments are described in section 5.1 and the setup of the experiments in section 5.2, followed by their results in section 5.3. In chapter 6 the results are discussed in relation to the research questions. Finally, in chapter 7 our conclusion is presented in section 7.1 and thoughts on possible future research in section 7.2.

Chapter 2

Background

This chapter presents the background theory that our work is based on. This includes the HUNT4 study, and the sleep analysis methods PSG and actigraphy. In addition the relevant machine learning techniques and evaluation measurements are described.

2.1 HUNT4

The Trøndelag health study, HUNT[1], has collected data from the inhabitants of Trøndelag County through questionnaires and biological samples since 1984. The fourth iteration of the study, HUNT4, was completed in 2019, and approximately 35,000 of the participants also participated in activity monitoring. Two sensors of the type Axivity AX3 data logger[2] were worn by the subjects for one continuous week. This is an accelerometer that logs acceleration in three dimensions. One of the sensors was placed on the lower back and the other on the thigh, approximately 10 cm above the knee.

To categorize the acceleration signals into activities, a framework for human activity recognition is used. A sample of 3 hours of acceleration data can be seen in figure 2.1, with the first graph corresponding to the back sensor and the second to the thigh sensor. The acceleration is measured in the range $\pm 8g$. Horizontal lines are seen in periods with little movement. The black dots in the plot, marking the predicted activity, classifies lying down, but does not distinguish between sleeping and lying while awake. Sleep classification could be incorporated in this model if a sufficiently accurate method is implemented.

[1] `https://www.ntnu.edu/hunt/about-hunt`; accessed 2020-10-14

[2] `https://axivity.com/product/ax3`; accessed 2020-11-23

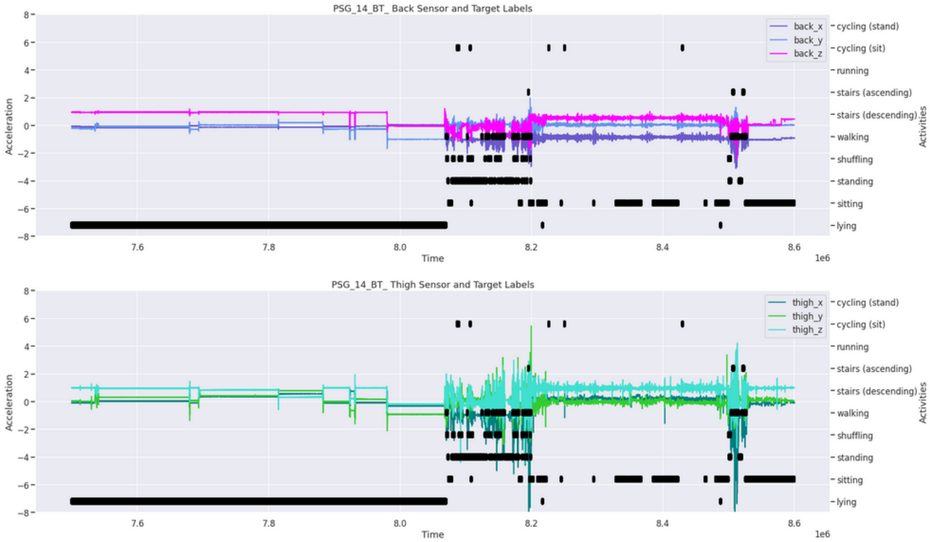


Figure 2.1: Example plot of acceleration data from the back (top) and thigh (bottom) for approximately 3 hours. The x-axis represents the time by samples, and the y-axis shows acceleration in g. The black dots are the predicted activity for each time step by our current model.

2.2 Polysomnography

The gold standard of sleep analysis is called polysomnography, or PSG. It is usually conducted in a sleep clinic or hospital, where measurements of body functions are recorded while the subject is sleeping. From these measurements a sleep technician can manually classify the sleep stage of the subject, and evaluate the sleep with high precision. Several sleep related disorders, for instance sleep apnea and insomnia, are diagnosed with PSG. Some of the most important measurements of PSG are [Ibáñez et al., 2018]:

- Electroencephalogram (EEG) - brainwave activity
- Electrooculogram (EOG) - eye movement
- Electromyogram (EMG) - muscle activity (face and body)
- Electrocardiogram (EKG) - heart rate and rhythm

These recordings are needed to differentiate between the various sleep stages. For example, REM sleep is recognised by its characteristic eye movements and brain activity, which are similar to that of wakefulness. The movements of the body and face are naturally important to separate REM sleep from wakefulness. As PSG requires costly equipment and expertise, it is not usually used for several consecutive nights or in large scale studies. Additionally, the setting of being in a sleep clinic, and not at home in your own bed, and being physically restricted by the equipment, can affect the quality of the sleep, and the measurements are therefore possibly not a good representation of a normal night of sleep. Another consequence of the high cost of PSG is that it is usually restricted to people with suspected sleep disorders. For this reason, only a small amount of all recorded PSG data is collected from healthy sleep.

2.3 Actigraphy

Actigraphy is a method for monitoring movement with a body-worn actigraph unit. The main module of an actigraph is the accelerometer, but additionally some actigraphy units contain other sensors, for instance temperature sensors. Actigraphy is used for sleep analysis, and because the units are inexpensive and can be worn for many consecutive days without interference, it is suitable for long term usage and large scale studies. The low intrusiveness of actigraphy prevents it of having a large impact on the sleep quality, and in addition actigraphy has the benefit of being usable in the subject's natural setting. The limitations of actigraphy compared to PSG are the results of the missing brain activity, heart rate monitoring and other measurements of body functions, which makes the method less accurate.

The acceleration data from the actigraph can be analyzed in a variety of ways. Some of the renowned methods are algorithms based on activity counts, [Cole et al., 1992] [Sadeh et al., 1994], while many new methods rely on machine learning algorithms. Several papers presenting machine learning methods for sleep classification will be presented in chapter 3.

2.4 Classification methods

All of the classification methods used in our experiments are supervised machine learning algorithms. Machine learning is a type of artificial intelligence, where the actor improves its performance based on experience. In supervised learning the model is given training examples consisting of both input and output data. By learning the connection between the input and the correct output, new input data can be classified. The input is usually represented by several numerical features for each data point, and the output is in our case represented by a single value, for example sleep(1) or wake(0). Each classification algorithm uses a number of hyperparameters, which have to be set manually to optimize the model, as they are not learned during training. The procedure for deciding the hyperparameters is described in section 4.5.

The following sections describe the classifications methods used in this project: Random Forest, XGBoost, K-NN and SVM. Additionally, decision trees are explained, as they are used in Random Forest and XGBoost. The relevant hyperparameters for the classifiers used in our experiments are summarized in table 2.1, and explained in the section of the respective method.

Table 2.1: Relevant hyperparameters for each of the classifiers.
RF = Random Forest, XGB = XGBoost.

Method	Hyperparameter	Description
RF	n_estimators	Number of trees
	max_depth	Maximum depth of each tree
	min_samples_leaf	Minimum number of samples in a leaf node
	min_samples_split	Minimum number of samples in internal node to split it
XGB	n_estimators	Number of trees
	colsample_bytree	Ratio of columns used in each tree
	gamma	Minimum loss reduction to split a leaf node in the tree
	learning_rate	Boosting learning rate
	max_depth	Maximum depth of each tree
	min_child_weight	Minimum sum of instance weight in a child node
K-NN	n_neighbors	Number of neighbors to consider
	p	Power parameter for distance metric
	weights	Uniform or distance based weighting of neighbors
SVM	C	Regularization parameter
	gamma	Kernel coefficient (for some kernels)
	kernel	Kernel type - how to split the classes

2.4.1 Decision trees

Decision trees are tree-shaped graph models with a root node, leaf nodes and internal nodes [Quinlan, 1986]. For each node, except the leaf nodes, some feature of the current data point is tested. The edges out of the node are the possible outcomes, connected to the next layer of nodes. The leaf nodes represent the labels or output values of the data. The tree is built from the training data, such that starting from the root node, going through the correct nodes based on the feature values of the data point, the training data ends in a leaf node with the correct label.

Some of the important parameters of a decision tree are the maximum depth of the tree, and the minimum number of samples in a leaf node. The maximum depth is a boundary of the number of layers in the tree. More layers increases the chance of overfitting, meaning that the model has learned too many details of the training data to generalize well when given new data. The minimum number of samples in a leaf node limits the splitting of leaf nodes with few samples. Similarly to the maximum depth, the probability of overfitting to the training data will increase if a small amount of samples is allowed in the leaf nodes.

2.4.2 Random Forest

Random Forest (RF) is an ensemble model, which means it is based on a simpler machine learning model, specifically the decision tree [Ho, 1995]. It uses the ensemble method bootstrap aggregating, or bagging, where each decision tree is made from a subset of the samples in the training data, and the same data sample can be used several times in one decision tree. When the Random Forest model is classifying a new unseen data sample, each of the decision trees will give a prediction, and the class with the most votes is chosen.

As the Random Forest model consists of decision trees, most of the hyperparameters of the decision tree applies to Random Forest as well. *Max_depth* and *min_samples_leaf* are important for the same reasons as for the decision tree. *Min_samples_split* is another parameter closely related to the minimum samples of a leaf, but sets a lower limit for the number of samples in an internal node, and not the resulting leaves. This is also important to avoid overfitting. The number of estimators - decision trees - should also be set. Generally more trees are better, as this will reduce the chance of overfitting, but the model will eventually reach a limit where the improvements are insignificant.

2.4.3 XGBoost

XGBoost (XGB) is also an ensemble algorithm, but unlike Random Forest it uses boosting [Chen and Guestrin, 2016]. This means that the weak learners, for instance the decision trees, are added one at a time to the complete model. The incorrectly classified data of one tree is prioritized in the next step, such that the total loss of the model is minimized. The loss of a model can be calculated in a variety of ways, penalizing difference in the true values and the predicted values. The models are weighted so that the final result is combining the strength of each classifier.

For the parameters of XGBoost, a higher number of estimators usually leads to an increase in performance. *Colsample_bytree*, the ratio of columns used when generating each tree, affects the randomness of the model, which influences the model's handling of noise in the data. *Gamma*, *max_depth* and *min_child_weight* are all affecting the potential of overfitting by restricting the method's ability to fit perfectly to the training data. The *learning_rate* decides how much impact each new training sample is given, so this is also important for generalizing well to unseen data.

2.4.4 K-nearest neighbors

In K-nearest neighbors (K-NN) the training samples are simply saved in a space with n dimensions, where n is the number of features. To classify new instances, they are compared with the k closest data points already known [Dudani, 1976]. The predicted class of the new observation is the most common class of the neighbors.

The value k is a hyperparameter set by the user. The ideal value of k is related to the nature of the data, and it is found through experimentation. The parameter p is used for computing the distance between the samples. The default distance metric in the K-NN implementation of Scikit-learn[3] is the Minkowski metric, calculated as follows:

$$M = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} \tag{2.1}$$

The value of p is usually set to 1, giving the Manhattan distance, or 2, resulting in the Euclidean distance. The parameter *weights* can be set to *uniform*, giving all neighbors equal value in the decision of class, or *distance*, giving closer samples a larger weight. Again, the optimal choice depends on the problem and the data.

