DEGREE PROJECT IN TECHNOLOGY,
FIRST CYCLE, 15 CREDITS
STOCKHOLM, SWEDEN 2019

Evaluating machine learning methods for detecting sleep arousal

KTH Bachelor Thesis Report

Anton Ivarsson & Jacob Stachowicz

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

Abstract

Sleep arousal is a phenomenon that affects the sleep of a large amount of people. The process of predicting and classifying arousal events is done manually with the aid of certified technologists, although some research has been done on automation using Artificial Neural Networks (ANN). This study explored how a Support Vector Machine performed(SVM) compared to an ANN on this task. Polysomnography (PSG) is a sort of sleep study which produces the data that is used in classifying sleep disorders. The PSG-data used in this thesis consists of 13 wave forms sampled at or resampled at 200Hz. There were samples from 994 patients totalling approximately $6.98 \cdot 10^{10}$ data points, processing this amount of data is time consuming and presents a challenge. 2000 points of each signal was used in the construction of the data set used for the models. Extracted features included: Median, Max, Min, Skewness, Kurtosis, Power of EEG-band frequencies and more. Recursive feature elimination was used in order to select the best amount of extracted features. The extracted data set was used to train two ”out of the box” classifiers and due to memory issues the testing had to be split in four batches. When taking the mean of the four tests, the SVM scored ROC AUC of 0,575 and the ANN 0.569 respectively. As the difference in the two results was very modest it was not possible to conclude that either model was better suited for the task at hand. It could however be concluded that SVM *can* perform as well as ANN on PSG-data. More work has to bee done on feature extraction, feature selection and the tuning of the models for PSG-data to conclude anything else. Future thesis work could include research questions as ”Which features performs best for a SVM in the prediction of Sleep arousals on PSG-data” or ”What feature selection technique performs best for a SVM in the prediction of Sleep arousals on PSG-data”, etc.

Keywords

Machine learning, Sleep Arousal, Polysomnography, Automatic Detection, Bachelors Thesis, Big Data

Evaluering av maskininlärningsmetoder för detektion av sömnstörningar

Abstract

Sömnstörningar är en samling hälsotillstånd som påverkar sömnkvaliteten hos en stor mängd människor. Ett exempel på en sömnstörning är sömnapne. Detektion av dessa händelser är idag en manuell uppgift utförd av certifierade teknologer, det har dock på senare tid gjorts studier som visar att Artificella Neurala Nätverk (ANN) klarar att detektera händelserna med stor träffsäkerhet. Denna studie undersöker hur väl en Support Vector Machine (SVM) kan detektera dessa händelser jämfört med en ANN. Datat som används för att klassificera sömnstörningar kommer från en typ av sömnstudie kallad polysomnografi (PSG). Den PSG-data som används i denna avhandling består av 13 vågformer där 12 spelats in i 200Hz och en rekonstruerats till 200Hz. Datan som används i denna avhandling innehåller inspelningar från 994 patienter, vilket ger totalt ungefär $6.98 \cdot 10^{10}$ datapunkter. Att behandla en så stor mängd data var en utmaning. 2000 punkter från vare vågform användes vid konstruktionen av det dataset som användes för modellerna. De attribut som extraherades innehöll bland annat: Median, Max, Min, Skewness, Kurtosis, amplitud av EEG-bandfrekvenser m.m. Metoden *Recursive Feature Elimination* användes för att välja den optimala antalet av de bästa attributen. Det extraherade datasetet användes sedan för att träna två standard-konfigurerade modeller, en SVM och en ANN. På grund av en begränings av arbetsminne så var vi tvungna att dela upp träningen och testandet i fyra segment. Medelvärdet av de fyra testen blev en ROC AUC på 0,575 för en SVM, respektive 0,569 för ANN. Eftersom skillnaden i de två resultaten var väldigt marginella kunde vi inte dra slutsatsen att endera modellen var bättre lämpad för uppgiften till hands. Vi kan dock dra slutsatsen att en SVM *kan* prestera lika väl som ANN på PSG-data utan konfiguration. Mer arbete krävs inom extraheringen av attributen, attribut-eliminationen och justering av modellerna. Framtida avhandlingar skulle kunna göras med frågeställningarna: “Vilka attributer fungerar bäst för en SVM inom detektionen av sömnstörningar på PSG-data” eller ”Vilken teknik för attribut-elimination fungerar bäst för en SVM inom detektionen av sömnstörningar på PSG-data”, med mera.

Nyckelord

Maskininlärning,
Sömnstörningar, Polysomnografi, Automatisk Detektion, Kandidatuppsats, Big Data

Acknowledgements

We are profoundly grateful to our supervisor Arvind Kumar. Without his aid and support this project would not have been possible.

Authors

Anton Ivarsson and Jacob Stachowicz
School of Electrical Engineering and Computer Science
KTH Royal Institute of Technology

Place for Project

Stockholm, Sweden

Examiner

Örjan Ekeberg Stockholm, Sweden
KTH Royal Institute of Technology

Supervisor

Arvind Kumar Stockholm, Sweden
KTH Royal Institute of Technology

Contents

1 Introduction

Getting enough sleep is a vital part of maintaining health and preventing disease. Inadequate sleep can lead to a variety of negative outcomes including impaired memory and learning ability, obesity, irritability, cardiovascular dysfunction, hypotension, diminished immune function [19] and depression[22].

A phenomenon which can cause poor sleep quality is *sleep arousal* and is a physiological and psychological state that can be loosely described as a disturbance of sleep. These arousals do not always force the affected to a point of perception, they can trigger a shift from deep sleep to lighter sleep. In lighter sleep stages a patient is more likely to wake up by other disturbances [23]. There are many causes of sleep arousals, one of which is lack of oxygen in the brain due to sleep apnea. Sleep apnea is a sleep disorder characterized by pauses in breathing or shallow breathing during sleep.

Evaluating abnormalities like sleep arousals are done with data from a type of sleep study called Polysomnography(PSG). PSG monitors a range of signals derived from a subject, including electroencephalography (EEG), electrooculography (EOG), and surface electromyography (EMG)[1].

PhysioNet is a research resource for complex physiological signals and is supported by the National Institute of General Medical Sciences (NIGMS) and the National Institute of Biomedical Imaging and Bioengineering (NIBIB) under NIH grant number 2R01GM104987-09[24]. This thesis is based on PSG-data from a challenge from PhysioNet, namely: ”You Snooze, You Win: the PhysioNet/Computing in Cardiology Challenge 2018”.

This thesis entails using PSG-data predict arousal events using two different machine learning methods and evaluating model performance. The process of predicting and classifying arousal events is done manually and is therefore time-consuming. A method of performing automatic prediction an classification would be a valuable tool for professionals in the field.

1.1 Problem Statement

As will described later, there has been work done on the automation of detecting sleep arousals. Most of this work has been done with some form of artificial neural network (ANN). There has been some success on the usage of support vector machines (SVM) for similar tasks but not on this exact type of PSG-data. Seeing that there were no comparison between ANN and SVM on PSG-data, we found our research question:

“Can a support vector machine perform better than an artificial neural network on Polysomnography data?”

The purpose of this thesis is determining which of the selected machine learning models performs better on PSG-data for detecting sleep arousals. A program which can classify target regions automatically on large amounts of data will be delivered.