2.4.5 Support vector machine

Support vector machine (SVM) is a method that separates the samples of two categories with as much distance as possible, as this has shown to minimize the likelihood of wrongly classifying new samples [Cortes and Vapnik, 1995]. The position of each sample is defined by the features, with one dimension for each feature. The border between the classes is placed to ensure that most of the samples are on the correct side of the border. To handle cases where the classes are not separable with a line, a kernel function is used. This will transform the data by adding dimensions to ensure that the classes can be separated.

The *kernel* hyperparameter sets the type of kernel used by the SVM, and some of the most used kernels are *linear*, *poly* (polynomial) and *rbf* (radial basis function). Some of the kernels use a parameter, *gamma*, to decide how far the

[3] `https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html`; accessed 2021-12-08

influence of a sample reaches. This parameter can be set to *auto* or *scale*, where the latter depends on the variation of the input data. Even with a good choice of kernel, the model will often not classify all samples correctly. The parameter C controls the regularization of the algorithm. With a high value, C will penalize each wrongly classified sample more than with a low value. C should be low enough to generalize well, but high enough to learn from the training data.

2.5 Evaluation of methods

To evaluate the performance of the classification methods, several metrics can be calculated. A single metric can not describe all aspects of the performance, but by using a collection of evaluation methods, we will have a clearer picture of the results. One evaluation method is the confusion matrix, describing the amount of positive samples correctly classified as positive (true positive), negative samples correctly classified as negative (true negative), and the positive and negative samples incorrectly classified (false negative and false positive, respectively). Figure 2.2 shows a confusion matrix for sleep–wake classification, where sleep is the positive class. These four categories of classified samples are used when deriving the following methods of evaluation.

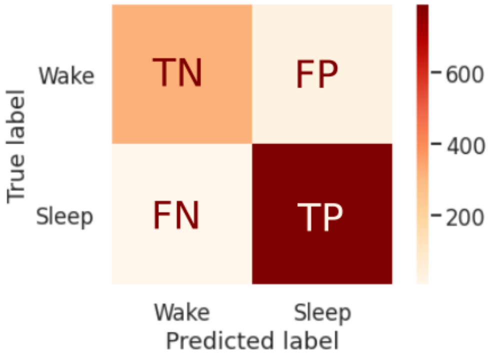


Figure 2.2: Confusion matrix. TN: true negative, FP: false positive, FN: false negative, TP: true positive.

The *sensitivity* of a classification method, also known as *recall*, is the proportion of positive samples correctly classified as positive:

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

The *specificity* of a method is the proportion of negative samples correctly classified as negative:

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

The *precision* is the proportion of the samples classified as positive that are actually positive:

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

The *accuracy* is the total proportion of correctly classified samples:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{2.5}$$

The *F1-score* is a measure that weights the sensitivity (recall) and precision of the model. Because of this, a model that classifies all the samples as the negative class will not have a high F1-score, even though the accuracy may be high.

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{2.6}$$

The Receiver Operating Characteristic curve (*ROC* curve) is a graph where the sensitivity, the true positive rate, is plotted on the y-axis and (1 − specificity), the false positive rate, is plotted on the x-axis. The area under this curve, known as *AUC*, tells the probability of the model predicting the likelihood of a random positive sample to be positive higher than the likelihood of a random negative sample to be positive.

Chapter 3

Related work

This chapter presents the strategy of the structured literature review, in addition to a short summary of the relevant papers found through the literature search. Finally, the findings are compared and related to our work.

3.1 Structured literature review protocol

A structured literature review is conducted to give an overview of the current state-of-the-art in sleep classification using machine learning on accelerometer data. We use three sources: Google scholar[1], IEEE Xplore[2] and Springer Link[3]. Table 3.1 contains the search terms used. The terms in each group have similar semantic meaning, so to find papers including at least one of the terms in each group, the complete search term was:

('Sleep study' OR 'Sleep detection' OR 'Sleep classification' OR 'Sleep analysis')
AND ('Actigraphy' OR 'Accelerometer' OR 'Actigraph')
AND 'Machine Learning'

The Boolean operators 'and' and 'or' are supported by all the search engines. The relevant findings from our previous work, [Gryvill, 2020], are included in this review, as well as new findings using the same protocol. In addition a few other papers were found in the process of writing this thesis.

[1] `https://scholar.google.com/`

[2] `https://ieeexplore.ieee.org/`

[3] `https://link.springer.com/`

Table 3.1: Search terms

	Group 1	Group 2	Group 3
Term 1	Sleep study	Actigraphy	Machine Learning
Term 2	Sleep detection	Accelerometer	
Term 3	Sleep classification	Actigraph	
Term 4	Sleep analysis		

3.2 Selected literature

In the following sections the contents of the papers found in the literature review are presented. The first section describes the papers where sleep–wake classification is the main focus, while the second section contains the papers with other related classification tasks. The summary in the end of the chapter compares the papers and discuss their relevance to our work.

3.2.1 Sleep–wake classification

The papers in this section all contain research where the main experiments were classification of sleep and wake.

Sano et al. [2019] compared the performance of long short-term memory (LSTM) with three other machine learning methods on sleep–wake classification. The other methods were a simple neural network, a logistic regression model, and SVM. Different combinations of sensor data from a wrist sensor and the subject's phone were also compared. The best results were achieved with accelerometer and skin temperature data, both from the wrist sensor. The LSTM model resulted in higher accuracy for all combinations of data.

Khademi et al. [2019] compared personalized versions of machine learning models with generalized models. This was done with naive Bayes, regularized logistic regression, Random Forest, AdaBoost, and XGBoost. The methods were evaluated by comparing total sleep time (TST), wake after sleep onset (WASO), sleep onset latency (SOL), sleep efficiency (SE), and number of awakenings (NA). For these parameters the personalized models outperformed the general versions for most algorithms. Random Forest and XGBoost outperformed the other methods with an accuracy of 0.85 and 0.86, respectively.

Li et al. [2020] developed an unsupervised method for sleep–wake classification using a hidden Markov model (HMM). The method is individualized, only analyzing the data of the current subject to classify the subject's data. To evaluate the method, it was compared with the Actiwatch software algorithm and a supervised method trained on another dataset. The HMM method outperformed the other methods, with a mean accuracy of 0.587.

Cho et al. [2019] presented a deep neural network, a combination of a convolutional neural network (CNN) and LSTM, trained on raw data. The method was compared with four other machine learning methods trained on features, and two traditional sleep detection algorithms. The experiments showed that the deep neural network had better accuracy (0.8877 ± 0.0397), recall (0.9296 ± 0.0503) and precision (0.9039 ± 0.0238) than the other methods for sleep–wake classification. Sleep diaries were used as the ground truth.

Li and Nakamura [2019] compared sleep–wake classification of an accelerometer on the trunk and a wrist accelerometer. The data from the trunk was classified with SVM using a Gaussian kernel. For the wrist data the traditional Cole-Kripke algorithm was used. The SVM method gained high accuracy (0.932 ± 0.043), sensitivity (0.917 ± 0.083), and specificity (0.941 ± 0.049) with the Cole-Kripke results as ground truth.

Palotti et al. [2019] compared several methods for sleep–wake classification with accelerometer data. In addition to the machine learning methods extra trees, logistic regression, linear SVM, perceptron, LSTM, and CNN, several traditional methods were also tested. Especially the deep learning methods LSTM and CNN performed well, with accuracy of 0.829 ± 0.010 and 0.831 ± 0.010, respectively.

Yildiz et al. [2019] confirmed that activity data and light data from a wrist-worn activity monitor analyzed with machine learning can be used for sleep–wake classification in elderly and cognitively impaired. The machine learning method used in this work was an LSTM neural network. It achieved a specificity of 0.377 and sensitivity of 0.602. The data was from night time only, so the majority of the data was sleep. Because of this, the specificity was expected to be low.

Fallmann et al. [2020] introduced a machine learning method for sleep–wake classification using actigraphy and heart rate variability in addition to personal information: gender, race and health status. The machine learning method was a neural network. They proposed training separate models for different groups of people, and their personalized models outperformed the general model.

Lüdtke et al. [2021] compared using an HMM on accelerometer data with PSG for sleep–wake classification. The results were compared with two conventional methods for sleep recognition and the two simple machine learning algorithms linear discriminant analysis and logistic regression. The dataset consisted of 20 subjects with sleep-related diagnoses. HMM outperformed the other models, with an accuracy of 0.790.

Liu et al. [2020] proposed a method for unsupervised and personalized sleep–wake classification using HMM. They used a commercial wearable device, so the accessible activity was in the form of step count (*Fitbit step count*). In addition they used heart rate data from the device. An HMM using a combination of activity and heart rate was compared with HMMs for activity and heart rate separately. Their method classified more epochs as wake, compared to Fitbit's algorithm. The authors claim that this can indicate a more accurate classification, because commercial wearable devices tend to overestimate sleep. No PSG data was recorded to confirm the improvement.

Banfi et al. [2021] developed a sleep–wake classification method with CNN using raw accelerometer data as input. One of their main goals was to make a method with low computational cost. PSG was used as ground truth. The performance of their method was compared with 6 machine learning methods given 12 features: SVM, Random Forest, Naive Bayes, AdaBoost, GradientBoost and Perceptron. Their CNN method outperformed the others on the measurements Cohen's Kappa, F1, concordance and sensitivity, but not specificity.

3.2.2 Other classification tasks

For the following papers, classification of sleep stages or other tasks related to sleep were investigated.

Ferree et al. [2019] improved detection of time in bed using leg-worn accelerometers. This was done by using body orientation, activity and time with a combination of a Bayesian classifier and a decision tree. This resulted in an accuracy of 0.97 for detecting segments of time in bed.

Faerman et al. [2020] investigated the connection between experienced sleep quality and actigraphy with Lasso penalized regression and Random Forest. Personal data, for instance age, body mass index and education level, was also used as features for the machine learning methods. No clear association between the actigraphy data and experienced sleep quality was found.

Fedorin et al. [2019] proposed a method for three- and four-class sleep stage classification using photoplethysmogram (PPG) in addition to accelerometer data. Their proposed algorithm used linear discriminant analysis, and it was compared with the results found in other papers using Random Forest methods and SVM. For four classes the mean accuracy was 0.77, and for three classes it was 0.84. This was higher than the accuracy found in the papers used for comparison.

Willetts et al. [2018] proposed a Random Forest method for multi-class classification of activity, with sleep as one of the classes. To enable the Random Forest model to make use of the temporal aspect of the time series, an HMM was used to encode the temporal structure. Cameras and sleep diaries were used as the ground truth. 0.97 of the the minutes spent sleeping were correctly classified.

Walch et al. [2019] evaluated the use of Apple Watch with their own open-sourced app. Combinations of features from movement, heart rate and circadian rhythm were tested. Both sleep–wake and three-class classification were used in the experiments. Random Forest, logistic regression, K-NN and neural net were all tested. For sleep–wake classification the best result was achieved when all features were included, with an accuracy of approximately 0.8 for all four models compared with PSG.