The available data contains 13 different types of arousals however the problem is limited to binary classification. This thesis will evaluate a closed set of classification models, ANN and SVM. Both of these methods will be described in detail later.

2 Theoretical Background

The theoretical background necessary for understanding this report is presented here.

2.1 Data set

Data for the PhysioNet challenge was contributed by the Massachusetts General Hospital's (MGH) Computational Clinical Neurophysiology Laboratory (CCNL), and the Clinical Data Animation Laboratory (CDAC). The data set includes 1985 subjects which were monitored at an MGH sleep laboratory for the diagnosis of sleep disorders. It is important to note that not all subject had sleep problems like sleep apnea. The data taken from this challenge was the labeled partition which contained records from 994 subjects [25]. 13 different physiological signal readings were recorded from each of the subjects during their stay in the laboratory, these signals will from now on also be referred to as wave forms. Table 1 describes the wave forms.

Signal Name	Units	Signal Description
SaO_2	%	Oxygen Saturation
ABD	μV	Electromyography, a measurement of abdominal movement
CHEST	μV	Electromyography, measure of chest movement
Chin1-Chin2	μV	Electromyography, a measure of chin movement
AIRFLOW	μV	A measure of respiratory airflow
ECG	mV	Electrocardiogram, a measure of cardiac activity
E1-M2	μV	Electrooculography, a measure of left eye activity
O2-M1	μV	Electroencephalography, a measure of posterior activity
C4-M1	μV	Electroencephalography, a measure of central brain activity
C3-M2	μV	Electroencephalography, a measure of central brain activity
F3-M2	μV	Electroencephalography, a measure of frontal brain activity
F4-M1	μV	Electroencephalography, a measure of frontal brain activity
O1-M2	μV	Electroencephalography, a measure of posterior brain activity

The subjects sleep in the laboratory lasted approximately 7-8 hours, all wave forms except SaO_2 was sampled at 200 Hz, however SaO_2 was re-sampled to 200Hz. Consequently the data for each patient consists of approximately $5.4 \cdot 10^6$ samples per wave form. Figure 1 is an example plot of the wave forms before processing, for an arbitrary point and patient. The total amount of samples in the labeled set is approximately $6.98 \cdot 10^{10}$.

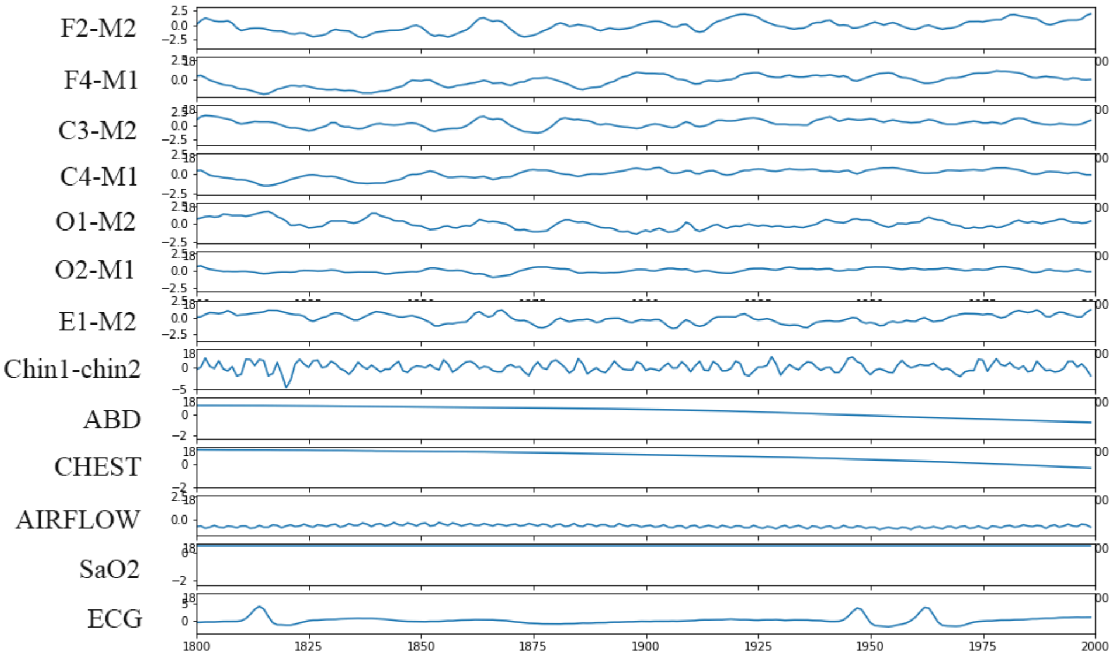

Figure 1: Example plot of wave forms.

Certified sleep technologists at MGH annotated the wave forms with the presence of arousals that interrupted the sleep of the subject. The annotated arousals were classified as either: spontaneous arousals, RERA, bruxisms, hypoventilations, hypopneas, apneas (central, obstructive and mixed), vocalizations, snores, periodic leg movements, Cheyne-Stokes breathing or partial airway obstructions.

2.2 Preprocessing Techniques

2.2.1 Normalization

Normalization of data is when the scale of all data points is transformed to fit between a certain interval [8]. The interval used in this project is between -1 and 1. The normalization technique used in this project is called *standard score*, it is calculated by subtracting the mean of the data from the data point in focus and dividing the result with the standard deviation of the data. Equation 1 shows the mathematical notation for the standard score, where X is the observed data point, X_N the normalized data point, μ the mean value of all data points and σ the standard deviation for all of the data points.

$$X_N = \frac{X - \mu}{\sigma} \tag{1}$$

Normalization is done so the different data intervals becomes comparable for the classifiers. The process of normalization puts all signal intervals on the same scale.

2.3 Description Of Features and related Terms

Here follows the descriptions of the non trivial statistical techniques used in this project. The trivial terms are mean, median, max, min, standard deviation and variance of the data.

2.3.1 Skewness

The Skewness of a probabilistic distribution is a statistical measure that shows how asymmetric the distribution for a real valued stochastic variable is [15]. The skewness is zero if a normal distribution is symmetric around the mean. The skewness is negative if the distribution has a longer ”tail” to the left than to its right, similarly the skewness is positive if the tail to the right is longer.

2.3.2 Kurtosis

Kurtosis is a measure of the probability for the more extreme outcomes for a stochastic variable in a given probability distribution, more exactly kurtosis is a description of the shape or size of the tail in a distribution [28]. Kurtosis is similar to skewness in the regard that it describes the shape of the distribution.

2.3.3 Covariance Matrix

A covariance matrix is a symmetrical and square matrix that shows how different data samples correlate with each other.

Covariance in statistics and probability is theory a measure of the correlation between two stochastic variables. The covariance is positive if the larger values in a stochastic variable X correlate with the larger values in a stochastic variable Y, and if a similar pattern exists for the smaller values for X and Y [29].

Equation 2 shows the mathematical definition of covariance between two real-valued stochastic variables X and Y. Where $E(X) = \mu$ and $E(Y) = \nu$ represents the *expected values* of X respectively Y.

$$Cov(X, Y) = E\big[(X - \mu)(Y - \nu)\big] \tag{2}$$

The row positions and the column positions in the covariance matrix represent the same sequence of data samples, or stochastic variables. In a normalized covariance matrix, with the scale zero to one. One in the position [i, j] means that the data points j and i highly correlate with each other and zero means that the data points are not dependant upon each other at all.