Sundararajan et al. [2021] used Random Forest for sleep–wake classification, sleep stage classification, and non-wear detection, using PSG as the ground truth. Their results were compared with traditional sleep detection algorithms. Their Random Forest model achieved better results than the traditional methods, with an F1-score of 0.7591 for sleep–wake classification. However, the results for sleep stage classification were poor.

Hu and Shou [2021] investigated three-class classification of wake, REM and non-REM using only data from accelerometers. They proposed a method using LSTM, and compared the result with other machine learning methods. The LSTM model achieved an average accuracy of 0.65, and the authors state that this result demonstrates that actigraphy data is useful for classifying wake, REM and non-REM without other sensors.

3.3 Summary

To summarize, this section addresses the research question:
RQ 1.1: What are the major contributions within sleep detection with accelerometer data using machine learning methods during the last three years?

The deep learning methods LSTM and CNN are performing well on sleep classification tasks, according to several of the related papers in this review [Sano et al., 2019], [Cho et al., 2019], [Palotti et al., 2019], [Yildiz et al., 2019], [Banfi et al., 2021], [Hu and Shou, 2021]. This tells us that deep learning is an important contribution in the field of sleep detection. Nevertheless, these methods will not be investigated further in this work, as a large amount of data is needed in deep learning to achieve good results.

Other high performing methods are XGBoost and Random Forest [Khademi et al., 2019], [Willetts et al., 2018], [Sundararajan et al., 2021], which are highly relevant for our work. As the data segmentation of Khademi et al. [2019] is similar to ours, with epochs of 30 s and sliding window of 21 epochs, the best forming methods, XGBoost and Random Forest, are used in our experiments as well. HMM also performs well, according to Li et al. [2020], Lüdtke et al. [2021] and Liu et al. [2020], but we have chosen to focus on more traditional machine learning methods in this project, such that the same datasets, consisting of features of time segments, could be used for all methods. The problem of incorporating time dependencies, which can be handled by an HMM, is solved in our work by segmenting the data in to features calculated from several subsequent data points, as described in section 4.3.

As the HUNT4 dataset consists of accelerometer data and temperature data, the possibilities of data features are limited. The personal information of each subject is not available in any of our datasets, so the methods of Fallmann et al. [2020] are not possible to test in this project. No PPG data is available in the HUNT4 data, so the results of Fedorin et al. [2019] are not relevant, as the goal is to use our classifier on data from HUNT4. Similarly, the features used in Walch et al. [2019] are not relevant to our work, as the heart rate was not measured in the HUNT4 data collection. The work of Sano et al. [2019] is however highly relevant, as the combination of accelerometer and skin temperature data was superior to the other data combinations.

The sensor placements of the actigraphs in Li and Nakamura [2019] and Ferree et al. [2019] are interesting, as they are the only papers where the actigraph was not placed on the wrist. As the sensor on the trunk in the work of Li and Nakamura [2019] achieved similar results as the wrist sensor, we could expect that adding the wrist sensor data in our training sets would not make a large impact on the results. However, they do not combine the data of the two sensors, so the sensors could have achieved similar performance without giving the classification algorithm the same information. According to Ferree et al. [2019], a leg-worn actigraph results in a high performance of detection of time in bed, but this is known to be a much simpler task than sleep detection.

The datasets in the experiments of Lüdtke et al. [2021] and Yildiz et al. [2019] both consist of subjects with presumably abnormal sleep, as they have some type of sleep disorder or cognitive impairment. One of the training datasets in our work also falls into this category, as all the participants were diagnosed with a sleep disorder. For this reason, these results are good indicators for our work. However, the experiments in these two papers only used data with abnormal sleep, while we compare this with expected normal sleep data.

In contrast to all the papers described in this section, we are evaluating classification of actigraphy data from multiple body-worn sensors jointly. The back and thigh sensor placements are not used in any of the related papers, as the wrist is evidently the placement of the sensor in most experiments. For this reason it is interesting to combine the wrist data with the data from the back and thigh, as this can say something about what sensor location is ideal. In addition, we investigate the influence of including temperature data. This is previously done by Sano et al. [2019], with good results, but not in combination with several actigraphs.

Chapter 4

Methodology

In this chapter the datasets used in our experiments are described. The main procedure of our work is explained, as well as details regarding the data processing, specifically the segmentation of the data and the feature generation.

4.1 Datasets

Two machine learning datasets were created from available raw data, and these are used for training and testing of our methods. The raw datasets both consist of data collected from accelerometers and thermometers on the back, thigh and wrist, and the placement of the sensors is shown in figure 4.1. The sensors are the same as the ones used in HUNT4. PSG was also recorded for all the subjects in the datasets, and the labels are based on the PSG recordings.

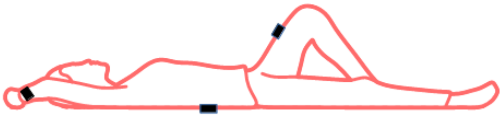


Figure 4.1: Sensor placement on the wrist, back and thigh.

In both datasets the sleep is categorized into five stages: wake, N1, N2, N3 and REM, where N1 and N2 are light sleep, and N3 is deep sleep. An example of the data is presented in figure 4.2. The data labels are coded as numbers, where 801 is wake, 802 is N1, 803 is N2, 804 is N3 and 805 is REM. For our experiments on sleep–wake classification, all data categories except wake are relabeled as 'sleep'.

For the other experiments, wake data is dropped. For REM sleep classification all non-REM data is relabeled to one category. Similarly, when classifying light sleep N1 and N2 are relabeled ’light’ and the rest ’not light’, and for deep sleep all except N3 are relabeled ’not deep’.

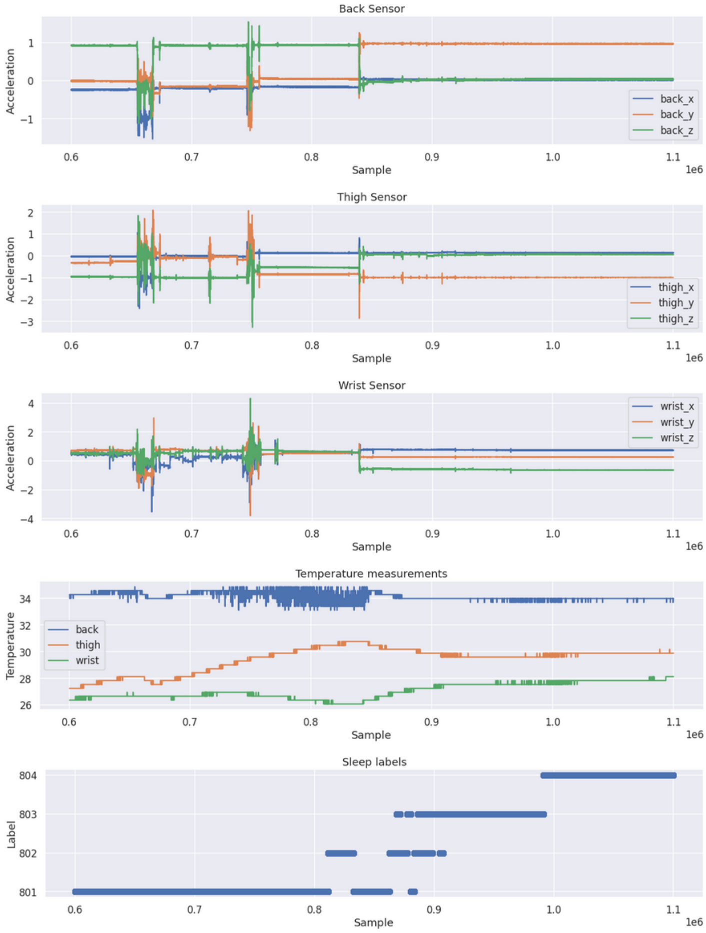


Figure 4.2: Example of data plot.

4.1.1 Sleep disorder data

This data is collected in a sleep lab at St Olavs Hospital, Trondheim, Norway, from 19 subjects, all diagnosed with a sleep disorder, for one night each. Data from 11 of these subjects is used in our experiments, as we are only able to extract the data from these subjects with our framework. The distribution of sleep and wake for each subject is visualized in figure 4.3, and the distribution of light, deep and REM sleep is shown in figure 4.4. In both figures it is evident that the dataset is unbalanced. For all subjects except subject 30, the number of wake epochs is less than 50 % of the sleep epochs, and all except subject 6 and 30 have light sleep as the majority sleep stage, significantly larger than deep and REM sleep. Because of the sleep disorders in these subjects, it is expected that the sleep patterns and movements are not representative for the general population.

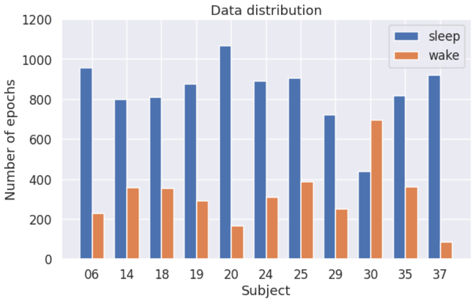


Figure 4.3: Distribution of sleep and wake for each subject in the sleep disorder dataset.

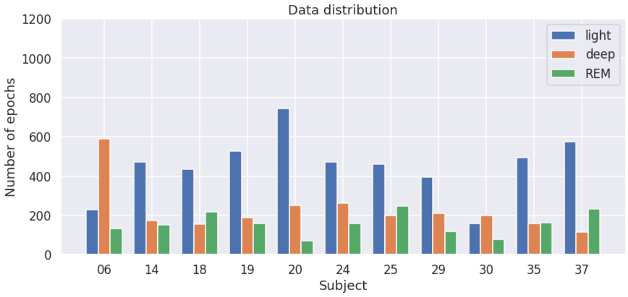


Figure 4.4: Distribution of sleep stages for each subject in the sleep disorder dataset.

4.1.2 Healthy sleep data

The 18 subjects of this dataset, collected in Oslo, Norway, does not have any reported sleep disorders. We use the data from 17 of the subjects, as the data of one subject is formatted differently, and would require extensive work. The distribution of sleep and wake is shown in figure 4.5, and the distribution of sleep classes can be seen in figure 4.6. In this dataset all subjects have a majority of sleep, but there is much variety in the relative amount of sleep compared to wake. For instance, for subject 5 the amount of wake is approximately 70 % of the amount of sleep, while for subject 17 the wake samples are only around 15 % of the sleep samples. All subjects have a majority of light sleep compared to the other sleep stages.

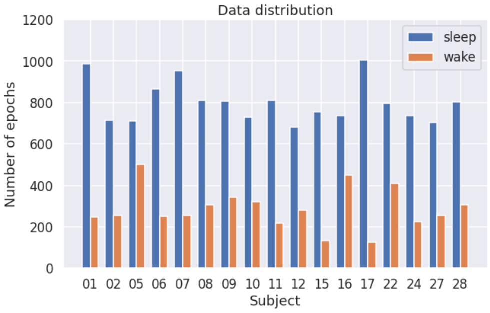


Figure 4.5: Distribution of sleep and wake for each subject in the healthy sleep dataset.