Figure 2 shows a covariance matrix with dimension 13 x 13 matrix. Each row and column represent different data samples.

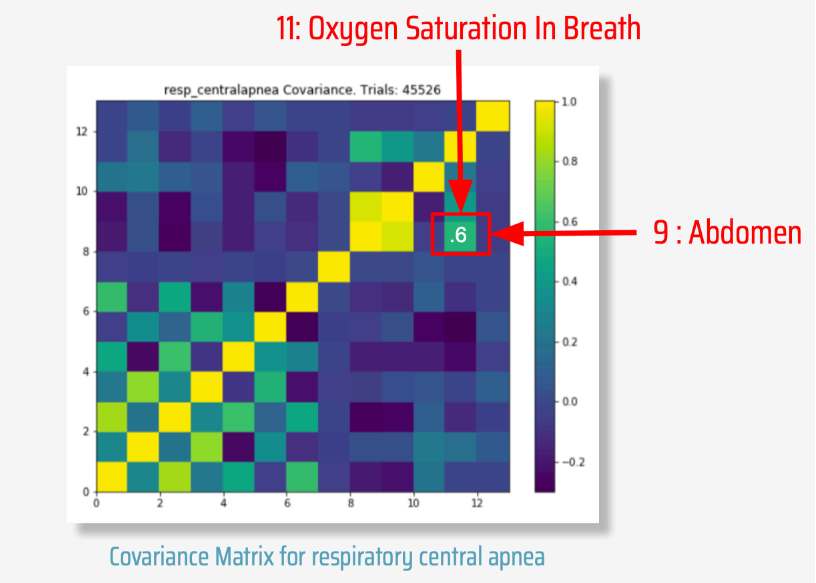

Figure 2: A covariance matrix that indicates that data point 9 and 11 have a correlation of 0.6 with each other.

2.3.4 Fast Fourier Transform

The Fourier transform (FT) is an technique for decomposing an a-periodic signal into several underlying sinus waves of different frequencies. FT extracts the amplitudes of different frequency waves hidden in a irregular signal. The amplitudes can thereafter be illustrated in a graph, where the y-axis is the power of the amplitude and the x-axis represents the frequency in Hz. The maximum amount of different frequencies are always equal to the amount of data points in the recorded signal. The Fast Fourier Transform (FFT) algorithm is an improvement of FT in the regard that it makes the calculation of FT faster and is therefore more suited for modern computing. Figure 3 is an illustration of how the FT works.

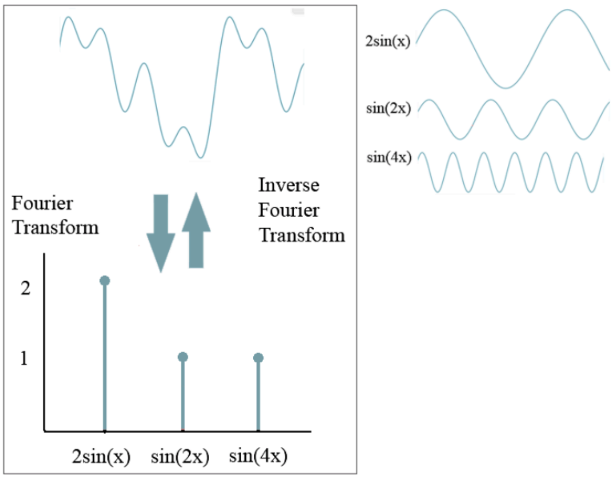

Figure 3: Fourier Transform.

2.4 Recursive Feature Elimination

Recursive feature elimination (RFE) is a technique for selecting the best X-number of features, where X is a number selected by the user. The algorithm selects a number of features by recursively considering smaller and smaller sets of features [7]. A cost function J is calculated for training samples, J is usually bound or an approximation of the ideal objective. An estimator that assigns weights(e.g., the coefficients of a linear model) to features is used, weight w_i is assigned to feature i. The weight represents the significance of that feature. RFE follows these three steps:

- The estimator is trained by optimizing the weights w_i with respect to J.
- The ranking criterion w_i^2 is calculated for each feature.
- The feature with the lowest ranking criterion is removed.

This process is repeated until the desired number of features is acquired. [13]

2.5 Evaluation Techniques

In this project one evaluation technique was used, namely *Area Under Curve Receiver Operating Characteristic.*

Receiver operating characteristic (ROC) is a statistical graph that shows how well a binary classifier is at distinguishing two classes, under all thresholds of classification [21]. ROC is a curve that is determined by the accuracy of the true positive and true negative classifications. A true positive is when a model classifies class 1 correctly and a false negative is when a model classifies class 0 incorrectly [12].

The x-axis represents the false positive rate (FPR) and the y-axis represents the true positive rate (TPR). Equation 3 and 4 shows the mathematical notations for TPR and FPR. Where TN and FP represents the amount of true positives and false positives respectively.

$$TPR = \frac{TN}{TN + FP} \tag{3}$$

$$FPR = \frac{FP}{FP + TN} \tag{4}$$

Area under curve (AUC) is the two-dimensional area under the ROC-curve and is a sum of all results for the different classification thresholds. The performance scale goes between 0 and 1. Figure 4 illustrates an example of the AUC-ROC graph.

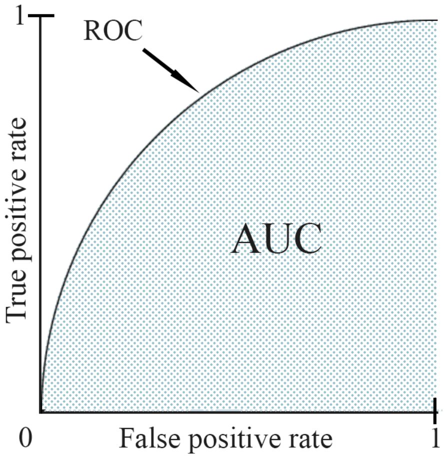

Figure 4: The AUC-ROC is used in classification problems as a measurement of how good the the chosen model is at classifying the different classes.

2.6 Machine Learning Models

There were two different machine learning methods used in this project; Support Vector Machine and Artificial Neural network. These chosen models are described below.

2.6.1 Support Vector Machine

Support Vector Machine (SVM) is a machine learning technique that is used for binary discriminative classification[5]. The SVM uses a decision boundary to separate data points between classes. The decision boundary is a hyper plane defined by an algebraic formula. A straight line dividing two classes on a two dimensional surface space is an example of a decision boundary of a polynomial algebraic formula. The black curve in figure 5 is a polynomial decision boundary of degree 2.