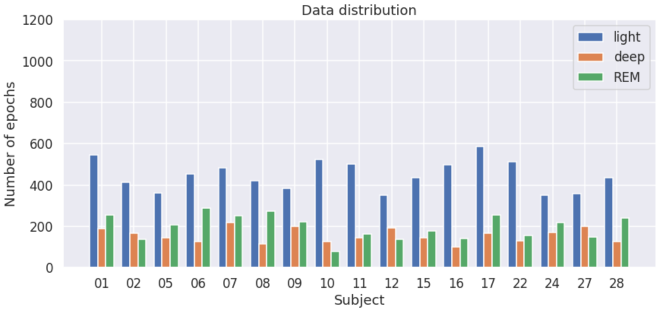


Figure 4.6: Distribution of sleep stages for each subject in the healthy sleep dataset.

4.2 Model development process

The structure of our experimental process is presented in figure 4.7. Raw acceleration data from the x-, y- and z-axis of the sensors and temperature data were recorded at a frequency of 100 Hz. As in our previous work, the feature generation and data segmentation are adapted from Hay [2019], and this is done for both datasets. The processing is explained in detail in section 4.3.

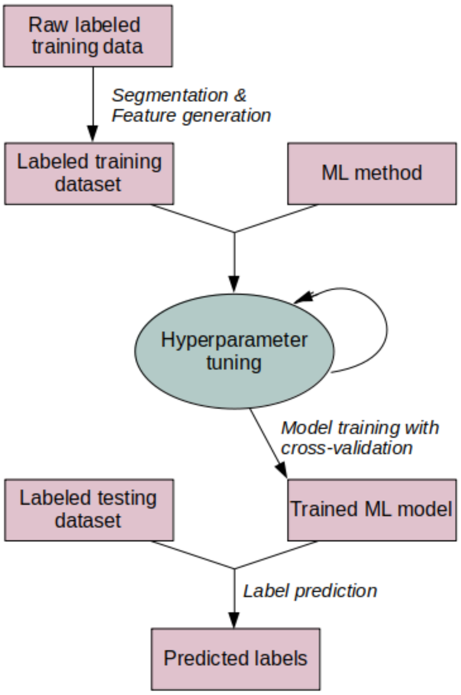


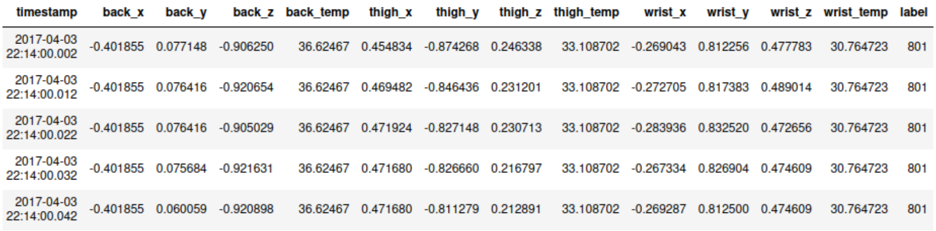
Figure 4.7: Experiment procedure

Based on the results of our literature review in chapter 3 and our previous work [Gryvill, 2020], we decide to use the algorithms Random Forest, XGBoost, K-NN and SVM. These have all shown to be good methods for sleep classification. By using multiple algorithms separately, the experiments are more robust, as one unsuited method will not weaken the results of the other methods. Another important aspect of the methods is the fact that Random Forest and XGBoost are tree-based methods, while K-NN and SVM are distance-based. While the tree-based models will look into one feature at the time and may ignore nonrelevant features, the distance-based models will look at the placement of the data sample compared to the other samples, all features together. This will probably give different results, depending on what type of method is most fitting for the classification tasks.

The machine learning models are trained on the sleep disorder dataset, and to exploit the strengths of the machine learning algorithms, the hyperparameters are tuned as described in section 4.5. Both training and tuning are done with leave-one-group-out cross-validation, and the performance of the models using the sleep disorder data is evaluated with the cross-validation. For a final evaluation of the models, they are trained on the entire sleep disorder dataset and tested on the healthy sleep data.

4.3 Segmentation and feature generation

The raw data consists of 3 values from each of the 3 accelerometers for each $\frac{1}{100}$ of a second, in addition to a temperature measurement for each sensor. An example of the raw data is shown in figure 4.8. The label, represented by a number, and the timestamps are extracted to separate files, as they are not used in the features. The labels, derived from PSG, are used as the ground truth when evaluating the predictions of the algorithms.



timestamp	back_x	back_y	back_z	back_temp	thigh_x	thigh_y	thigh_z	thigh_temp	wrist_x	wrist_y	wrist_z	wrist_temp	label
2017-04-03 22:14:00.002	-0.401855	0.077148	-0.906250	36.62467	0.454834	-0.874268	0.246338	33.108702	-0.269043	0.812256	0.477783	30.764723	801
2017-04-03 22:14:00.012	-0.401855	0.076416	-0.920654	36.62467	0.469482	-0.846436	0.231201	33.108702	-0.272705	0.817383	0.489014	30.764723	801
2017-04-03 22:14:00.022	-0.401855	0.076416	-0.905029	36.62467	0.471924	-0.827148	0.230713	33.108702	-0.283936	0.832520	0.472656	30.764723	801
2017-04-03 22:14:00.032	-0.401855	0.075684	-0.921631	36.62467	0.471680	-0.826660	0.216797	33.108702	-0.267334	0.826904	0.474609	30.764723	801
2017-04-03 22:14:00.042	-0.401855	0.060059	-0.920898	36.62467	0.471680	-0.811279	0.212891	33.108702	-0.269287	0.812500	0.474609	30.764723	801

Figure 4.8: Sample of raw data.

The data is segmented into epochs of 30 seconds, as PSG data usually is in this format and it is used in several of the related papers [Fedorin et al., 2019], [Khademi et al., 2019], [Palotti et al., 2019], [Walch et al., 2019], [Yildiz et al., 2019], [Fallmann et al., 2020], [Sundararajan et al., 2021], [Banfi et al., 2021], [Hu and Shou, 2021]. The features are calculated for each window of 21 epochs, 10.5 minutes, because this gives one target epoch in the center and a normal window size compared to other papers. We use a sliding window, illustrated in figure 4.9, which means that the first window consists of the first 21 epochs, and the second window does not include the first epoch, but the following 21.

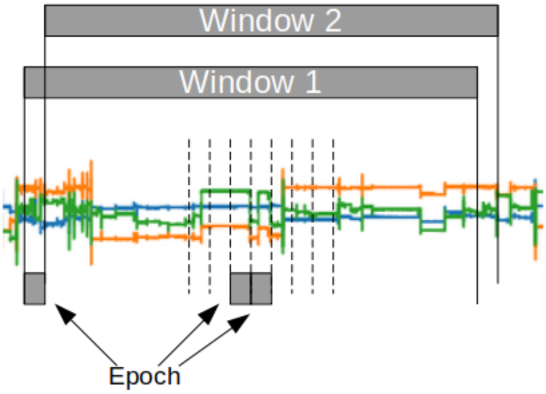


Figure 4.9: Illustration of data segmentation. The blue, orange and green lines are the acceleration data of the x-, y- and z-axis, respectively, from one sensor.

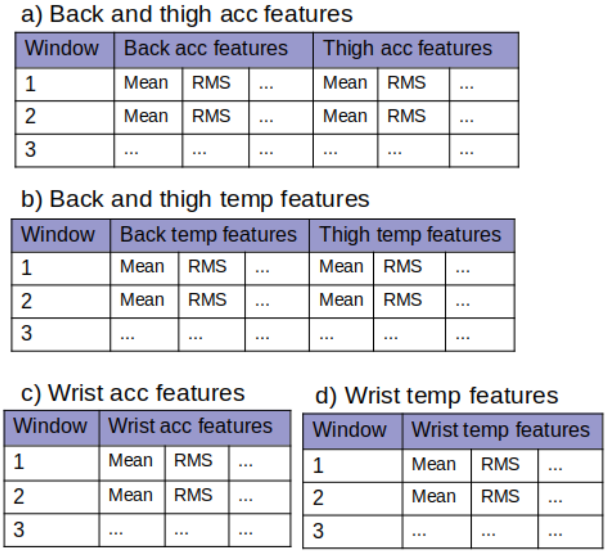
The body movements of a subject during sleep and wakefulness are highly individual. For this reason, the changes in the values are more relevant than the values them self in this type of time series data. The generated features represent aspects of the change in the data series, for instance the mean value and standard deviation, which are more general for different subjects. The datasets used for training and testing of the machine learning models does not include the raw data, only the features derived from them. A list of all features is presented in table 4.1. These are based on the work of Hay [2019], as the features led to promising results for similar data in her work. For the back and thigh accelerometer data, the features are generated for each window of the x-, y- and z-axis of both sensors. In addition, the features are generated for the norm of the combined data. When data from the wrist sensor is included, the feature generation is also used on each of the wrist sensor axis, and on the norm of the back, thigh and wrist data combined. The same features are applied for the temperature data. The *energy* feature is calculated for each axis separately, as for the other features, and additionally for each accelerometer sensor combined. For the experiments with only back and thigh accelerometer data, this results in 100 features in total. Including the temperature data of the back and thigh sensor, the total number is 128, and for acceleration and temperature data from all three sensors, a total of 199 features are extracted.

Table 4.1: Features. Adapted from Hay [2019].

Feature	Description
Mean	The average value of the data.
Root mean square	Square root value of the mean square of the data.
Standard deviation	Standard deviation of the data. A measure used to quantify the amount of variation or dispersion of a set of data values.
Kurtosis	Kurtosis of the data. A measure of the peakedness of the data.
Skewness	The skewness of the distribution of the data. A measure of the asymmetry of the probability distribution.
Sum of values	Sum of values of the data.
Coefficient of variation	The ratio of the standard deviation to the mean of the data.
Zero crossings	The number of zero crossings in the data.
Interquartile range	The difference between the 75th and 25th percentile.
Min-max-mean	The average of the differences between local minimums and maximums.
Energy	The signal's energy.
Maximums - central	Number of local maximums in the data for the central epoch.
Maximums - first 10	Number of local maximums in the data for the combined first 10 epochs.
Maximums - last 10	Number of local maximums in the data for the combined last 10 epochs.

4.4 Final dataset structure

The raw data of the two datasets are segmented and features extracted for each subject separately. Because we are experimenting with different subsets of data features, they are stored in multiple files. This structure is shown in figure 4.10. The features extracted from the back and thigh accelerometer data are stored in one file, while the temperature features of the back and thigh are stored in a separate file. Similarly, the accelerometer features of the wrist are kept in one file, and the wrist temperature features in another. The relevant tables are combined when the data is used by the machine learning models. In all of the files, the features are columns and the time windows are rows. The amount of data for each subject in the dataset is described in appendix A.



a) Back and thigh acc features

Window	Back acc features			Thigh acc features		
1	Mean	RMS	...	Mean	RMS	...
2	Mean	RMS	...	Mean	RMS	...
3	...	...	...	...	...	...

b) Back and thigh temp features

Window	Back temp features			Thigh temp features		
1	Mean	RMS	...	Mean	RMS	...
2	Mean	RMS	...	Mean	RMS	...
3	...	...	...	...	...	...

c) Wrist acc features

Window	Wrist acc features		
1	Mean	RMS	...
2	Mean	RMS	...
3	...	...	...

d) Wrist temp features

Window	Wrist temp features		
1	Mean	RMS	...
2	Mean	RMS	...
3	...	...	...