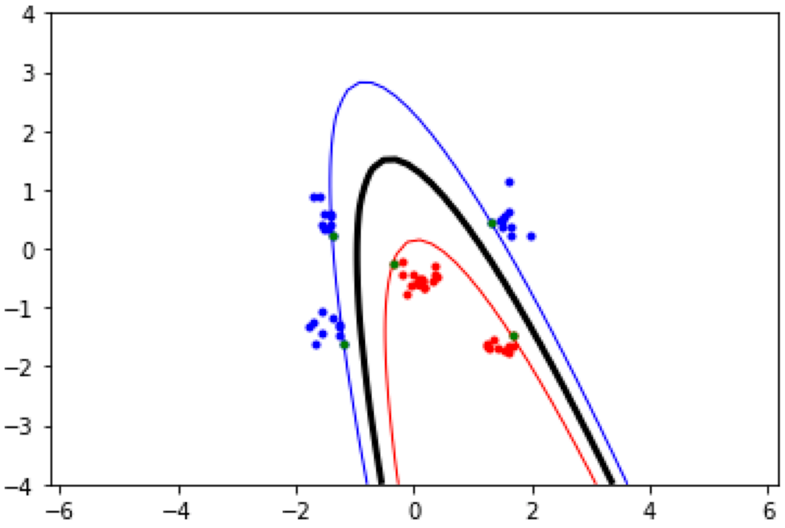

Figure 5: A SVM with a polynomial decision boundary represented on a two-dimensional feature space.

The building blocks of a SVM are:

- *Kernels*
  The algebraic type for a hyper plane is defined by its kernel. A SVM has one of the following three types of kernels: polynomial, radial basis function or sigmoid [9].
- *Gamma-values*
  The gamma value is a measure of how precisely the the model should be fitted to the training data. A higher gamma-value forces the model to fit the training data more accurately, this not always good due to the bias-variance

trade-off[9]. If a model is highly fitted to the training data it tends to become over-fitted and gains a high variance [26]. A over-fitted model is too specific to the training data and will perform worse at classifying the test-data. If a gamma value is too small it leads to a larger bias, which means that the model assumes a formula for the hyper plane which may not represent the optimal class boundary.

- *C-values*
  The C-value is a penalty term and a tuning parameter. The C-value determines the distance from the decision boundary to the margins and sets a threshold for allowed miss-classifications [9]. The margins in figure 5 are represented by the red and blue curves. Each data point that is positioned outside the class margin is assigned a penalty value. The C-value decides how large the sum of all penalty value are allowed to become.

- *Degree*
  This parameter only applies to a SVM with a polynomial kernel and describes what degree the polynomial equation for the decision boundary has.

2.6.2 Artificial Neural Network

Artificial Neural network (ANN) is a group name for several machine learning methods which all use neural networks for binary and non-binary classification. The neural network is made up by a large amount of connections between different neurons which contain different weighted values. The ANN model is used when a non-linear correlation exists between the input vector and the output. The training process of the model recognizes the intricate relationships between input and output [2]. Neurons are the processing units of an ANN. The inspiration behind the ANN is the biological neural networks in the brain of living animals and humans, even though they differ in how they work. An ANN consists of an input layer, an output layer and one or several hidden layers, see figure 6. The input vector consists of known values from variables such as features [10]. The output vector contains measures of probability for each classification class. The hidden layers contain neurons, and are the engine of the ANN.

A feed-forward artificial neural network is considered the simplest and the most commonly used ANN. A feed-forward ANN does not form cyclic relationships in its network. If the classification result is not accurate enough in a feed-forward ANN, then optimization by tweaking values in the neurons is implemented by a iterative or recursive technique called back propagation[11]. The ANN used in this project is a feed-forward ANN.

The amount of neurons i each of the hidden layers are decided through a variety of methods, e.g experimentation, intuition [4]. Each neuron contains an activation function that determines an output given a specific input [6]. The continuous scale of the output calculated by the activation function ranges between a predefined spectrum, usually between 0 and 1 or -1 and 1. The activation function has similarities with the SVM kernel in the regard that both are fundamentally a type of algebraic formula that calculates the output, for example both the SVM and ANN commonly use the radial basis function or the sigmoid function [17].

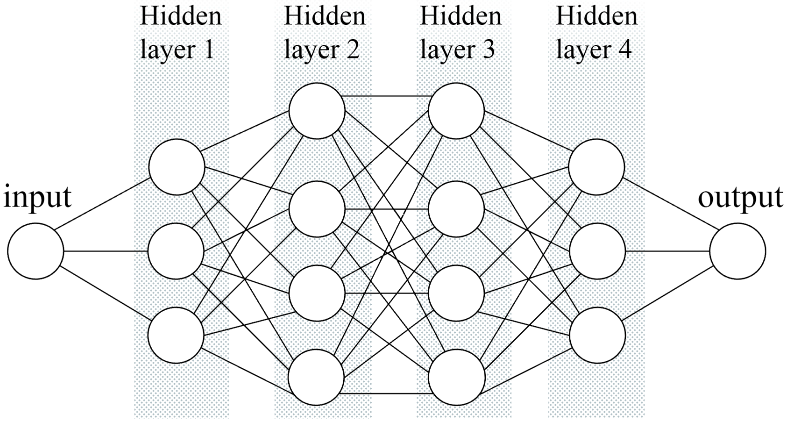

Figure 6: An illustration of the neurons in an ANN with its interconnections.

3 Literature Overview

Detecting and classifying sleep stages and sleep arousals with PSG data largely maintain a manual task even though there has been some work done on the subject of automation. Prior to the PhysioNet challenge mentioned in the introduction, the research mostly focus on arousals related to Sleep Apnea. Biswal et al. claim to have an algorithm that performs equal to expert sleep technologists for classifying sleep stages, sleep apnea and limb movement [3]. Yildirim et al. showed similar results on the classification of sleep stages using PSG-data[30]. Both of these publications do show that machine learning techniques, specifically deep learning, can be effectively applied to PSG-data but do not consider non-apneic sleep arousal events. Unlike research prior to the challenge, the publications from the challenge focus on both apneic and non-apneic events. Even though most of the previous work with machine learning, PSG-data and PSG adjacent data has been done with some form of ANN, there has been some success using SVM:s. For instance, the 2011 work of Singla et al. showed with a comparative study that SVM can outperform ANN for classifying eye events using one EEG signal namely FP1-F3 with three features: kurtosis coefficient, maximum amplitude and minimum amplitude for a five second window[27].

Khandoker et al. conducted a similar comparative study on the classification on apneic arousals on EEG and EMG data. This study showed that the ANN performed better than the SVM, using three signals: EEG: C2/A2 and C4/A1 central derivations and submental EMG. The researchers extracted 40 features including power of EEG-band frequencies, maximum and minimum value for a three-second interval [18].

It is unclear how the accuracy was measured in the papers by both Khandoker et al. and Singla et al in that there are no mention of the rate of correct true positives compared to rate of true negatives. This in conjunction with the problems being quite different makes them hard to compare, but it is clear that a SVM can outperform an ANN in certain situations and vice versa. It is also worth mentioning that as both of these studies did not work on complete PSG-data, as PSG-data generally contains more signals than EEG and EMG; It is still unclear how the SVM generally performs on complete PSG-data.

As of the 23rd of February there have been 23 publications based on the data from the PhysioNet challenge. A majority of the publications used some form of ANN in the classifications of the arousals, with many of them achieving high AUC ROC scores. None of the 23 publications based on the from the PhysioNet challenge used a SVM.

Howe-Patterson et al. was declared the winner of the PhysioNet challenge and had achieved an AUC ROC score of 0.931 using a using a Dense Convolutional Neural Network[14].

Seeing that there were no comparison between ANN and SVM on PSG-data, it could be concluded that such a comparison would be interesting.

4 Method

This chapter will start with information about the tools used and a description of how the models were trained.