Figure 4.10: Dataset structure for one subject. RMS = root mean square, acc = acceleration, temp = temperature.

4.5 Hyperparameter tuning

As mentioned in section 4.2, the hyperparameters of the classifiers are tuned to fit our data and the classification tasks. Without this step, the algorithms will not have the optimal setup, which will lead to a negative effect on the results. The hyperparameters of the classifiers are chosen using grid search, where multiple values are given for each parameter and all combinations are tested. This is run in multiple iterations to ensure that the best values are found. The method is set to improve the F1-score of the classifier, as this measurement considers the balance of precision and recall. Relevant hyperparameters for each of the four classifiers are described in table 2.1. These parameters are tuned, and the optimal values for the hyperparameters in each experiment, found through the grid search, are given in appendix B.

Chapter 5

Experiments and Results

The experiments executed in our research are presented in this chapter with all necessary information to repeat them. The results are described in the end of the chapter.

5.1 Experimental Plan

The experiments are conducted to answer the research questions of goal 2 and 3 in chapter 1. They are divided into four groups: sleep–wake classification, light sleep classification, deep sleep classification and REM sleep classification. For each group, the four classification methods Random Forest, XGBoost, K-NN and SVM, are tested.

The sets of features used in each experiment are illustrated in table 5.1. To answer research question 2.1 and 3.1, a baseline classification of only features from the back and thigh accelerometer data is conducted, experiment 1 in the table. This is compared with classification where the temperature data of the back and thigh sensor are included in the input features, experiment 2. For research question 2.2 and 3.2, experiment 2 is compared with experiment 3, consisting of features from both accelerometer data and temperatures from the back, thigh and wrist sensors.

The models are trained on the sleep disorder dataset after hyperparameter tuning, using leave-one-group-out cross-validation, where the data of one subject is one group. All data of one subject is held out to ensure that the algorithm is not trained on data from that subject already, as this would give unrealistically good results compared to data from new unseen subjects. As PSG data is normally collected from subjects with presumed sleep disorders, most labeled datasets in this field does not represent a healthy population. For a final evaluation the models are tested on the healthy sleep dataset. This dataset is probably more similar to the HUNT4 data, so this can give an indication of how models trained on sleep disordered subjects perform on the HUNT4 dataset. The accuracy, F1-score, AUC, sensitivity and specificity of the models are compared to evaluate them.

Table 5.1: Feature sets used in the experiments.
Acc = acceleration, temp = temperature, exp = experiment.

	Back & thigh	Back, thigh & wrist
Acc	Exp 1	–
Acc & temp	Exp 2	Exp 3

5.2 Experimental Setup

All code is written in Python with the machine learning library Scikit-learn[1]. As the XGBoost algorithm is not included in Scikit-learn, this method is imported from its own library[2].

5.3 Experimental Results

In this section all the results of our experiments are presented. The first subsection contains the results from the sleep disorder dataset, where the values are the mean values from the cross-validation after tuning the hyperparameters. The final subsection presents the results where the trained models are tested on the healthy sleep data. All the results are presented by accuracy, F1-score, AUC, sensitivity and specificity.

5.3.1 Testing on sleep disorder data

The sleep disorder dataset is mainly used for training of the models, so the results presented are the mean of the results for each round of classification with the data of one subject as the test data. The best results, using the hyperparameters in appendix B, are presented in the following sections.

[1] `https://scikit-learn.org/stable/`, accessed 2021-11-08

[2] `https://xgboost.ai/`, accessed 2021-11-08

Sleep–wake classification

The results of all the sleep–wake classification are presented in table 5.2. The top section of the table contains the results with only acceleration data from the back and thigh sensors. The second section presents the results for the experiment including the temperature data from back and thigh, and the results with all data from all three sensors are shown in the bottom part of the table.

Table 5.2: Sleep–wake classification results. RF = Random Forest, XGB = XGBoost.

		Accuracy	F1	AUC	Sensitivity	Specificity
Exp 1	RF	**0.89**	**0.92**	**0.92**	0.97	0.65
Thigh &	XGB	**0.89**	**0.92**	0.91	0.95	**0.72**
back	K-NN	0.85	0.89	0.91	0.92	0.66
	SVM	0.88	**0.92**	0.90	**0.99**	0.58
Exp 2	RF	**0.89**	**0.92**	**0.92**	**0.98**	0.64
Thigh &	XGB	**0.89**	**0.92**	**0.92**	0.96	**0.70**
back with	K-NN	0.86	0.90	0.91	0.94	0.65
temperature	SVM	0.87	0.91	0.91	0.97	0.61
Exp 3	RF	**0.91**	0.93	0.93	**0.97**	**0.72**
Thigh, back	XGB	**0.91**	**0.94**	**0.94**	**0.97**	**0.72**
& wrist with	K-NN	0.87	0.91	0.93	0.94	0.69
temperature	SVM	0.89	0.92	0.92	**0.97**	0.67

As described in the table, Random Forest and XGBoost achieve the best accuracy for all subsets of data, with the highest score of 0.91 for both methods in experiment 3, consisting of thigh, back and wrist data with temperature. For the F1-score, Random Forest, XGBoost and SVM all achieve 0.92 with only the thigh and back accelerometer data. The same score is achieved for Random Forest and XGBoost when including the temperature, while SVM scored 0.91, slightly lower. XGBoost stands out with the best F1-score, 0.94, with all data included. The highest AUC score is achieved by XGBoost with features from all the data. The sensitivity is highest for SVM in experiment 1, and the best specificity of 0.72 is achieved by XGBoost for thigh and back, and XGBoost and Random Forest with all data. K-NN is performing worse than the other methods for most metrics, especially accuracy and sensitivity.

For all methods it is evident that including the temperature data is not of great significance. However, including the wrist data results in a minor improvement for all metrics of all the algorithms except for the sensitivity of Random Forest and SVM.

Light sleep classification

For classification of light and non-light sleep using the sleep disorder data, the results of the cross-validation are shown in table 5.3.

Table 5.3: Light sleep classification results. RF = Random Forest, XGB = XGBoost.

		Accuracy	F1	AUC	Sensitivity	Specificity
Exp 1	RF	0.52	0.59	0.58	0.67	0.39
Thigh &	XGB	**0.56**	**0.62**	**0.60**	**0.70**	0.44
back	K-NN	0.52	0.57	0.51	0.62	0.41
	SVM	0.54	0.57	0.55	0.62	**0.46**
Exp 2	RF	0.54	0.61	0.57	0.71	0.37
Thigh &	XGB	**0.56**	0.61	**0.60**	0.67	**0.48**
back with	K-NN	0.52	0.57	0.52	0.63	0.41
temperature	SVM	0.53	**0.69**	0.54	**1.00**	0.01
Exp 3	RF	0.58	0.63	0.64	0.71	0.45
Thigh, back	XGB	**0.61**	0.64	**0.66**	0.69	**0.53**
& wrist with	K-NN	0.56	0.67	0.58	0.88	0.21
temperature	SVM	0.53	**0.69**	0.60	**1.00**	0.00

All the best metric values are in the bottom part of the table, experiment 3, including all the data features. XGBoost achieves the highest accuracy, AUC and specificity, respectively 0.61, 0.66 and 0.53. The highest F1-score and sensitivity, 0.69 and 1.00 respectively, are from SVM both for all data and for thigh and back data with temperature. It should be noted that, however, when achieving a sensitivity of 1.00 the specificity is close to 0.

Deep sleep classification

Deep sleep classification results in the scores presented in table 5.4.

The results for deep sleep show that the highest score of each metric is achieved by either Random Forest or XGBoost. The highest accuracy of 0.72 is shared by Random Forest in experiment 1, and XGBoost in experiment 3. The best

Table 5.4: Deep sleep classification results. RF = Random Forest, XGB = XGBoost.

		Accuracy	F1	AUC	Sensitivity	Specificity
Exp 1	RF	**0.72**	0.20	**0.61**	0.15	**0.95**
Thigh &	XGB	0.68	**0.30**	0.59	0.29	0.86
back	K-NN	0.65	**0.30**	0.56	**0.32**	0.80
	SVM	0.66	0.26	0.53	0.25	0.83
Exp 2	RF	0.70	**0.27**	**0.62**	0.23	0.90
Thigh &	XGB	**0.71**	**0.27**	**0.62**	0.23	**0.91**
back with	K-NN	0.66	**0.27**	0.55	**0.27**	0.82
temperature	SVM	0.67	0.24	0.54	0.23	0.84
Exp 3	RF	0.71	0.19	**0.71**	0.17	**0.94**
Thigh, back	XGB	**0.72**	**0.35**	0.70	**0.33**	0.89
& wrist with	K-NN	0.68	0.28	0.56	0.26	0.85
temperature	SVM	0.68	0.23	0.59	0.21	0.87

F1-score and sensitivity of 0.35 and 0.33 respectively, are achieved by XGBoost with all features, Random Forest with all features have the highest AUC of 0.71, and experiment 1 with Random Forest have the highest specificity of 0.95.

The AUC-score of all algorithms except K-NN improve drastically when the wrist data is included in the features. The same pattern is not seen in the other metrics.

REM sleep classification

For the classification of REM and non-REM sleep, the results of the experiments are presented in table 5.5.

The REM-sleep classification results in K-NN outperforming the other methods with respect to F1-score, AUC and sensitivity for all experiments. These results are slightly higher for each addition of features, except for the AUC. Random Forest has the highest accuracy and specificity of 0.80 and 0.98 respectively, both for the baseline experiment. For the experiment with all features included, XGBoost has the highest accuracy and specificity, but the results are lower than the best of Random Forest. In general it should be noted that all the AUC-scores are close to 0.5 or below. None of the sensitivity scores are higher than 0.21, and the highest F1-score is 0.19.

Table 5.5: REM classification results. RF = Random Forest, XGB = XGBoost.

		Accuracy	F1	AUC	Sensitivity	Specificity
Exp 1 Thigh & back	RF	**0.80**	0.02	0.46	0.01	**0.98**
	XGB	0.76	0.09	0.44	0.08	0.92
	K-NN	0.69	**0.17**	**0.51**	**0.19**	0.82
	SVM	0.67	0.15	0.45	0.14	0.80
Exp 2 Thigh & back with temperature	RF	**0.79**	0.02	0.44	0.01	**0.98**
	XGB	0.78	0.08	0.48	0.06	0.96
	K-NN	0.68	**0.18**	**0.49**	**0.19**	0.80
	SVM	0.65	0.13	0.43	0.14	0.78
Exp 3 Thigh, back & wrist with temperature	RF	0.75	0.14	0.46	0.11	0.90
	XGB	**0.77**	0.17	0.49	0.14	**0.91**
	K-NN	0.69	**0.19**	**0.51**	**0.21**	0.81
	SVM	0.73	0.10	0.45	0.10	0.88

5.3.2 Testing on healthy sleep data

After training with cross-validation on the sleep disorder dataset, with the results presented in the former sections, the models are trained on the entire sleep disorder dataset and tested on the healthy sleep dataset.

Sleep–wake classification

The results of the sleep–wake classification of the healthy sleep dataset are presented in table 5.6.

The best accuracy and F1-score for sleep–wake classification on healthy sleep data is 0.85 and 0.90, respectively, achieved by Random Forest in the experiments with temperature included, experiments 2 and 3. The same F1-score is also achieved by Random Forest in experiment 1 and XGBoost in all experiments. Random Forest and XGBoost also have the highest AUC-score of 0.86 in experiment 3. A sensitivity of 0.98 is achieved by Random Forest in all experiments, XGBoost in experiment 1 and 3, and SVM with only thigh and back acceleration features. The highest specificity, 0.49, is reached by XGBoost in experiment 2, and Random Forest in experiment 3. The results for this experiment have very little change when using different sets of features. The most significant difference is in the specificity. The performance of all algorithms are also very similar, but Random Forest and XGBoost stands out as slightly better than the two distance-based methods.

Table 5.6: Sleep–wake results on the healthy sleep data, trained on the sleep disorder data. RF = Random Forest, XGB = XGBoost.

		Accuracy	F1	AUC	Sensitivity	Specificity
Exp 1 Thigh & back	RF	**0.84**	**0.90**	0.82	**0.98**	**0.46**
	XGB	**0.84**	**0.90**	**0.84**	**0.98**	**0.46**
	K-NN	0.81	0.88	0.81	0.97	0.35
	SVM	0.83	0.89	0.82	**0.98**	0.39
Exp 2 Thigh & back with temperature	RF	**0.85**	**0.90**	**0.84**	**0.98**	0.48
	XGB	0.84	**0.90**	**0.84**	0.97	**0.49**
	K-NN	0.82	0.89	0.83	0.96	0.43
	SVM	0.82	0.89	0.82	0.94	0.48
Exp 3 Thigh, back & wrist with temperature	RF	**0.85**	**0.90**	**0.86**	**0.98**	**0.49**
	XGB	0.84	**0.90**	**0.86**	**0.98**	0.44
	K-NN	0.82	0.89	0.84	0.97	0.41
	SVM	0.82	0.89	0.82	0.96	0.46

Light sleep classification

The light sleep classification results in the metrics presented in table 5.7.

For the AUC-score of the light sleep classification, the highest result is 0.53, achieved using XGBoost in experiment 1 and SVM in experiment 3. For the rest of the evaluation metrics the best results are from both of the experiments with temperature data, 2 and 3, with 0.56 for accuracy, 0.72 for F1-score, 1.00 for Sensitivity and 0.52 for specificity. SVM outperforms the other methods with respect to accuracy, F1 and sensitivity, while XGBoost has the highest performance for specificity.

Deep sleep classification

The trained deep sleep models give the results presented in table 5.8.

The table shows that Random Forest performs best according to accuracy, AUC and specificity, with the results 0.77, 0.67 and 0.94 respectively when all features are included, experiment 3. The same accuracy and specifity are also achieved by Random Forest in experiment 1. The best F1-score of 0.31 is achieved by XGBoost in experiment 3, and the highest sensitivity of 0.37 by SVM on the feature set containing thigh and back data with temperature, experiment 2.

Table 5.7: Light sleep classification results on healthy sleep data.
RF = Random Forest, XGB = XGBoost.

		Accuracy	F1	AUC	Sensitivity	Specificity
Exp 1	RF	0.51	0.57	0.50	0.58	0.42
Thigh &	XGB	**0.52**	0.56	**0.53**	0.54	**0.49**
back	K-NN	0.51	0.59	0.50	0.62	0.38
	SVM	**0.52**	**0.60**	0.50	**0.65**	0.34
Exp 2	RF	0.50	0.54	0.50	0.52	0.46
Thigh &	XGB	0.51	0.53	**0.51**	0.50	**0.52**
back with	K-NN	0.48	0.53	0.47	0.52	0.42
temperature	SVM	**0.56**	**0.72**	0.51	**1.00**	0.00
Exp 3	RF	0.51	0.55	**0.51**	0.54	0.47
Thigh, back	XGB	0.51	0.53	0.51	0.50	**0.52**
& wrist with	K-NN	0.53	0.63	0.50	0.71	0.31
temperature	SVM	**0.56**	**0.72**	**0.53**	**1.00**	0.00

Table 5.8: Deep sleep classification results on healthy sleep data.
RF = Random Forest, XGB = XGBoost.

		Accuracy	F1	AUC	Sensitivity	Specificity
Exp 1	RF	**0.77**	0.11	**0.61**	0.07	**0.94**
Thigh &	XGB	0.67	**0.30**	0.60	**0.36**	0.75
back	K-NN	0.71	0.22	0.54	0.21	0.83
	SVM	0.66	0.22	0.51	0.25	0.75
Exp 2	RF	**0.74**	0.21	**0.62**	0.17	**0.88**
Thigh &	XGB	0.72	**0.27**	**0.62**	0.27	0.83
back with	K-NN	0.68	0.24	0.53	0.25	0.79
temperature	SVM	0.62	**0.27**	0.52	**0.37**	0.68
Exp 3	RF	**0.77**	0.12	**0.67**	0.08	**0.94**
Thigh, back	XGB	0.72	**0.31**	0.65	**0.32**	0.82
& wrist with	K-NN	0.71	0.23	0.54	0.23	0.82
temperature	SVM	0.70	0.25	0.57	0.25	0.81

There is in general large variations for all metrics of Random Forest, XGBoost and SVM, comparing the different experiments. For instance the F1-score of Random Forest changes from 0.11 to 0.21 when the temperature is included, and it changes to 0.12 when the wrist data is included.

REM sleep classification

For the testing of the REM classification models on the healthy data, the results are as described in table 5.9.

Table 5.9: REM classification results on healthy sleep data. RF = Random Forest, XGB = XGBoost.

		Accuracy	F1	AUC	Sensitivity	Specificity
Exp 1	RF	**0.75**	0.04	0.54	0.00	**1.00**
Thigh and	XGB	0.74	0.04	0.55	0.02	0.98
back	K-NN	0.66	**0.22**	0.50	**0.20**	0.81
	SVM	0.69	0.15	**0.54**	0.11	0.89
Exp 2	RF	**0.73**	0.06	0.52	0.04	**0.96**
Thigh &	XGB	0.72	0.12	0.55	0.08	0.92
back with	K-NN	0.66	**0.22**	0.50	**0.20**	0.81
temperature	SVM	0.67	0.19	**0.54**	0.16	0.84
Exp 3	RF	**0.72**	0.15	0.55	0.10	**0.92**
Thigh, back	XGB	0.71	0.14	**0.57**	0.10	0.91
& wrist with	K-NN	0.65	**0.21**	0.49	**0.19**	0.80
temperature	SVM	0.68	0.16	0.49	0.12	0.86

The over all highest accuracy and specificity, 0.75 and 1.00 respectively, are achieved with Random Forest in the baseline experiment. The sensitivity is highest, 0.20, when K-NN is used on the thigh and back data, both in experiment 1 and 2, and the best F1-score of 0.22 is also achieved by K-NN in the same experiments. XGBoost with wrist data achieves an AUC-score of 0.57, while the other AUC-scores are slightly lower, but all around 0.5. It can be seen that the sensitivity in general is low and the specificity is high for all methods. This means that most of the non-REM data is correctly classified as non-REM, but not much of the REM data is correctly classified.

Chapter 6

Discussion

The results from the experiments described in the previous section are discussed in this chapter, in relation to the research questions of goal 2, to improve sleep–wake classification, and goal 3, to improve sleep stage classification.

6.1 Sleep–wake classification

The first research question for sleep–wake classification is

RQ 2.1 How does including temperature data in the dataset impact the classification results?

From our experiments with sleep–wake classification it is evident that including the temperature features gives a minor improvement of the performance for some of the models, but the score of the best performing models, Random Forest and XGBoost, does not change. This can be seen in table 6.1, showing the F1-score for all experiments on sleep and wake. This indicates that when the movements of the back and thigh are already known, the changes in skin temperature between sleep and wake are not relevant, or not similar enough for different subjects, to distinguish better between sleep and wake. Another possible explanation for the lack of improvement, is the choice of temperature features. This could be investigated further by experiments where temperature features are tested and feature importance is measured. Table 6.1 also shows that the results from the cross-validation on the sleep disorder dataset are close to the results of the healthy sleep data, which tells us that the models did not overfit to the sleep disorder data, and that the two datasets are very similar regarding sleep and wake.

Table 6.1: Sleep–wake F1 results. SDD = sleep disorder data, HSD = healthy sleep data, RF = Random Forest, XGB = XGBoost.

		SDD	HSD
Exp 1 Thigh & back	RF	**0.92**	**0.90**
	XGB	**0.92**	**0.90**
	K-NN	0.89	0.88
	SVM	**0.92**	0.89
Exp 2 Thigh & back with temperature	RF	**0.92**	**0.90**
	XGB	**0.92**	**0.90**
	K-NN	0.90	0.89
	SVM	0.91	0.89
Exp 3 Thigh, back & wrist with temperature	RF	0.93	**0.90**
	XGB	**0.94**	**0.90**
	K-NN	0.91	0.89
	SVM	0.92	0.89

RQ 2.2 How does including wrist sensor data in the dataset impact the classification results?

To answer this research question, it should be noted that by including the wrist features, the increase in performance is higher than the increase for the first two experiments, but only slightly. However, there is an increase in accuracy, F1, AUC and specificity for all algorithms with the sleep disorder data, shown in table 5.2, which indicates that there are some aspects of the wrist data that are characteristic for sleep and wake. Nevertheless, the results without the features from the wrist data are high, and when testing on the healthy dataset, the improvement is not as evident, as shown in table 5.6. Because of this, the impact of the wrist data on sleep–wake classification is not of great importance.

Through our experiments we do not have results for every possible set of features. For this reason the value of the wrist data is not fully investigated, and it could, for instance, be the case that the wrist sensor alone would outperform the feature sets we have used in our experiments. This is however not very likely, as the tree-based methods, Random Forest and XGBoost, builds multiple trees with different subsets of the data, and various features are used in the splits of the trees. If the wrist features were much more informative than the back and thigh features, it is likely that the results of experiment 3 would have been significantly better than experiment 2. As this is not the case, we assume that the wrist data would not give drastically better results on its own. In summary, the best results from sleep–wake classification are achieved with the wrist data included, but all experiments perform well.

For both datasets the tree-based methods, Random Forest and XGBoost, have the best achievements of most experiments and metrics in the sleep–wake classification. This indicates that tree structures are well suited for this task.