4.1 Tools And Environment

The programming language used in this project was python, the tests were run on a desktop with 16GB of RAM. Libraries used:

- Machine learning, evaluation metrics and signal processing: scikit-learn(sklearn).
- Computations: Numpy.
- Reading .mat and .mat7 matlab files: WFDB.
- Memory analyzing: Pympler.

4.2 Extracting features

A single point in the raw PSG-data does not give much information without adjacent points. Therefore, the features wwere calculated from intervals of 10 seconds. That is, the features was calculated from a matrix with the dimensions $13 * 2000$.

4.3 Training The Models

The extracted data was split in a training set and a test set at, 70% and 30% respectively. The ratio between the two class instances, sleep-arousal and not sleep arousal, was about 1:22. This kind of class imbalance often hinder the predictive power of machine learning models [16]. Having made the observation that our data was heavily imbalanced, a balancing algorithm was implemented to the part of the data that was used for training. The method used is called *under sampling* which essentially means removing a large portion of the instances of one class. The algorithm removed instances of the non arousal class resulting in a ratio of 1:2. Figure 7 illustrates the result of the balancing.

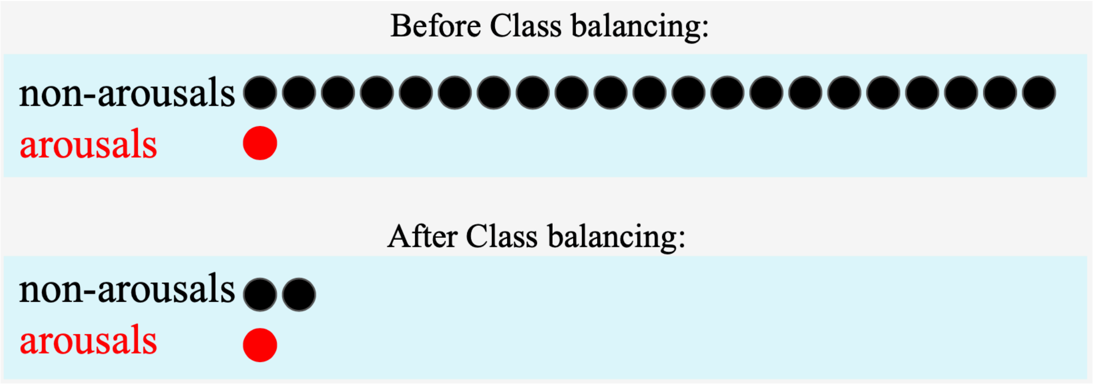

Figure 7: Before and after class balancing.

4.4 Model Parameters

The parameters used in training the classifiers are presented here.

4.4.1 Parameters For The Artificial Neural Network

- Optimizer: Limited-memory BFGS.
- Alpha: $1 * 10^{-5}$.
- Neurons in first hidden layer: 5.
- Hidden Layers: 2.

4.4.2 Parameters For The Support Vector Machine

- Kernel: Radial Basis Function.
- C: 1.

4.5 Evaluating The Models

As classes in the test data were imbalanced, looking at the percentage of correctly classified samples would be misleading. Classifying all the non arousal samples correctly and all the arousal samples incorrectly would approximately yield an accuracy of 21/22 %, purely looking at percentage; As a consequence used ROC AUC to measure accuracy. Non-incremental models were used, meaning that they had to be trained on all the data at once. Running the first tests on all the samples there were significant performance issues, therefore an analysis of the memory used in our program was run.

Analyzing memory on the training data it was noted that:

- Data point(vector of features) was 15280 bytes.
- Class of a data point was 8 bytes.

Equation 5 shows the total amount of data points extracted, assuming an average stay in the lab of 7.5 hours. Equation 6 shows the amount of total memory needed to load entire data set.

$$samples = (994 * 7.5 * 3600 * 200)/2000 = 2683800 \tag{5}$$

$$size = samples * 15280 + samples * 8 {\approx} 4.103 * 10^{10} bytes {\approx} 41GB \tag{6}$$

Consequently to use all the available data, both training and evaluation had to be split up. The data was split in four parts. The models were trained and tested on each of these partitions.

5 Result

This chapter will describe the work done, lastly a results table will presented. Figure 8 shows an overview of the work done.

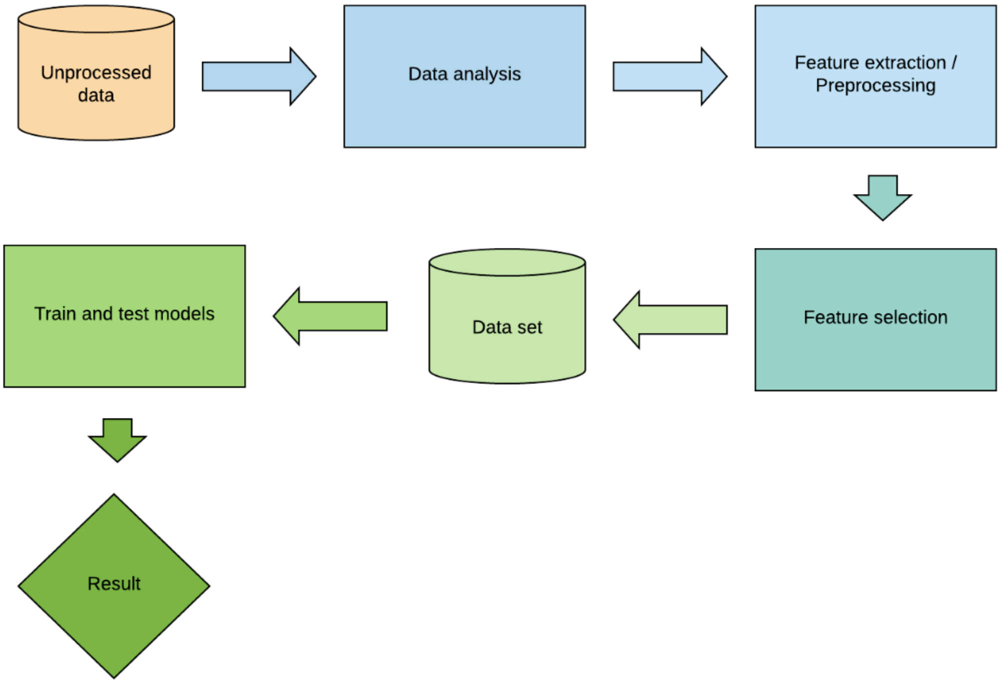

Figure 8: Project overview.

5.1 Data Analysis

For each of the 13 types of arousals average wave forms were calculated for intervals 10 seconds before the arousal up the the recorded time of the arousal. Figure 9 shows an example waveform for an RERA arousal of type Central Apnea the last two seconds of the interval. An average random waveform was also calculated, to serve as a baseline for some comparisons. The average random waveform is an average waveform of randomly selected positions.