6.2 Sleep stage classification

For classification of sleep stages we would like to know:

RQ 3.1 How does including temperature data in the dataset impact the classification results for light, deep and REM sleep, respectively?

By including the temperature features in the classification of light sleep, the SVM model has a big improvement of F1-score and sensitivity, but the specificity drops to 0.01. This means that almost all data points are classified as light sleep, which can be explained by the fact that most of the sleep data samples are light sleep, as shown in figure 6.1. Because of this, it is possible to achieve a high score on most performance metrics by classifying all samples as the majority class. As the majority class is also the positive class in this case, the F1 score is not drastically reduced by this. Apart from this major change for SVM, the results are very similar for the feature sets with and without temperature, as shown by the F1-scores in table 6.2. Similarly as for the sleep–wake classification, this can indicate that the temperature is not of great importance, or that our temperature features are inappropriate for the classification task. The results for the healthy sleep dataset shows that using the temperature data gives poorer results for Random Forest, XGBoost and K-NN. The reason for this could be a difference in temperature change for people with sleep disorders and without. This can be investigated by analyzing the temperature data further.

For all algorithms in the deep sleep classification, all metrics change when the temperature data is included. Some metrics improves while other metrics decline for all methods, however the changes are not in the same direction for all methods. This can be seen in table 5.4, and it tells us that the temperature data has different impact on the algorithms, and none of the methods have strictly better performance. In other words, the addition of temperature measurements does not clearly result in an improvement of the deep sleep classification. If the F1-score is given the highest priority, the best results are achieved with XGBoost and K-NN on the baseline feature set, as shown in table 6.2. The table also shows that with Random Forest, the healthy sleep data achieves significantly lower F1-scores than the sleep disorder data for all feature sets of the deep sleep classification. This indicates that the model is overfitting. It could also be the result of a difference in the deep sleep patterns of healthy and non-healthy sleep,

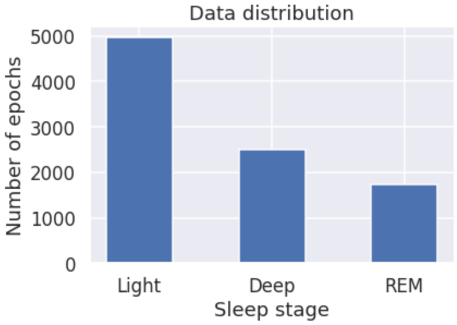


Figure 6.1: Sum of sleep stage epochs for all subjects in the sleep disorder dataset.

but the performance of XGBoost on the same data disproves this claim. The overfitting of the Random Forest model should have been avoided by retraining of the model, preferably on a more balanced version of the dataset.

The effects of including the temperature data in the REM classification experiment are in general small, similarly to the sleep–wake and light sleep classifications. However the low AUC-scores of all experiments, shown in table 5.5, suggest that the algorithms have difficulties separating the two classes. This was expected because the most distinctive aspects of REM sleep are the rapid eye movements and the brain activity, and none of these are detected by the accelerometers in our datasets. Additionally, REM sleep is a small part of the total sleep data, as shown in figure 6.1, so the algorithms tend to classify most of the data as non-REM. This explains the low F1-score and sensitivity, as REM sleep is the positive class in this case. By balancing the dataset, the results could possibly have been improved. However, for the deep sleep classification the performance was not as poor, even though the data distribution was only slightly better, which implies that the type of data is the main cause of the poor results. Comparing the F1-scores of REM classification of the sleep disorder data with the healthy sleep data, there is no clear pattern in the differences. The variation is, however, quite big, for instance the F1-score of XGBoost changes from 0.09 to 0.08 in the sleep disorder dataset, but it changes from 0.04 to 0.12 in the healthy sleep dataset. This can be caused by the poor ability to differentiate between the classes, leading to more randomness in the classification.

RQ 3.2 How does including wrist sensor data in the dataset impact the classification results for light, deep and REM sleep, respectively?

Inclusion of the wrist data features in the light sleep classification improves

Table 6.2: Sleep stage classification F1 results. SDD = Sleep disorder data, HSD = healthy sleep data, RF = Random Forest, XGB = XGBoost.

		Light sleep		Deep sleep		REM sleep	
		SDD	HSD	SDD	HSD	SDD	HSD
Exp 1 Thigh & back	RF	0.59	0.57	0.20	0.11	0.02	0.04
	XGB	**0.62**	0.56	**0.30**	**0.30**	0.09	0.04
	K-NN	0.57	0.59	**0.30**	0.22	**0.17**	**0.22**
	SVM	0.57	**0.60**	0.26	0.22	0.15	0.15
Exp 2 Thigh & back with temperature	RF	0.61	0.54	**0.27**	0.21	0.02	0.06
	XGB	0.61	0.53	**0.27**	**0.27**	0.08	0.12
	K-NN	0.57	0.53	**0.27**	0.24	**0.18**	**0.22**
	SVM	**0.69**	**0.72**	0.24	**0.27**	0.13	0.19
Exp 3 Thigh, back & wrist with temperature	RF	0.63	0.55	0.19	0.12	0.14	0.15
	XGB	0.64	0.53	**0.35**	**0.31**	0.17	0.14
	K-NN	0.67	0.63	0.28	0.23	**0.19**	**0.21**
	SVM	**0.69**	**0.72**	0.23	0.25	0.10	0.16

all metrics of XGBoost and all except sensitivity for Random Forest using the sleep disorder dataset, as shown in table 5.3. This indicates that the wrist data is providing useful information that separates light sleep from the other sleep stages. K-NN has similar improvements, but with a large decline in specificity. As the light sleep class is the majority class, the improvement in the K-NN classifier is clearly made at the expense of correctly classifying the non-light sleep. This is a general problem when the classes are highly unbalanced. For the SVM classifier the only change is an improvement of AUC-score. In general, it is evident that the wrist data has a positive impact on classification of light sleep. Table 6.2 shows that Random Forest, XGBoost and K-NN all achieve better F1-scores using the sleep disorder data compared to the healthy sleep data. This indicates overfitting for the three models. SVM, is however performing better on the separate testing data.

For deep sleep classification, the large increase of AUC shown in table 5.4 for Random Forest, XGBoost and SVM suggests that the wrist data is valuable for separating deep sleep from the other sleep stages. The F1-score and sensitivity are however poor, as the data is unbalanced. The increased performance of all metrics except specificity for XGBoost reveals that this algorithm in particular is well suited for classification of deep sleep, and that inclusion of wrist data is beneficial. By comparing the sleep disorder data and the healthy sleep data in table 6.2, it is evident that there is some overfitting in several models. This is especially visible in Random Forest, with feature set 2 and 3, achieving F1-scores of 0.27 and 0.21 for the sleep disorder data and healthy sleep data respectively in experiment 2, and 0.19 and 0.12 for the same datasets in experiment 3.

As stated earlier, an AUC of around 0.5 means that the model is not successfully separating the classes. This is the case for REM classification with all tested methods, also when the wrist data is included, as described in table 5.5. A noteworthy impact of the wrist data, however, is the large increase of the F1-score and sensitivity for Random Forest and XGBoost. This tells us that even though the methods are not better at separating REM from non-REM, more REM samples are correctly classified, possibly because more samples are classified as REM in total. The improvement of the F1-score is also clear in the Random Forest results using the healthy subjects data, where the F1-score changes from 0.06 to 0.16. Even with this large improvement of results when the wrist data is included, REM sleep classification is not well performed by any of the machine learning methods in our tests.

Chapter 7

Conclusion and future work

This chapter contains the conclusion of our work, followed by some suggestions for future extensions of our work.

7.1 Conclusion

In this project we have run experiments with machine learning methods on actigraphy data, with various sets of features. For sleep–wake classification our results indicate that including temperature data features does not improve the classification of accelerometer data. There is, however, higher performance when the data from the wrist sensor is used in addition to the back and thigh sensor in the baseline experiment. As the HUNT4 dataset does not have wrist data, we would like to know whether the back and thigh data is sufficient for classification of sleep and wake. The results were better for all three sensors, so it would probably give more accurate results for the HUNT4 data if wrist measurements were available. However, the improvement is small and the baseline results are good for sleep–wake classification, even when tested on a different dataset. For instance the XGBoost classifier achieved an accuracy, F1-score, AUC, sensitivity and specificity of 0.84, 0.90, 0.84, 0.98 and 0.46, respectively, when tested on back and thigh accelerometer data from an unseen dataset.

The effect of temperature and wrist data on the results for light sleep were very similar to general sleep classification, but the results are in general much lower for all metrics. For deep sleep, the ability to differentiate between the classes improves with temperature and wrist data, but the accuracy does not. Additionally, the F1 and sensitivity are low for all experiments with deep sleep, with a maximum of 0.35 for both metrics. REM sleep turned out to be the most difficult classification task, with the lowest F1-score, 0.19 at best.

In general, it is evident that actigraphy data alone, even with temperatures and three sensors, is insufficient for sleep class classification, at least with the machine learning models and features used in the project. However, the results for light sleep is significantly better than the results of deep sleep and REM. Inclusion of wrist data does give slightly better results for light sleep, while deep sleep and REM are not strongly related to the wrist data.

The overall results of the sleep disorder data cross-validation and the tests on the healthy sleep data shows that models trained on data from poor sleepers achieve similar performance on data from healthy sleepers. This should be confirmed by new experiments with other datasets, to ensure that it is not just the case for our two datasets. However, it is promising for classification of the HUNT4 data with a model trained on sleep disorder data.

7.2 Future Work

Based on the work presented in this thesis, several paths can be taken in future work. Some possibilities are presented in this section.

7.2.1 Feature engineering for temperature features

Our conclusions for inclusion of temperature data in all the experiments are dependent on the features we have extracted from the temperature data. We decided to use the same features for the temperature as for the accelerometer data, even though these may not be the optimal features. To explore other feature options the data should be analyzed and the importance of several features should be evaluated. Another consideration is the size of the sliding time window used by the features. As temperature generally changes at a lower rate than the body movements, the use of larger window sizes should be evaluated. More relevant features and window size could result in better results for the classifiers.

7.2.2 Semi-supervised learning

Supervised learning is the only type of machine learning used in our work. For future extensions semi-supervised learning could be implemented and evaluated. This involves using unlabeled data in addition to the labeled data, to improve the performance of the model. As the HUNT4 data is unlabeled, this could be used as a part of the training data for a semi-supervised model. With the amount of data in the HUNT4 dataset, this could lead to a large improvement. However, the machine learning models would be restricted to use features from the back and thigh, as there is no wrist data in the HUNT4 dataset.

7.2.3 Hidden Markov model

From the literature review it is evident that Hidden Markov models are suitable for sleep classification. Despite this, it was not implemented in our work. HMM is suitable for classification tasks where each data point is dependent on the previous data. This is true for sleep data, as the probability of one sample being 'sleep' is higher if it is known that the previous sample was 'sleep'. We have represented the temporal dependencies by using features that summarize the changes over several subsequent data points. However, HMM may give better results, as it explicitly uses the result of one classified data point in the next classification.

7.2.4 Multi-class classification

Even though our sleep stage classification results were not very promising, multi-class classification for light, deep and REM sleep could give interesting results. One benefit of this method is that the imbalance of the data classes will be smaller, which could improve the performance of classification for the minority classes, REM and deep sleep. A consequence of using a model that is able to separate between many classes is the increased complexity compared to a binary classifier.

7.2.5 Training on all data

In this project, the sleep disorder dataset was used for training and the healthy sleep dataset was used for testing. In future projects this setup can be altered, for instance by using leave-one-group-out cross validation on the data of both datasets combined. It could also be investigated whether the models have higher accuracy when they are trained on all data or just one of the datasets when it is tested on the HUNT4 data. This would require a manual evaluation, as the HUNT4 dataset is not labeled.

Bibliography

Banfi, T., Valigi, N., di Galante, M., d'Ascanio, P., Ciuti, G., and Faraguna, U. (2021). Efficient embedded sleep wake classification for open-source actigraphy. *Scientific Reports*, 11(345).

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. Association for Computing Machinery.

Cho, T., Sunarya, U., Yeo, M., Hwang, B., Koo, Y. S., and Park, C. (2019). Deep-actinet: End-to-end deep learning architecture for automatic sleep-wake detection using wrist actigraphy. *Electronics*, 8(12):1461.

Cole, R. J., Kripke, D. F., Gruen, W., Mullaney, D. J., and Gillin, J. C. (1992). Automatic Sleep/Wake Identification From Wrist Activity. *Sleep*, 15(5):461–469.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Dudani, S. A. (1976). The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6(4):325–327.

Faerman, A., Kaplan, K. A., and Zeitzer, J. M. (2020). Subjective sleep quality is poorly associated with actigraphy and heart rate measures in community-dwelling older men. *Sleep Medicine*, 73:154–161.

Fallmann, S., Chen, L., and Chen, F. (2020). Enhanced multi-source data analysis for personalized sleep-wake pattern recognition and sleep parameter extraction. *Personal and Ubiquitous Computing*.

Fedorin, I., Slyusarenko, K., Lee, W., and Sakhnenko, N. (2019). Sleep stages classification in a healthy people based on optical plethysmography and accelerometer signals via wearable devices. In *2019 IEEE 2nd Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, pages 1201–1204.

Ferree, T., Moynihan, M., and Gozani, S. (2019). Applications of machine learning to improve time in bed detection by leg-worn actigraphy. *Sleep*, 42(Supplement_1):A405–A406.

Gryvill, M. (2020). Machine learning methods for sleep-wake classification of accelerometer data. *Specialization project, Norwegian University of Science and Technology, Trondheim, Norway.*

Hay, A. (2019). Machine learning methods for sleep-wake classification using two body-worn accelerometers. *Master Thesis, Norwegian University of Science and Technology, Trondheim, Norway.*

Hillman, D. R. and Lack, L. C. (2013). Public health implications of sleep loss: the community burden. *The Medical journal of Australia*, 199(8):S7–S10.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282.

Hu, J. and Shou, H. (2021). Three-level sleep stage classification based on wrist-worn accelerometry data alone. *bioRxiv.*

Ibáñez, V., Silva, J., and Cauli, O. (2018). A survey on sleep assessment methods. *PeerJ*, 6(e4849).

Khademi, A., El-Manzalawy, Y., Master, L., Buxton, O. M., and Honavar, V. G. (2019). Personalized sleep parameters estimation from actigraphy: A machine learning approach. *Nature and Science of Sleep*, 11:387–399.

Li, L. and Nakamura, T. (2019). An epidemiological sleep study based on a large-scale physical activity database. *IEEE 1st Global Conference on Life Sciences and Technologies (LifeTech)*, pages 292–293.

Li, X., Zhang, Y., Jiang, F., and Zhao, H. (2020). A novel machine learning unsupervised algorithm for sleep/wake identification using actigraphy. *Chronobiology International*, 37(7):1002–1015.

Liu, J., Zhao, Y., Lai, B., Wang, H., and Tsui, K. L. (2020). Wearable device heart rate and activity data in an unsupervised approach to personalized sleep monitoring: Algorithm validation. *JMIR Mhealth Uhealth*, 8(8).

Lüdtke, S., Hermann, W., Kirste, T., Beneš, H., and Teipel, S. (2021). An algorithm for actigraphy-based sleep/wake scoring: Comparison with polysomnography. *Clinical Neurophysiology*, 132(1):137–145.

Palotti, J., Mall, R., Aupetit, M., Rueschman, M., Singh, M., Sathyanarayana, A., Taheri, S., and Fernández-Luque, L. (2019). Benchmark on a large cohort for sleep-wake classification with machine learning techniques. *NPJ Digital Medicine*, 2.

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, 1:81–106.

Reimer, M. A. and Flemons, W. W. (2003). Quality of life in sleep disorders. *Sleep Medicine Reviews*, 7(4):287–365.

Sadeh, A., Sharkey, M., and Carskadon, M. A. (1994). Activity-Based Sleep-Wake Identification: An Empirical Test of Methodological Issues. *Sleep*, 17(3):201–207.

Sano, A., Chen, W., Lopez-Martinez, D., Taylor, S., and Picard, R. W. (2019). Multimodal ambulatory sleep detection using lstm recurrent neural networks. *IEEE Journal of Biomedical and Health Informatics*, 23(4):1607–1617.

Sundararajan, K., Georgievska, S., te Lindert, B. H. W., Gehrman, P. R., Ramautar, J., Mazzotti, D. R., Sabia, S., Weedon, M. N., van Someren, E. J. W., Ridder, L., Wang, J., and van Hees, V. T. (2021). Sleep classification from wrist-worn accelerometer data using random forests. *Scientific Reports*, 11(24).

Walch, O., Huang, Y., Forger, D., and Goldstein, C. (2019). Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. *Sleep*, 42(12).

Willetts, M., Hollowell, S., Aslett, L., Holmes, C., and Doherty, A. (2018). Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 uk biobank participants. *Scientific Reports*, 8:7961.

Yildiz, S., Opel, R. A., Elliott, J. E., Kaye, J., Cao, H., and Lim, M. M. (2019). Categorizing sleep in older adults with wireless activity monitors using lstm neural networks. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3368–3372.

Appendix A

Training dataset summary

The following tables, A.1 and A.2, contain a summary of the sleep disorder dataset and the healthy sleep dataset, respectively.

Sleep disorder data

Table A.1: Summary of sleep disorder data for each subject. *Samples* are number of 100 Hz samples from the raw data, *windows* are number of time windows in the extracted feature dataset.

Subject	Samples	Windows
06	3,628,000	1184
14	3,546,000	1157
18	3,577,000	1167
19	3,576,000	1167
20	3,780,000	1235
24	3,678,000	1201
25	3,959,000	1294
29	3,000,000	975
30	3,481,000	1135
35	3,619,000	1181
37	3,099,000	1008

Healthy sleep data

Table A.2: Summary of healthy sleep data for each subject. *Samples* are number of 100 Hz samples from the raw data, *windows* are number of time windows in the extracted feature dataset.

Subject	Samples	Windows
01	3,777,000	1234
02	2,988,000	971
05	3,723,000	1216
06	3,420,000	1115
07	3,705,000	1210
08	3,420,000	1115
09	3,531,000	1152
10	3,237,000	1054
11	3,159,000	1028
12	2,964,000	963
15	2,742,000	889
16	3,636,000	1187
17	3,474,000	1133
22	3,696,000	1207
24	2,970,000	965
27	2,952,000	959
28	3,396,000	1107

Appendix B

Final values of hyperparameters

This chapter contains the values of the hyperparameters for each classifier after tuning. The parameters for sleep–wake classification are presented in table B.1. The values for light sleep, deep sleep and REM sleep are presented in table B.2, B.3 and B.4, respectively. Experiment 1, 2 and 3 are described in table 5.1.

Table B.1: Final values of hyperparameters for sleep–wake classification. RF = Random Forest, XGB = XGBoost.

Method	Hyperparameter	Experiment 1	Experiment 2	Experiment 3
RF	n_estimators	400	250	100
	max_depth	10	10	50
	min_samples_leaf	1	1	3
	min_samples_split	2	7	2
XGB	n_estimators	100	200	200
	colsample_bytree	0.3	1	0.5
	gamma	0.3	1	0.005
	learning_rate	0.25	0.2	0.05
	max_depth	4	4	2
	min_child_weight	2	4	4
K-NN	n_neighbors	200	110	450
	p	1	1	1
	weights	'distance'	'uniform'	'uniform'
SVM	C	0.001	0.001	0.001
	gamma	'scale'	'scale'	'scale'
	kernel	'linear'	'linear'	'linear'

Table B.2: Final values of hyperparameters for light sleep classification. RF = Random Forest, XGB = XGBoost.

Method	Hyperparameter	Experiment 1	Experiment 2	Experiment 3
RF	n_estimators	500	200	75
	max_depth	40	5	15
	min_samples_leaf	2	2	1
	min_samples_split	5	5	7
XGB	n_estimators	150	100	300
	colsample_bytree	1	0.5	0.7
	gamma	0.15	0.3	10
	learning_rate	0.4	0.1	0.5
	max_depth	10	8	4
	min_child_weight	6	8	2
K-NN	n_neighbors	2	31	500
	p	2	1	2
	weights	'distance'	'uniform'	'uniform'
SVM	C	10	0.01	0.01
	gamma	'auto'	'scale'	'auto'
	kernel	'rbf'	'rbf'	'rbf'

Table B.3: Final values of hyperparameters for deep sleep classification. RF = Random Forest, XGB = XGBoost.

Method	Hyperparameter	Experiment 1	Experiment 2	Experiment 3
RF	n_estimators	250	25	225
	max_depth	40	50	30
	min_samples_leaf	2	3	4
	min_samples_split	7	2	20
XGB	n_estimators	200	200	200
	colsample_bytree	0.5	0.1	0.7
	gamma	0.15	0.0	0.3
	learning_rate	0.6	0.5	0.5
	max_depth	6	8	4
	min_child_weight	6	4	3
K-NN	n_neighbors	9	9	3
	p	2	1	2
	weights	'distance'	'uniform'	'uniform'
SVM	C	30	20	55
	gamma	'auto'	'auto'	'auto'
	kernel	'rbf'	'rbf'	'rbf'

Table B.4: Final values of hyperparameters for REM sleep classification. RF = Random Forest, XGB = XGBoost.

Method	Hyperparameter	Experiment 1	Experiment 2	Experiment 3
RF	n_estimators	150	50	5
	max_depth	25	20	30
	min_samples_leaf	2	1	1
	min_samples_split	2	5	6
XGB	n_estimators	200	250	150
	colsample_bytree	0.5	0.5	0.5
	gamma	10	100	0.5
	learning_rate	0.35	0.35	0.6
	max_depth	4	6	2
	min_child_weight	4	2	1
K-NN	n_neighbors	2	1	1
	p	1	1	2
	weights	'distance'	'uniform'	'uniform'
SVM	C	10000	10000	40
	gamma	'auto'	'auto'	'auto'
	kernel	'rbf'	'rbf'	'rbf'

NTNU

Norwegian University of
Science and Technology