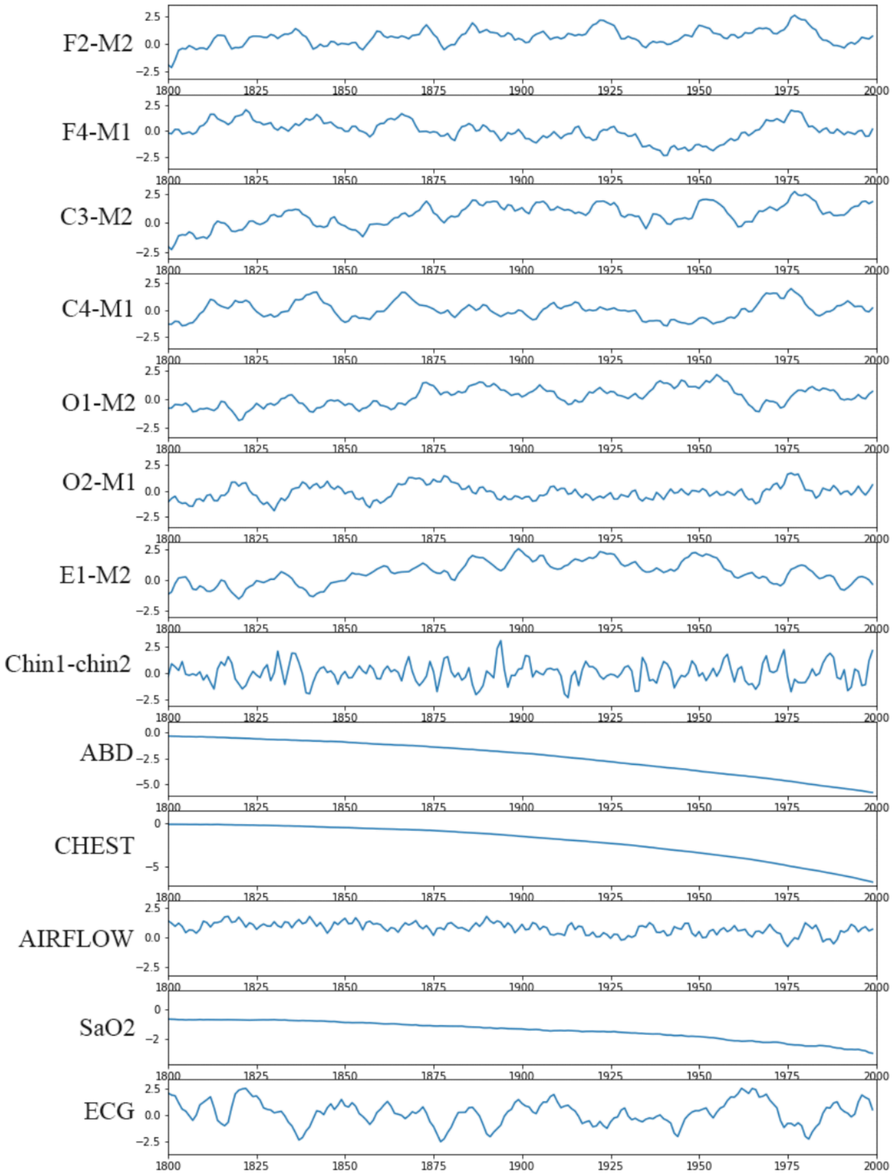

Figure 9: Average waveform: Arousal - RERA Central Apnea.

For these average wave forms the following was done:

- Normalize each signal and calculate covariance matrix. See figure 10.
- Calculate the Fourier transform.
- Normalize the Fourier transform and calculate covariance matrix. See figure 11.

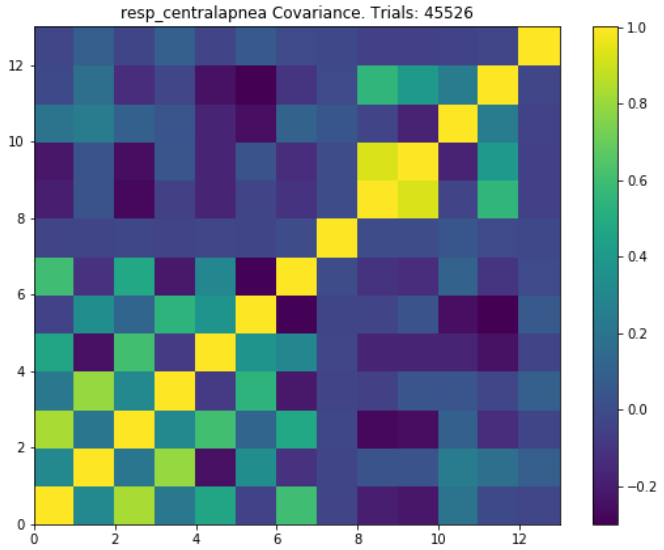

Figure 10: Average Covariance: Arousal - RERA Central Apnea.

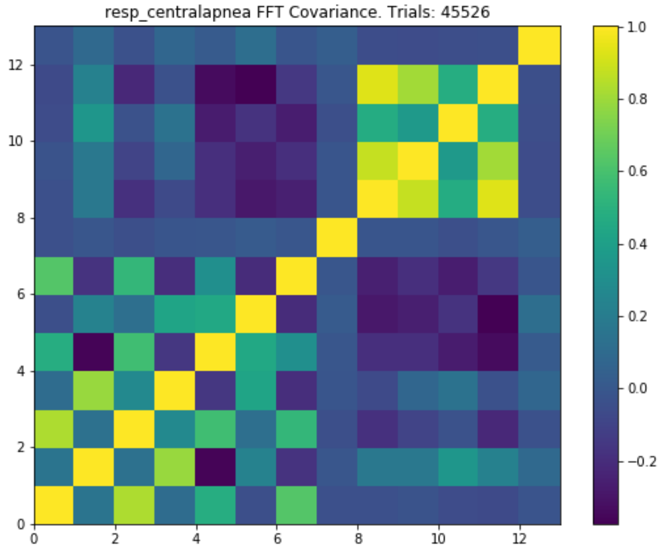

Figure 11: Average Fourier transform Covariance: Arousal - RERA Central Apnea.

To examine how much statistical value these structures had, they were compared that of each arousal type to the same structures calculated for a random positions in the data. Figure 12 shows the difference of the average random covariance matrix to that of an example arousal, bright means that it differs from the average random covariance. As these covariance matrices showed significant differences it could be concluded that these structures had value and should be used in the feature extraction.

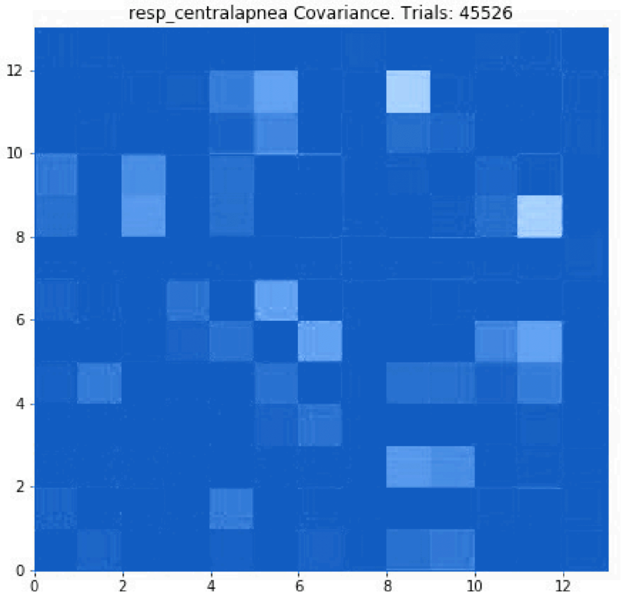

Figure 12: Average covariance with average random covariance overlay: Arousal - RERA Central Apnea.

5.2 Feature Extraction and Preprocessing

Table 1 shows extracted feature categories and the amount of features in each category. For all features except the Fourier transform features the interval was normalized before constructing the data point.

Table 1: Extracted features

Category	Count
Statistical features	325
Covariance matrix distances	13
Points of interest	13
Power of EEG-band frequencies	65
Total	416

5.2.1 Statistical Features

The statistical features used for each signal were the mean, median, max, min, standard deviation, variance, skewness and kurtosis. The correlation between between the first half of a signal with the later half was also tested. A mean of the Fourier transform was also used.

5.2.2 Covariance Matrix Distance

A covariance matrix was calculated for each interval, this matrix was compared to each covariance matrix of the average waveform before each type of arousal. This comparison produced a numerical value for each different type of arousal that represents a similarity to the covariance matrix to each type of arousal; The mean, min and max values of these distances was used as features. Respectively the same procedure was performed for the Fourier transform of the interval.

5.2.3 Points Of Interest

After surveying the average wave forms it could be concluded that the last point for each signal often showed a deviation from the average and could therefore be a good addition to the statistical features.

5.2.4 Power Of EEG-band Frequencies

Frequencies for each of the 13 signals were calculated using Fourier Transform. The amplitudes of the different frequencies were then grouped into the standard five EEG bands; delta (0-4 hz), theta (4-8 hz), alpha (8-12 hz), beta (12-30 hz) and gamma (30-45 hz). Each of these were used as features.

5.3 Feature Selection

Training a model with arbitrary features can cause overfitting, a way to reduce a models variance is using feature selection [20].RFE with a SVM was used to select the best number of features. To know which amount of features to select, a survey of the ROC AUC for different amounts of selected features was conducted. Figure 13 shows ROC AUC measurements for different classifiers, dependent on different amounts of selected features. Note that as the RFE was done on an SVM, future work could include implementing RFE with each of the classifiers. After this it was clear that the best number of features to use was *78*.

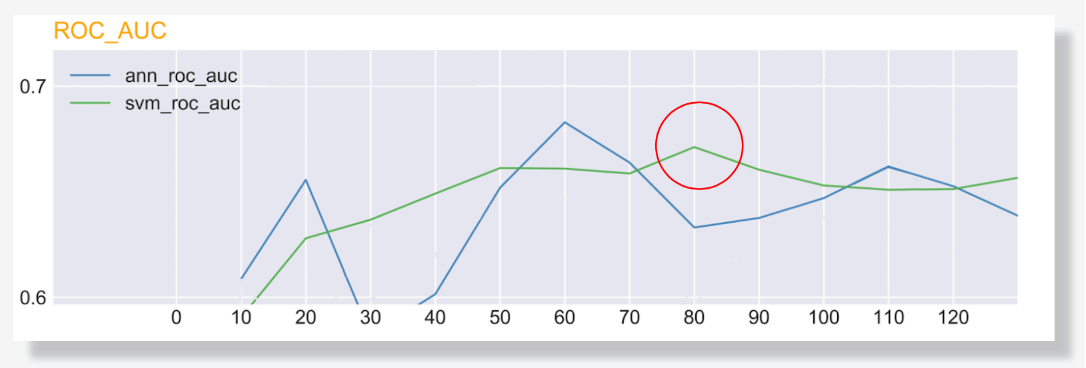

Figure 13: ROC AUC for the SVM and the ANNs, dependant on different amounts of selected features. The x-axis is the AUC ROC score and the y-axis is the amount of the selected best features.

5.4 Final Figure

The final four tests were made on 175 training samples and 75 test samples each. The results shown in table 2, are the mean results of the four tests.

Table 2: Results

Classifier	ROC AUC 1	ROC AUC 2	ROC AUC 3	ROC AUC	Mean ROC AUC
ANN	0.548	0.522	**0.630**	0.574	0,569
SVM	**0.563**	**0.530**	0.615	**0.591**	**0,575**

6 Conclusions

The results in table 2 show that the SVM performed marginally better than the ANN. This project has a bias towards SVM in the sense that the amount of features used was optimal for the SVM and not the ANN. The ANN was forced to train on the same selection of *best* features as the one found specifically selected for the SVM. There was also not enough time to properly tune the models. The results are inconclusive due the similarity in the results between SVM and ANN, the bias towards SVM, and the lack of tuning the models by altering their parameters and testing the changes on data. The conclusion that either model performed better on detecting sleep arousals in polysomnography data could not be made.

6.1 Discussion

There are several areas that could be improved in this project, adjustments of the model parameters and tests for evaluating the amount of valid information contained in each feature are 2 things that would have a large impact. Many of the areas that are lacking in this project come down the fact the project was too general. The delimitations of the project seemed reasonable starting the project further along it could be concluded that many of the areas that are lacking could be entire thesis projects by themselves. It is important to take in to account that both the ANN and SVM practically were un-tuned to the type of data we used. Tuning the parameters of an ANN was a task which we did not have time to perform adequately. There are several good guides and papers on the subject of tuning the parameters an SVM[9]. The implementation of a tuning algorithm for both classifiers was planned but only realized for the SVM. Due to time restraints, only 20 samples were tested, which is about 2 percent of the total data size. Hence these results could not be used, it might, however, have been better to use these rather than the defaults. The extraction of features was only run for the SVM. Extracting the best x features on 20 samples, where x is a number in a range of 1 to 390 with a step size of 5, took several hours. Doing this test on a more representable sample size, let us say 250 samples, would take up to several weeks if our calculations are right. Feature selection on SVM is usually done to generalize performance, shorten the computing run-time and constraint or classification issues in the data. However, there are sources that argue that feature selection isn't always necessary for all SVM problems and claim that it depends on the type and character of the data. There was no time to perform this test for the ANN as well, it would perhaps have been better to use all features to get a more fair comparison since we opted not to use the parameters derived from the test on a small sample size. Performing the evaluations on models trained on all features was not implemented, and might have been a good addition. Even though this project was inconclusive it can serve as a foundation for others interested in similar problems.

6.1.1 Future work

As mentioned there is a variety of areas in this project that could be improved, in order to build better classifiers and yield more conclusive results. As stated in the discussion, this paper makes up a good foundation for a scientifically valid comparison study between SVM and ANN for classifying sleep arousals using polysomnography data. But the project is in the current state too general.

These are the areas of improvement or of further exploration:

- Preprocessing
  There are some work left in preprocessing. Filtering and down sampling the frequencies in the signals are two examples. Using proper filtering and down sampling could reduce noise, shorten the run-time of the python program and reduce the size of RAM-memory needed. There is no indication if this would give a better classification performance or not.

- Feature extraction
  other features could be extracted, that could potentially be better than the existing features.

- Tuning the models.
  The models were tested without tuning them independently for this specific task. An tuning program was constructed in python for the Support Vector Machine but never tested on large data sets due to time limitations. It measures AUC ROC for the classifier while changing the C-value, Gamma-value, Kernel function and grade for the polynomial kernels. Lastly the results are saved in a table.
  A tuning algorithm for the artificial neural network was not implemented. Tuning both the classifiers is a necessary task for making a scientific comparison. The potential for the SVM and the ANN is not measured without tuning the classifiers. This is one of the biggest reasons that makes this study inconclusive at its current state.

- Feature selection methods.
  A questionable method was used for finding the best features. Firstly, the best features were only tested on a small amount of the data, namely 2 percent. Secondly, the best amount of features were only tested for the SVM, and the ANN was forced to work with the same amount, the creates a bias towards the SVM and furthermore makes the results inconclusive. A feature selection method that is optimized for linear regression was used. A suggestion for future work is to test different feature selection methods, e.g. *selectKbest* from the python library sklearn or other that are specifically made for a SVM and an ANN.

- K-Cross validation
  Implementing K-cross validation for a better calibration of the models. The purpose of cross validation is to determine the ability of a model to predict new information. Cross validation also gives information on if the model is

overfitted to training data or has a form of selection bias.

- Class balancing
  There are potentially better methods for class balancing than method used in this paper(under sampling). The method used in this paper changes the data set, this may not always be preferable. In some cases it is more preferable to rework the problem, using other methods for class balancing. An example of another technique that does not alter the data set in the same way is *class weighted* or *cost sensitive learning methods*.

- Evaluation methods
  AUC ROC was used in this project to measure performance. All the projects who participated in the Physionet challenge where evaluated using *The Area Under Precision-Recall Curve* (AUPRC). To compare the results an AUPRC evaluation method would would enable a possibility for comparison between this project a variety of similar projects connected to the Physionet challenge.

- More memory
  The limitations in not meeting the memory requirements made it impossible to perform training and testing on the whole data set. The code developed for this project needed at least 41 GB of RAM to be able to test all the data at once. A alternative to acquiring more RAM is to down sample the data, this in turn could affect the classification accuracy.

This project is inconclusive in its present state, mostly due to time constraints. It is on the other hand a good starting foundation for future work and testing in the area of comparing the accuracy of different machine learning classifiers using polysomnography data. Our program is not limited for just SVM and ANN, it is a good foundation for testing most of the machine learning classifiers. The areas for further research and improvements are preprosessing, feature extraction, feature selection and a necessary tuning of the classification model parameters after the type of the data set.

References

[1] Armon, Carmel. *Polysomnography*. Accesed: 2019-05-07.

[2] Asiri, Sidath. *Meet Artificial Neural Networks*. Accesed: 2019-05-13. URL: `https://towardsdatascience.com/meet-artificial-neural-networks-ae5939b1dd3a`.

[3] Biswal, Siddharth et al. “Expert-level sleep scoring with deep neural networks”. In: *Journal of the American Medical Informatics Association* 25.12 (2018), pp. 1643–1650.

[4] Brownlee, Jason. *UHow to Configure the Number of Layers and Nodes in a Neural Network*. Accesed: 2019-06-03. URL: `https://machinelearningmastery.com/how-to-configure-the-number-of-layers-and-nodes-in-a-neural-network/`.

[5] Cortes, Corinna and Vapnik, Vladimir. “Support-Vector Networks”. In: *Machine Learning*. 1995, pp. 273–297.

[6] Deep AI, Inc. *Activation Function*. Accesed: 2019-05-13. URL: `https://deepai.org/machine-learning-glossary-and-terms/activation-function`.

[7] developers, scikit-learn. *Feature ranking with recursive feature elimination*. Accesed: 2019-05-13.

[8] Dodge, Y. and Institute, International Statistical. *The Oxford Dictionary of Statistical Terms*. Oxford University Press, 2006. ISBN: 9780199206131. URL: `https://books.google.se/books?id=%5C_OnjBgpuhWcC`.

[9] Fraj, Mohtadi Ben. *In Depth: Parameter tuning for SVC*. Accesed: 2019-05-13. URL: `https://medium.com/all-things-ai/in-depth-parameter-tuning-for-svc-758215394769`.

[10] Genaro Salierno OMSCS Computer Science Interactive Intelligence, Georgia Institute of Technology (2019). *How do I define the input and output of neural network system?* Accesed: 2019-05-13. URL: `https://www.quora.com/How-do-I-define-the-input-and-output-of-neural-network-system`.

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

[12] Goodle. *Classification: ROC Curve and AUC*. Accesed: 2019-05-13. URL: `https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc`.

[13] Guyon, Isabelle et al. “Gene selection for cancer classification using support vector machines”. In: *Machine learning* 46.1-3 (2002), pp. 389–422.

[14] Howe-Patterson, Matthew, Pourbabaee, Bahareh, and Benard, Frederic. “Automated Detection of Sleep Arousals From Polysomnography Data Using a Dense Convolutional Neural Network”. In: *signal* 1 (2018), p. 2.

[15] Illowsky, B. and Dean, S. *Collaborative Statistics*. Open textbook library. Illowsky Publishing, 2008. ISBN: 9780978745073. URL: https://books.google.se/books?id=p2RtPQAACAAJ.

[16] Japkowicz, Nathalie and Stephen, Shaju. "The class imbalance problem: A systematic study". In: *Intelligent data analysis* 6.5 (2002), pp. 429–449.

[17] Jeeva, Manikandan. *The Scuffle Between Two Algorithms -Neural Network vs. Support Vector Machine*. Accesed: 2019-05-13. URL: https://medium.com/analytics-vidhya/the-scuffle-between-two-algorithms-neural-network-vs-support-vector-machine-16abe0eb4181.

[18] Khandoker, Ahsan H, Palaniswami, Marimuthu, and Karmakar, Chandan K. "Support vector machines for automated recognition of obstructive sleep apnea syndrome from ECG recordings". In: *IEEE transactions on information technology in biomedicine* 13.1 (2009), pp. 37–48.

[19] Lee, Miryoung et al. "Sleep disturbance in relation to health-related quality of life in adults: the Fels Longitudinal Study". In: *JNHA-The Journal of Nutrition, Health and Aging* 13.6 (2009), pp. 576–583.

[20] Munson, M Arthur and Caruana, Rich. "On feature selection, bias-variance, and bagging". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2009, pp. 144–159.

[21] Narkhede, Sarang. *Understanding AUC ROC Curve*. Accesed: 2019-05-13. URL: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.

[22] Nutt, David, Wilson, Sue, and Paterson, Louise. "Sleep disorders as core symptoms of depression". In: *Dialogues in clinical neuroscience* 10.3 (2008), p. 329.

[23] Peters, Brandon. *Arousal During the Stages of Sleep*. Accessed: 2019-03-22.

[24] PhysioNet. *PhysioNet*. Accesed: 2019-03-26. URL: https://physionet.org.

[25] PhysioNet. *PhysioNet*. Accesed: 2019-03-26. URL: https://physionet.org/challenge/2018/.

[26] Singh, Seema. *Understanding the Bias-Variance Tradeoff*. Accesed: 2019-05-13. URL: https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229.

[27] Singla, Rajesh et al. "Comparison of SVM and ANN for classification of eye events in EEG". In: *Journal of Biomedical Science and Engineering* 4.01 (2011), p. 62.

[28] Taylor, Courtney. *What is kurtosis*. Accesed: 2019-05-05.

[29] Weisstein, Eric W. *"Covariance." From MathWorld–A Wolfram Web Resource*. Accesed: 2019-05-13. URL: http://mathworld.wolfram.com/Covariance.html.

[30] Yildirim, Ozal, Baloglu, Ulas Baran, and Acharya, U Rajendra. “A deep learning model for automated sleep stages classification using psg signals”. In: *International journal of environmental research and public health* 16.4 (2019), p. 599.

www.kth.se